University of Gdańsk
Faculty of Chemistry
Department of Environmental Chemoinformatics

Viacheslav Muratov

A properties-to-omics predictive framework for hazard assessment of metal- and carbon-based nanomaterials

Ph.D. Thesis

Supervisor:
Prof. dr hab. Tomasz Puzyn

Co-supervisor:
Dr inż. Karolina Jagiełło

Gdańsk 2026

Uniwersytet Gdański
Wydział Chemii
Katedra Chemii i Radiochemii Środowiska

Viacheslav Muratov

Predykcyjne podejście łączące właściwości materiałów z danymi omicznymi w ocenie zagrożeń związanych z nanomateriałami na bazie metali i węgla

Rozprawa doktorska

Promotor:
Prof. dr hab. Tomasz Puzyn

Promotor pomocniczy:
Dr inż. Karolina Jagiełło

Gdańsk 2026

Table of Contents

Acknowledges **2**

List of publications included in the dissertation **3**

Abbreviations **5**

Summary **7**

Streszczenie **9**

1. Introduction **11**

1.1 Engineered nanomaterials: forms and applications 11

1.2 Nanomaterials exposure scenarios in the environment 13

1.3 Pulmonary responses to nanoparticle exposure 15

1.4 Chemical risk assessment and nanosafety frameworks 21

1.5 Biomarkers and toxicogenomics in nanosafety 24

1.6 Computational toxicology and machine learning for high-dimensional data 28

2. Research Challenge, Hypothesis, Objectives **32**

2.1 Research Challenge and Main Hypothesis 32

2.2 Detailed Research Hypotheses 33

2.3 Detailed Research Objectives 34

3. Results **36**

3.1 Linking MWCNTs properties to pulmonary transcriptomic responses through the Transcriptomic Response Index 36

3.2 Transcriptomic Response Index across Titanium-based ENMs 43

3.3 Titanium-based ENMs exposure biomarkers 53

4. Conclusion **61**

5. References **63**

6. Co-author declarations **80**

7. Publications included in the dissertation **88**

Acknowledges

This work was funded via the National Science Center in the frame of the TransNANO project (UMO-2020/37/B/ST5/01894). Figures for this work were prepared using the software https://BioRender.com. I acknowledge Epic Games, Inc. for providing access to Unreal Engine 5 under their standard licensing terms.

I thank my supervisor prof. dr hab. Tomasz Puzyn and co-supervisor dr inż. Karolina Jagiełło for their support in conducting this doctoral project. My appreciation also goes to my colleagues from the Laboratory of Environmental Chemoinformatics.



List of publications included in the dissertation

1. **Muratov, V**.; Jagiełło, K.; Puzyn, T., TRIumph in Nanotoxicology: Simplifying Transcriptomics into a Single Predictive Variable, *Nanoscale Horiz* **2025**, *10*, 3116–3126. DOI: 10.1039/D5NH00330J.
2. **Muratov, V**; Jagiełło, K.; Puzyn, T, An Integrative Omics-Machine Learning Framework for Predicting Pulmonary Responses to Titanium Dioxide Nanoparticles, *Nanoscale,* **2026**, *18*, 7021-7033. DOI: 10.1039/D5NR04657B
3. **Muratov, V**.; Jagiełło, K.; Mikolajczyk, A.; Danielsen, P. H.; Halappanavar, S.; Vogel, U.; Puzyn, T., The Role of Machine Learning in Predicting Titanium Dioxide Nanoparticles Induced Pulmonary Pathology Using Transcriptomic Biomarker, *J Hazard Mater* **2025**, *493*, 138240. DOI: 10.1016/j.jhazmat.2025.138240.

International Conference Presentations

1. **Muratov, V.**; Jagiełło, K.; Puzyn, T. (2025): Beyond Animal Studies: AI-Driven Toxicogenomics for Next-Generation Titanium Dioxide Nanoparticle Safety. *4th International Online Conference on Materials*, 3–6 November 2025, Basel, Switzerland, Oral
2. **Muratov, V.**; Jagiełło, K.; Puzyn, T. (2026): From Atomic Physics to Predictions: AI-Assisted Screening of Heavy-Metal Removal by Metal-Oxide Nanomaterials. *The 1st International Online Conference on Atoms*, 29–30 January 2026, Basel, Switzerland, Oral
3. **Muratov, V.**; Jagiełło, K.; Puzyn, T. (2026): Computational Toxicogenomic Assessment of Inhalation Risks from Titanium Dioxide Sanding-Dust Particulate Emissions. *The 1st International Online Conference on Environments*, 2–3 March 2026, Basel, Switzerland, Oral
4. **Muratov, V**, Jagiełło, K., Mikołajczyk, A., Puzyn, T. (2024): Exploring the Role of Machine Learning in Identifying Potential Transcriptomic Biomarkers in Pulmonary Diseases Caused by Titanium Dioxide Nanoparticles. *11th International Conference on Nanotoxicology - NanoTox 2024*, 23-25 September 2024, Venice, Italy, Poster

5. **Muratov, V**.; Jagiełło, K.; Puzyn, T. (2026): Comprehensive Transcriptomic Response Index for Decoding Nanomaterial-Induced Pulmonary Responses. *SETAC Europe 36th Annual Meeting*, 17–21 May 2026, Maastricht, The Netherlands, Poster

Other publications

1. Jagiełło, K.; Ciura, K.; **Muratov, V**.; Merugu, S.; Halappanavar, S.; Danielsen, P.; Nicklas, R.; Mikolajczyk, A.; Vogel, U. Grouping Nanoparticles Based on Properties and Transcriptomic Response: Are We Dealing with a Single Nanoform or a Set of Nanoforms with Common Pulmonary Hazards? *Environ. Sci.: Nano*, **2026**, *13*, 1085–1096. DOI: 10.1039/D5EN01090J
2. Snigur, D.; Dubovyi, V.; Arabadji, M.; Bevziuk, K.; **Muratov, V**.; Barbalat, D.; Zhukovetska, O. A Novel Fast Room-Temperature Cloud Point Extraction Coupled to Graphite Furnace Atomic Absorption Spectroscopy for Copper(II) Traces Determination. *Int J Environ Anal Chem* **2024**, *104* (17), 5254–5263. DOI: 10.1080/03067319.2022.2118593.
3. Nesterkina, M.; **Muratov, V**.; Ognichenko, L.; Kravchenko, I.; Kuz'min, V. Quantitative Structure–Activity Relationship Study on Prolonged Anticonvulsant Activity of Terpene Derivatives in Pentylenetetrazole Test. *Open Chem* **2021**, *19* (1), 1184–1192. DOI: 10.1515/chem-2021-0108.

ABBREVIATIONS

0D / 1D / 2D / 3D	zero-/one-/two-/three-dimensional
AD	applicability domain
AOP / AOPs	Adverse Outcome Pathway(S)
ASINA	Anticipating Safety Issues at the Design Stage of NAno Product Development
BET	Brunauer–Emmett–Teller specific surface area
CNT/F	carbon nanotubes and nanofibers
DEG / DEGs	differentially expressed gene(s)
DF4nanoGrouping	decision-making framework for the grouping and testing of nanomaterials project
DNA	deoxyribonucleic acid
ENM / ENMs	engineered nanomaterial(s)
FAIR	Findable, Accessible, Interoperable and Reusable
FDR	false discovery rate
GA	genetic algorithm
GRACIOUS	Grouping, Read-Across, Characterization and Classification Framework for Regulatory Risk Assessment of Manufactured Nanomaterials and Safer Design of Nano-Enabled Products project
GridSearchCV	grid search with cross-validation
IATA / IATAs	Integrated Approach(es) to Testing and Assessment
LightGBM	Light Gradient Boosting Machine
ML	machine learning
MoA	mode of action
mRNA	messenger ribonucleic acid
MWCNT / MWCNTs	multi-walled carbon nanotube(s)
NAM / NAMs	new approach methodology/methodologies
NP / NPs	nanoparticle(s)
OECD	Organization for Economic Co-operation and Development
PC / PCs	principal component(s)
PCA	principal component analysis
PoD / PoDs	point(s) of departure
PTML	perturbation-based machine learning

qPCR	quantitative polymerase chain reaction
qRT-PCR	quantitative reverse-transcription polymerase chain reaction
QSAR / nano-QSAR	Quantitative Structure–Activity Relationship / Quantitative Structure–Activity Relationship for nanomaterials
Q^2	external validation coefficient
Q^2_{CV}	cross-validated determination/predictivity coefficient
R^2	coefficient of determination
$RMSE_C$	root mean square error of calibration
$RMSE_{CV}$	root mean square error of cross-validation
$RMSE_{EXT}$	root mean square error of external validation
RF	random forest
RNA	ribonucleic acid
RNA-seq	RNA sequencing
ROC-AUC	area under the receiver operating characteristic curve
SbD	Safe-by-Design
scRNA-seq	single-cell RNA sequencing
SiRMS	Simplex Representation of Molecular Structure
SMOTE	Synthetic Minority Oversampling Technique
SSbD	Safe-and-Sustainable-by-Design
SWCNT / SWCNTs	single-walled carbon nanotube(s)
TGx	toxicogenomics
TiO_2-NP / TiO_2-NPs	titanium dioxide nanoparticle(s)
tPOD	transcriptomic point of departure
TRI	Transcriptomic Response Index
UV	ultraviolet
XGBoost	Extreme Gradient Boosting

Summary

The work described in this dissertation develops a predictive framework that links properties of engineered metal- and carbon-based nanomaterials to quantitatively predictable changes in transcriptomic profiles, with relevance to lung toxicity. This work tackles a key challenge in the field of nanosafety: traditional animal studies are costly and tend to focus on late-stage endpoints, while transcriptomic readouts are able to detect early molecular outcomes in the body, but the sheer complexity and dimensionality of gene-expression data make it difficult to use such data in routine risk assessments. To address this limitation, the dissertation integrates characterizations of nanomaterials, data from toxicogenomic studies, and machine learning approaches along with Organization for Economic Co-operation and Development (OECD)-approved validation strategies into a single, coherent computational workflow. In this way, the dissertation contributes to ongoing efforts to establish machine learning approaches that are applicable across diverse classes of engineered nanomaterials, while combining data-driven predictive modeling with mechanistic interpretability to support the development of next-generation *in silico* New Approach Methodologies (NAMs) for safety assessment.

This work examined two representative classes of toxicologically relevant nanomaterials – titanium dioxide nanoparticles (representatives of metal-based/metal-oxide nanomaterials) and multi-walled carbon nanotubes (representatives of high-aspect-ratio, carbon-based nanomaterials) – that differ markedly in chemical composition, shape, longevity, and anticipated mechanisms of action. These nanomaterials, therefore, provide a useful test case for assessing whether a broadly applicable machine learning framework can be established and optimized for predicting transcriptomic alterations across diverse classes of engineered nanomaterials. The central hypothesis of the dissertation is that high-dimensional transcriptomic data can be reduced to predictive, biologically interpretable endpoints suitable for mechanistic hazard assessment and potency ranking, and that these endpoints, together with selected transcriptomic biomarkers, can be modeled from nanomaterial physicochemical properties and exposure-related descriptors using validated machine learning approaches. Collectively, the work described in this dissertation extends the reach of nanosafety modeling beyond single endpoint assays and singular gene predictions, demonstrating the feasibility of developing and validating machine learning (ML) workflows that predict both particular transcriptomic biomarkers and broad indices of global transcriptional

activity from nanomaterial descriptors. The developed framework simplifies the predictions of early molecular responses in processes related to inflammation, fibrosis, oxidative stress, and tissue repair. Importantly, the predictive models were constructed according to the regulatory-established guidelines of internal and external validation, and the estimation of the applicability domain of the trained models. The rigorous validation underlines the potential applicability of the developed framework in the context of regulatory toxicology, Integrated Approaches to Testing and Assessment (IATA), and hazard evaluation based on non-animal methods.

This dissertation links the fields of materials chemistry and systems toxicology by establishing a scalable methodology that integrates physicochemical characterization of exposed nanomaterials with omics datasets, dimensionality-reduction techniques, and validated machine learning approaches, in order to provide an applicable workflow for converting complex transcriptomic information into a mechanistically interpretable endpoint. The resulting approach has the potential to reduce the need for large-scale animal experimentation, aid early warning of potentially hazardous properties of nanomaterials, contribute to the ranking of their potencies, and help to inform the development of safer nanomaterials.

Streszczenie

W niniejszej rozprawie opisano proces opracowywania modelu prognostycznego, który łączy właściwości inżynieryjne nanomateriałów na bazie metali i węgla z możliwymi do ilościowego przewidywania zmian w profilach transkryptomicznych, mających znaczenie dla toksyczności płucnej. Praca ta podejmuje kluczowe wyzwanie w dziedzinie bezpieczeństwa nanomateriałów: tradycyjne badania na zwierzętach są kosztowne i zazwyczaj koncentrują się na punktach końcowych w późnych stadiach, podczas gdy odczyty transkryptomiczne pozwalają wykrywać wczesne zmiany molekularne w organizmie, jednak ogromna złożoność i wielowymiarowość danych dotyczących ekspresji genów utrudniają wykorzystanie takich danych w rutynowej ocenie ryzyka. W odpowiedzi na to wyzwanie, w niniejszej pracy doktorskiej zintegrowano informacje o strukturze i właściwościach nanomateriałów, dane z badań toksykogenomicznych oraz metody uczenia maszynowego wraz ze strategiami walidacji zatwierdzonymi przez OECD w ramach jednego, spójnego procesu obliczeniowego. W ten sposób praca ta wnosi wkład w trwające wysiłki na rzecz ustanowienia metod uczenia maszynowego, które będą miały zastosowanie w odniesieniu do różnych klas nanomateriałów inżynieryjnych, łącząc jednocześnie modelowanie predykcyjne oparte na danych z interpretowalnością mechanistyczną w celu wsparcia rozwoju modeli *in silico* nowej generacji do oceny bezpieczeństwa nanomateriałów (NAM).

W niniejszej pracy zbadano dwie reprezentatywne klasy nanomateriałów istotne z punktu widzenia toksykologii – nanocząstki dwutlenku tytanu (przedstawiciele nanomateriałów na bazie metali/tlenków metali) oraz wielościenne nanorurki węglowe (przedstawiciele nanomateriałów na bazie węgla o wysokim współczynniku kształtu) – które znacznie różnią się składem chemicznym, kształtem, trwałością oraz przewidywanymi mechanizmami działania. Nanomateriały te stanowią zatem użyteczny przypadek testowy do oceny, czy można stworzyć i zoptymalizować szeroko stosowaną strukturę uczenia maszynowego do przewidywania zmian transkryptomicznych w różnych klasach nanomateriałów inżynieryjnych. Główna hipoteza niniejszej rozprawy zakłada, że wielowymiarowe dane transkryptomiczne można zredukować do predykcyjnych, biologicznie interpretowalnych punktów końcowych, odpowiednich do mechanistycznej oceny zagrożenia i klasyfikacji siły działania, oraz że punkty te, wraz z wybranymi biomarkerami transkryptomicznymi, można modelować na podstawie właściwości fizykochemicznych nanomateriałów i deskryptorów związanych z narażeniem, wykorzystując zweryfikowane metody

uczenia maszynowego. Podsumowując, prace opisane w niniejszej rozprawie rozszerzają zakres modelowania bezpieczeństwa nanomateriałów poza testy oparte na pojedynczych wskaźnikach i prognozy dotyczące pojedynczych genów, wykazując wykonalność opracowania i walidacji procesów uczenia maszynowego, które na podstawie deskryptorów nanomateriałów pozwalają przewidywać zarówno konkretne biomarkery transkryptomiczne, jak i szerokie wskaźniki ogólnej aktywności transkrypcyjnej. Opracowana struktura ułatwia prognozowanie wczesnych reakcji molekularnych w procesach związanych ze stanem zapalnym, zwłóknieniem, stresem oksydacyjnym i odbudową tkanek. Co istotne, modele prognostyczne zostały skonstruowane zgodnie z ustalonymi wytycznymi regulacyjnymi dotyczącymi walidacji wewnętrznej i zewnętrznej oraz oszacowania zakresu zastosowania wytrenowanych modeli. Rygorystyczna walidacja podkreśla potencjalną przydatność opracowanej struktury w kontekście toksykologii regulacyjnej, zintegrowanych podejść do badań i oceny (IATA) oraz oceny zagrożeń opartej na metodach bez wykorzystania zwierząt.

Niniejsza rozprawa łączy dziedziny chemii materiałowej i toksykologii systemowej poprzez ustanowienie skalowalnej metodologii, która integruje charakterystykę fizykochemiczną narażonych nanomateriałów z zestawami danych omicznych, technikami redukcji wymiarowości oraz zweryfikowanymi podejściami uczenia maszynowego, w celu zapewnienia praktycznego schematu postępowania służącego przekształcaniu złożonych informacji transkryptomicznych w punkt końcowy możliwy do interpretacji pod kątem mechanizmów działania. Opracowane podejście może potencjalnie ograniczyć potrzebę przeprowadzania eksperymentów na zwierzętach na dużą skalę, wspomóc wczesne ostrzeganie o potencjalnie niebezpiecznych właściwościach nanomateriałów, przyczynić się do ustalenia rankingu ich szkodliwości oraz pomóc w opracowywaniu bezpieczniejszych nanomateriałów.

1. Introduction

1.1 Engineered nanomaterials: forms and applications

Engineered nanomaterials (ENMs) are characterized by having at least one dimension in the range of roughly 1-100 nm. The behavior of materials in this size-dependent regime is primarily affected by the possession of a high surface-to-volume ratio, and quantum confinement occurs [1, 2]. ENMs vary in shape, their physicochemical properties, applicability, and the way they interact biologically – for instance, properties such as strength, optical characteristics, and chemical reactivity often depend on the considered scale [1]. ENMs are generally classified according to the number of dimensions that remain within the nanoscale range. In this context, "dimensionality" refers to how many external dimensions of a material are confined to approximately 1-100 nm. Zero-dimensional (0D) nanomaterials have all three dimensions in the nanoscale range; spherical nanoparticles are a typical example. One-dimensional (1D) nanomaterials have two dimensions in the nanoscale range, while the third extends beyond it (i.e., nanowires and nanotubes). Two-dimensional (2D) nanomaterials have only one nanoscale dimension, usually thickness, whereas the other two dimensions are larger; nanofilms, plates, and sheets fall into this category [3]. Composite materials are the only form in which three-dimensional objects appear [4]. In many cases, commercial products containing ENMs are based on three-dimensional (3D) matrices (for example, coatings, polymers, or construction materials). When ENMs are embedded in such matrices, their potential to be released, exposed, and transformed is different from dispersions of "free" nanoparticles. ENMs are also frequently classified according to their chemical origin or composition [5], and the categories of ENMs include metal nanoparticles (i.e., silver, gold, and copper), metal-oxide nanoparticles (i.e., titanium dioxide, zinc oxide, and the iron oxides), carbon-based nanoparticles (including single- or multi-walled carbon nanotubes (SWCNT/MWCNT)), nanopolymers, and nanocomposites [3, 5].

Studies indicate a significant increase in the use of ENMs across a range of industries such as pharmaceutics, energy, and optics [3, 6, 7]. This wide applicability is possible since ENMs exhibit several properties – optical, catalytic, mechanical, and electrical – which are not readily achieved with ordinary micron-sized or bulk materials. For instance, ENMs can be employed to design drugs that selectively target particular cells or organs and can improve the efficacy of cancer therapy [8]. In the

field of nanomedicine, ENMs can serve as targeted carriers, as contrast agents for imaging, or as theranostic systems, where diagnosis and therapy are combined in a single nanoformulation [3, 8]. Beyond medicine, ENMs are used as additives in textiles, in cement, or in food, where they can act either as dyes or as strengtheners [9–11]. The electronics sector is also an area exploiting ENMs – for example, the application of carbon nanotubes as electrodes and in wearable technologies [12]. Studies together with current market figures corroborate the potential of ENMs – for example, the size of the ENM market was USD 16.3 billion in 2023, and it is expected to grow to more than USD 50.8 billion by 2033 [13].

1.1.2 Titanium dioxide nanoparticles

Titanium-based ENMs accounted for more than 23.7% of the total nanomaterials market in 2025, which makes them the most widely used ENMs worldwide [13]. This widespread use reflects their relatively low production cost, chemical stability, and favorable optical and surface properties. Titanium dioxide nanoparticles (TiO_2-NPs) exist mainly in the rutile and anatase phases [14] and are widely applied as white pigments and ultraviolet (UV)-blocking agents in coatings, paints, plastics, cosmetics, and sunscreens [15]. Moreover, owing to their photocatalytic and antimicrobial activity, TiO_2-NPs are increasingly explored in technical applications such as self-cleaning materials and water-treatment membranes, where they can improve hydrophilicity, reduce fouling, and enhance antimicrobial performance [16]. In addition to these industrially relevant uses, TiO_2-NPs have also been investigated in nanomedicine, where surface functionalization may improve their suitability for targeted therapeutic applications [17].

Recent work indicates that TiO_2-NPs may offer an alternative means of enhancing both the antimicrobial activity and the hydrophilicity of membranes that are used for treating water [18]. Embedding TiO_2-NPs in either polymeric or ceramic membranes can lead to improvements in resistance to fouling, increased permeability to water, and can also prolong the life of the membrane, which is of particular importance for technologies aimed at purifying water.

1.1.3 Multi-walled carbon nanotubes

Cylindrical nanotubes known as MWCNTs consist of several layers of graphene [19]. The layered form affects stiffness, defect tolerance, and the mechanical behavior of the tubes. An MWCNT may be regarded as a coaxial bundle of SWCNTs of varying diameters assembled into a multilayer tube [19]. The diameter of these tubes is

usually between 2 nm and 100 nm, and the length can extend to the micrometer scale [20].

Evidence from preclinical work suggests that intratumoral injection of MWCNTs slows tumor progression and is associated with an increase in the numbers of CD8+ and CD4+ T cells in the spleen, which supports immune responses within draining tumor tissue and lymph nodes [21]. When used in combination with cadmium oxide nanoparticles, MWCNT-based formulations maintained HeLa cell viability above 50% under the tested conditions, suggesting sufficient cytocompatibility to warrant further investigation for drug-delivery and other therapeutic applications [22]. Although no clinical investigations of MWCNTs have yet been carried out, a recent review has shown that they have considerable promise in the treatment of breast cancer [23]. Taken together, the preclinical data imply that appropriate functionalization of the surface of MWCNTs and optimization of dose may widen the margin within which MWCNTs can be used safely and effectively in nanomedicine aimed at treating tumors.

Electrode systems based on MWCNTs share distinctive traits, such as high sensitivity, stability, and flexibility, which favor good electrochemical performance and have important implications for the development of wearable sensors [24]. The quantum confinement properties of MWCNTs also make them attractive candidates for incorporation into supercapacitor technology [25]. When MWCNTs are modified, they are able to furnish well-defined routes for electrical conduction, which in turn enhances their suitability for applications in vapor detection as well as in electronic devices [26]. The success of MWCNTs' employment in sensors and in supercapacitors is largely attributable to their high electrical conductivity and excellent thermal stability [27]. For this reason, architectures based on MWCNTs are increasingly regarded as attractive building blocks for next-generation devices in energy storage and sensing, where lightweight, flexible, and durable conductive networks are required.

1.2 Nanomaterials exposure scenarios in the environment

1.2.1 Pulmonary exposure to ENMs scenarios

During many occupational and industrial processes, humans are exposed to TiO_2-NPs and MWCNTs. Employees who manufacture or work with these ENMs routinely inhale airborne nanoparticles either during production, handling, or in

subsequent stages of processing [28, 29]. In most workplaces, the aerosols that are encountered are not individual primary nanoparticles but rather agglomerates and aggregates. Nevertheless, these larger entities can still be of respirable size and therefore may deposit efficiently in the respiratory tract. Dust clouds produced by TiO_2-NPs and by MWCNTs increase the risk of inhalation for people working in the ENM-rich occupational environment, leading to particular concern about ENMs' safety in industrial settings (for instance, in Denmark, the occupational exposure limit for Printex90 black nanoparticles has been set at 3.5 mg/m^3[30]). Importantly, exposure limits for ENMs are usually expressed in terms of mass concentration, whereas biologically relevant dose metrics may also include particle number and surface area. For materials with a high aspect ratio, such as MWCNTs, shape is also important because length and diameter influence aerodynamic behavior, deposition, and clearance within the lung [31, 32].

Following inhalation, ENMs may persist in lung tissue and, in some cases, translocate across the blood-air barrier into the systemic circulation [33, 34]. The degree to which ENMs become systemically distributed in the body is largely determined by characteristics of the material, as well as by the dose to which organisms are subjected and by the condition of the airway barrier. These factors influence both local pulmonary and systemic responses to ENM exposure. At the molecular and cellular levels, inhaled ENMs may induce oxidative stress, deoxyribonucleic acid (DNA) damage, and inflammatory signaling, whereas persistent or unresolved injury may contribute to tissue-level consequences such as fibrosis [35]. In addition, it is important to recognize that secondary exposure can occur when ENMs move through tissue barriers into the bloodstream and then concentrate in other organs such as the spleen, liver, or kidneys [36].

1.2.2 The respiratory system in the context of inhaled nanomaterials

Inhalation leads to deposition of ENMs in different regions of the respiratory tract; the extent to which this occurs depends on properties such as size, shape, and other physicochemical characteristics of the particles [37]. Regional deposition, therefore, plays a decisive role in determining the local dose that is delivered, how long the particles remain within a particular part of the tract, and which cellular targets are encountered first. It is generally observed that ENMs deposit in three principal regions of the airway – roughly half of the 20 nm nanoparticles deposit in the alveolar region, whereas only about 15% of the inhaled nanoparticles end up in

the nasopharyngeal and tracheobronchial regions [38]. Because ultrafine particles are governed to a greater extent by Brownian diffusion, they tend to deposit efficiently in the distal, particularly alveolar, regions of the lung. The deposition of nanoparticles in the respiratory tract is largely governed by the thermodynamic or aerodynamic diameter of the particles that are inhaled [39]. Rodent studies of inhalation and of the biokinetics of TiO_2-NPs have shown that these particles penetrate deeply into lung tissue and elicit measurable biokinetic responses [39–41] – findings that underline the importance of inhalation experiments because they provide realistic information on where deposition takes place and what happens at the tissue level immediately after exposure. MWCNTs and other high aspect ratio nanoparticles have been shown to produce acute toxicity and other adverse effects when inhaled, and the biological effects are strongly influenced by the diameter and length of the particles because these dimensions determine surface activity, clearance, and deposition within the lung [42].

1.3 Pulmonary responses to nanoparticle exposure

1.3.1 Deposition, clearance, and particle overload

The initial factor determining how inhaled substances interact with lung tissue is the site of deposition of those materials, as well as the duration of their residency within that site [43, 44]. The physical properties of aggregates, particularly their size and density, play a major role in governing this process because they influence the aerodynamic behavior of inhaled particles and therefore determine how efficiently particles reach different regions of the respiratory tract [43]. For compact particles such as TiO_2-NPs, deposition is governed mainly by aerodynamic diameter, which reflects agglomerate size and density, whereas for fibrous or high-aspect-ratio materials such as MWCNTs, aerodynamic behavior is additionally influenced by length, diameter, and overall stiffness [45, 46]. These distinctions are consequential because clearance mechanisms differ by location: particles that land in the conducting airways are cleared relatively quickly by the mucociliary transport system, whereas particles that reach the alveoli depend mainly on clearance by macrophages [43, 46]. The same regional clearance pathways apply to both compact and fibrous ENMs, but high-aspect-ratio materials may be cleared less efficiently because their shape can hinder complete macrophage engulfment and thereby prolong retention [43, 46].

The literature consistently reports that nanoparticles behave differently from larger particles: studies of inhaled particles below 100 nm are taken up less efficiently by phagocytes than larger microparticles, tending instead to persist within epithelial or interstitial tissues and only occasionally to appear in the lymphatic or vascular circulation [43]. This pattern is consistent with observations for poorly soluble nanomaterials more broadly, for which translocation to secondary organs is often reported to account for less than 1% of total lung deposition, indicating that a substantial fraction of the material remains in the lung for prolonged periods [47]. Although patterns of clearance mustn't be uniform across all types of material [47, 48]. For example, inhaled iridium nano-aerosols have been reported to have lung retention half-times of several hundred days, with little dependence on size over the range of 10-75 nm, whereas other poorly soluble nanomaterials such as TiO_2, carbon black, and CeO_2 are more frequently reported to have shorter half-times, in the range of 40-60 days [47].

In periods when the amount of deposited material overwhelms the organism's capacity to clear it, conditions of overload arise [43, 44, 48]. Broad review of nanotoxicology describes lung particle overload as a situation where the amount of retained material exceeds that predicted by models based on linear kinetics, and may be associated with persistent inflammation, development of fibrosis, and the appearance of lung tumors [43].A mechanistic toxicokinetic study supports this, showing that low-, mid-, and high-exposure nanoparticle groups examined in that study were classified as non-overload, slight overload, and moderate overload, respectively [51]. More recent research on carbon black has characterized overload as a threshold beyond which the persistence of particles and their accumulation impair clearance and exacerbate lung damage – including inflammatory responses and fibrotic changes [44]. Although this overload framework has been described most explicitly for poorly soluble particulate materials, the pulmonary persistence of high-aspect-ratio nanomaterials such as MWCNTs is additionally influenced by fiber dimensions and biopersistence, which can also promote incomplete clearance and sustained inflammatory signaling [45, 46].

1.3.2 Acute and chronic inflammation in the lung

The most frequently reported lung tissue response to airborne substances is inflammation, which is often characterized by measurements of neutrophil influx together with levels of inflammatory mediators [43, 45]. Nevertheless, the expression of inflammatory phenotypes is not homogeneous across different types of material or across different forms of aerosols [49–51]. As shown for TiO_2-NPs inhalation, the

crystal form of the particles can influence which type of early inflammatory response is predominant [51]. Size-dependent immune effects were found with silver, where 15 nm particles produced a moderate inflammatory reaction characterized by a pronounced influx of neutrophils as well as increases in *IL-1β, MCP-1,* and *MIP-2* after 24 hours, whereas 410 nm particles did not provoke such a response [50].

Importantly, not all inflammatory responses persist [46, 49, 50]. An inhalation and instillation study of WO_3 describes the pulmonary reaction as transient and mild, noting that within the time periods studied, no signs of long-term inflammation or of fibrotic changes were apparent [49] – a finding that differs from that reported for more biopersistent fibrous substances [46]. For less biopersistent particles such as many TiO_2-NPs, inflammatory responses are often more transient, although sufficiently high retained burdens may still prolong inflammation. By contrast, fibrous carbon nanomaterials such as MWCNTs are more strongly associated with sustained macrophage activation, continued production of cytokines and reactive oxygen species, and persistent inflammatory infiltrates because their clearance is less efficient [46]. Similar distinctions appear throughout the wider literature on inflammation, where chronic inflammatory states are associated with later risks of developing conditions such as emphysema, fibrosis, and cancer [44, 45, 49].

Thus, the transition from acute to chronic inflammation depends not only on the intrinsic persistence of particles, but also on the magnitude of the retained lung burden and, in occupational settings, on repeated or continuous exposure that maintains that burden over time [44, 45, 49]. The analysis of retained surface area in particular makes clear distinctions between acute and chronic patterns of inflammation over both the timescale of exposures and of subsequent recoveries [45]. It also demonstrates that the persistence of lung burden may allow neutrophil levels to remain above those seen in controls for many months following exposure [45].

1.3.3 Oxidative stress, cell injury, and tissue remodeling

Studies repeatedly highlighted a close association between oxidative stress and cellular damage in cases of pulmonary toxicity [43, 45, 51]. A study of TiO_2 dust has noted that all small-agglomerate forms of the aerosol induced increases of oxidative damage, and are associated with increased levels of lactate dehydrogenase (LDH) [51]. By contrast, the larger aggregates tended to be characterized more by neutrophilic inflammation than by oxidative markers, leading the authors to conclude that the two physical forms of the particles produced distinct patterns of acute responses –

with the larger aggregates favoring recruitment of inflammatory cells, and the smaller ones being associated more with oxidative stress and with cytotoxic effects [51].

A chemically distinct response to silver nanoparticles was characterized, reporting that treatment with 15 nm particles was associated with increases in both total and reduced glutathione, elevations in LDH levels, and evidence for the presence of silver nanoparticles inside lung cells, whereas exposure to 410 nm particles did not trigger the same set of biochemical alterations [50]. The authors argued that differences in particle size might therefore underlie the observed differences in pulmonary inflammation as well as in cellular interactions, pointing specifically that the smaller particles were apparently more efficient at reaching the alveolar space [50].

The importance of these early-stage studies lies in the way that they continually link ongoing tissue damage with subsequent processes of tissue repair and remodeling [52]. In the review on CNTs, it is noted that extended periods of tissue damage can trigger activation of fibroblasts, differentiation of these cells into myofibroblasts, and increases in extracellular matrix synthesis, with consequent alterations to lung structure and to respiratory function [52]. Thus, whereas TiO_2-based studies in this section mainly illustrate how particle form can shift the balance between oxidative injury and neutrophilic inflammation in the acute phase, MWCNT-based studies are used to show how persistent injury can progress to fibroblast activation, extracellular matrix deposition, and focal fibrosis.

Evidence for tissue-level remodeling was also reported in the context of an inhalation experiment, where authors detected inhaled MWCNTs within macrophages, in subpleural mesenchymal cells, and within the collagenous matrix underlying focal subpleural lesions [53]. Moreover, they noted that high doses of these materials were associated with increases in fibrosis at both 2 and 6 weeks following exposure, and characterized these changes as focal, pointing out the presence of macrophage-rich mononuclear aggregates beneath the areas of fibrosis [53]. More broadly, descriptions of lung fibrosis based on the concept of Adverse Outcome Pathways (AOPs) tend to place extracellular matrix deposition downstream of persistent signals of inflammation and injury, resulting in septal thickening [54].

1.3.4 An adverse outcome pathway (AOP) perspective on nanoparticle-induced lung injury

The AOP framework is valuable for arranging disparate observations by linking early, molecular, or cellular events with subsequent outcomes at the organ or tissue

level within a clear, stepwise scheme [52, 54, 55]. In the context of fibrosis research, Clippinger et al. have proposed the use of AOPs as a way of structuring already known data, recognizing gaps in the current understanding, and guiding the choice of relevant *in vitro* readouts for assessing potential fibrotic hazards [55]. Similarly, the review focused on carbon nanotubes argues that AOPs provide a clear and transparent framework for spotting key mechanisms as well as for prioritizing tests to detect fibrogenic activity [52].

The most explicitly delineated signaling cascade relevant to lung fibrosis in the material cited is referred to as AOP 173, which charts starting with the engagement of lung cellular membranes by the inciting factor, and culminating in deposition of extracellular matrix and the development of fibrotic pathology [54]. The authors also note that disease processes triggered by MWCNTs appear to largely follow this general scheme, although the details of how MWCNTs activate lung cells are not yet fully worked out [54]. Within the literature, this fibrosis-oriented AOP is described most explicitly for MWCNTs; the AOP perspective provides a bridge between descriptive toxicology and mechanism-based hazard assessment, helping to explain why early molecular and transcriptomic readouts may be useful for predicting later adverse outcomes, and can also be applied to TiO_2-NPs-induced pathologies.

1.3.5 Dose metrics and particle properties as determinants of pulmonary toxicity

The extent of pulmonary injury depends not only on the concentration of the deposited material and on the intrinsic properties of the particles themselves, but also on the duration of exposure, the post-exposure period, and the persistence of the retained lung burden before effective clearance occurs [43, 45]. Thus, deposition dose should not be viewed as a static measure of concentration alone, but rather as a dynamic function of deposition, residence time, biopersistence, and clearance efficiency. This is particularly important in the lung, where the same deposited amount may lead to very different biological outcomes depending on whether the material is rapidly removed, slowly retained, or accumulates under overload conditions [43, 45].

In particular, the size and agglomerate size are key determinants of deposition because they govern aerodynamic and thermodynamic behavior within the respiratory tract and therefore influence which lung region receives the effective dose [43, 45]. These differences are toxicologically relevant because deposition in the alveolar region is generally associated with longer residence times and greater

dependence on macrophage-mediated clearance, whereas material deposited in the conducting airways may be removed more rapidly by mucociliary transport [46]. The biological consequences of particle size are illustrated by inhalation studies showing that small agglomerates of TiO_2 induced more serious oxidative damage and cytotoxicity, while larger aggregates were associated more with neutrophilic inflammation; similarly, smaller silver nanoparticles produced clearer inflammatory and biochemical responses than larger ones, indicating that size-dependent deposition and cell interaction can alter the pattern as well as the magnitude of pulmonary response.

Other physico-chemical characteristics further modulate how retained particles interact with lung cells and tissues. Surface charge influences the efficiency of cellular association and uptake, with cationic particles generally being more readily internalized by cells, which may intensify downstream inflammatory or cytotoxic responses [43]. Solubility is also critical, because rapidly dissolving particles may release toxic ions and thereby shift the mechanism of toxicity from particle-driven effects toward chemically mediated injury [43]. Surface reactivity, in turn, is strongly linked to the generation of reactive oxygen species and oxidative stress, and is therefore often associated with enhanced pulmonary inflammation even when the retained surface dose is relatively small [43, 45, 51]. Taken together, these properties do not merely alter particle-cell contact in a physical sense; they influence the likelihood of oxidative stress, membrane damage, cytokine release, persistent macrophage activation, and, ultimately, the progression toward chronic inflammation and tissue remodeling [43, 45, 51].

Accordingly, it is argued that the results of inhalation experiments ought to be discussed in terms of the amounts of material that are deposited or retained – ideally quantified in terms of mass, surface area, and particle number [43, 45]. Among these metrics, retained surface area appears particularly informative because it better captures the biologically active interface between the material and lung tissue than mass alone, especially for poorly soluble nanomaterials and high-surface-area particles [43, 45]. The evidence available for retained surface area as a metric of dose appears to be especially robust among the datasets considered [45]. Analyses of data from multiple studies on rats showed that measures of inflammation in response to TiO_2-NPs, carbon black, and MWCNTs varied in a sigmoidal manner as a function of retained surface area within the lungs.[45] This observation is consistent with the broader view developed in the preceding sections: persistent retained burden, rather than nominal exposure concentration alone, is a major determinant of whether

pulmonary responses remain acute and transient or instead progress toward chronic inflammation, overload, and fibrotic change [44, 45, 49].

Certain morphologies (i.e., long fibers, nanobelts, and platelet-like particles) are difficult to clear and tend to provoke frustrated phagocytosis, as well as being likely to trigger inflammatory or pleural reactions [43]. This is especially relevant for high-aspect-ratio materials such as MWCNTs, where length, diameter, and shape influence not only deposition but also the ability of macrophages to engulf and remove the material effectively [43, 46]. Rapidly dissolving particles can release toxic ions, which means that cationic particles are more readily internalized by cells, and that high surface reactivity is consistently associated with increased pulmonary inflammation, even when the amount of deposited surface is relatively small [43]. Importantly, no single physicochemical property is able to account for all cases of pulmonary inflammation across the range of nanoscale materials – although surface reactivity may serve as a kind of integrator of several important factors in this regard [43]. For this reason, pulmonary hazard assessment of nanomaterials should be based on an integrated consideration of dose metric, exposure duration, retained burden, clearance behavior, morphology, and surface-dependent reactivity, rather than on any single descriptor in isolation [43, 45, 51].

1.4 Chemical risk assessment and nanosafety frameworks

1.4.1 New Approach Methodologies (NAMs) to testing and assessment

The sheer number and variety of ENMs make it impractical to evaluate each one in detail through traditional approaches, and across the field, there is a persistent pressure to provide sufficient safety data and minimize the use of animal experiments [56–59]. The particular challenges of inhalation toxicology highlight that long-term, 90-day rodent studies are costly and have been characterized as imperfect surrogates for human responses, and at the same time, no single *in vitro* test or computational approach appears capable of fully describing the range of ways in which nanomaterials may cause damage to tissues [55, 60]. These emerging non-animal and mechanism-informed approaches are increasingly referred to as NAMs, which encompass *in vitro, in silico, in chemico*, and integrated testing strategies intended to

improve the efficiency, mechanistic relevance, and regulatory usefulness of hazard assessment [56–59].

A concept of the IATA was presented to fill the gaps in the regulatory risk assessment workflow and is described as a hypothesis-led ensemble of *in vitro* experiments, computational analyses, use of read-across, and considerations of toxicological concern thresholds and of exposure assessment, frequently analyzed within an AOP-inspired framework [57]. In this context, IATA can be regarded as one of the principal frameworks through which NAMs are organized and applied in regulatory toxicology. Contemporary approaches to nanosafety operationalize through decision trees and tiered testing programs customized to the route of exposure and the aims of the evaluation [61–63]. Under the auspices of Grouping, Read-Across, Characterization and Classification Framework for Regulatory Risk Assessment of Manufactured Nanomaterials and Safer Design of Nano-Enabled Products (GRACIOUS) and similar schemes, hypotheses for grouping of materials relevant to inhalation are organized around properties such as dissolution behavior, biopersistence, reactivity, and the potential to trigger inflammatory responses, with initial screening carried out at lower tiers [57, 61, 62] – a structure that also permits adjustment of the requirements for evidence to the needs of particular projects [62, 63].

Within these conceptual frameworks, neither grouping nor read-across is regarded as an informal workaround – their application is contingent upon clear identification of the relevant materials, scientifically defensible similarities, and pertinence to particular routes of exposure [56]. As a result, many modern classification systems are moving toward a synthesis of physicochemical similarity with data on exposure, fate, toxicokinetics, as well as mechanisms of action, and approaches that are driven by functional consideration, such as decision-making framework for the grouping and testing of nanomaterials (DF4nanoGrouping), are therefore presented as being closely allied to the broader group of IATAs [58, 62]. Accordingly, NAMs in nanosafety should be understood not as single replacement tests, but as coordinated evidence-generation strategies that integrate alternative experimental systems with computational and mechanistic approaches. The same sources repeatedly indicate that challenges of reducing uncertainty and ensuring data quality continue to be major limiting factors, and that harmonized testing protocols, standardized metadata, and approaches based on weighing evidence will be essential if these developments gain wider regulatory traction [59, 64, 65].

1.4.2 Safe(-and-Sustainable)-by-Design and the predictive tools potential

The concept of Safe-by-Design (SbD) is based on integrating safety considerations at the earliest stages of the innovation cycle, so that the properties of nanomaterials or the design of nano-enabled products can be modified before they reach later stages of development and commercialization. In practice, this means that potential hazards are considered not only after a material has been produced, but already during its design, synthesis, and intended application. SbD therefore aims to minimize adverse impacts proactively and to evaluate materials from a whole-life-cycle perspective, including production, use, and disposal [66, 67]. In this sense, SbD does not replace traditional risk assessment but rather complements it by using information about hazards and levels of exposure as a basis for setting design targets that favor safer options during the initial phases of research and development [63, 67].

Recent developments have expanded the traditional scope of SbD to include safety as an integral part alongside economic performance and environmental sustainability [66]. This broader perspective is often referred to as Safe-and-Sustainable-by-Design, reflecting the idea that a material should not only be less hazardous, but should also remain technically functional, economically realistic, and environmentally acceptable throughout its life cycle. In this context, the Anticipating Safety Issues at the Design Stage of NAno Product Development (ASINA) framework is notable for taking the classic three-pillar model of SbD further through iterative, data-driven optimization, the incorporation of life-cycle assessment, and by placing particular emphasis on the exposure conditions throughout the product's lifespan [66]. In other words, ASINA does not treat safety testing as a single final checkpoint, but as part of a repeated design-feedback process in which material characteristics, exposure scenarios, and sustainability considerations are evaluated together. These advances rely strongly on good practices in Findable, Accessible, Interoperable and Reusable (FAIR) data management as well as on the maintenance of traceability for characterization and for hazard data throughout the whole of the life cycle [66].

The need for predictive tools in this area is strong, and it is anticipated that high-throughput screening will support informed choices of materials before comprehensive toxicological work, while computational platforms are being engineered to combine a range of data types into decision-support tools aimed at the design of safer products [59, 63, 66, 68]. Such tools are particularly valuable in nanosafety because the number of possible material variants is very large, whereas

the resources required for full experimental testing are limited. Predictive models can therefore be used as early-tier support tools to prioritize materials, formulations, or exposure conditions that warrant more detailed investigation. Even so, caution is expressed in the literature: many *in silic*o predictive models for nanomaterials are considered to be at an early stage of development, and the applicability of such tools depends in large part on efforts toward standardization, validation, and the definition of clear limits of where such tools may be applied [59].

Early *in vitro* hazard assessments must be undertaken with the understanding that such studies are mainly aimed at signaling potential problems rather than fully substituting for more comprehensive higher-tier analyses [69]. These analyses refer to more complex experimental approaches, including advanced *in vitro* systems, *in vivo* studies, or integrated testing strategies that provide a broader and more physiologically relevant picture of hazard. Within the broader framework of the SbD approach, it appears that methods that more closely mimic physiological conditions may offer better predictive value for what will occur *in vivo* [69]. Nonetheless, the need for accurate deposition of dose figures, rigorous controls of potential interferences, sample preparations that are relevant to realistic exposures, and standardized methods of testing remains strong [69]. Accordingly, the practical value of predictive tools within SbD depends not only on model performance itself, but also on the quality, comparability, and biological relevance of the data used to build and evaluate such models.

1.5 Biomarkers and toxicogenomics in nanosafety

1.5.1 Biomarkers of exposure, effect, and susceptibility

Research involving workers who had been exposed to MWCNT or TiO_2-NPs has shown that early signs of inflammation, fibrosis, and oxidative stress can be detected within various biological matrices, including airway, blood, urine, and exhaled breath [70–72]. In particular, work with workers exposed to MWCNT revealed increased concentrations of several inflammatory cytokines in both sputum and serum, most notably higher levels of sputum *IL-1β, IL-4, IL-5, IL-6, IL-8,* and *TNF-a*, as well as higher levels of serum *IL-1β, IL-4,* and *TNF-a* when compared with controls [70]. Fatkhutdinova et al also reported increases in sputum levels of *KL-6* associated with

MWCNT exposure, while serum levels of *KL-6* did not differ significantly between those exposed to MWCNT and healthy workers [70].

In a larger investigation of U.S. workers that was deliberately focused on early-stage biomarker changes, a wide range of potential biomarkers were assessed in both sputum and blood, including those related to fibrosis, inflammation, oxidative stress, and cardiovascular pathology. Consistently, it was found that associations between carbon nanotubes and nanofibers (CNT/F) exposure and markers of sputum fibrosis and oxidative stress were stronger than associations with markers of sputum inflammation or cardiovascular disease [73]. Taken together, these data suggest that signals of early disease can be picked up among groups of workers exposed to CNT/F [70, 73].

The predominant focus identified in studies of workers exposed to TiO_2-NPs was on oxidative stress pathways, rather than on the establishment of specific dose biomarkers for particular materials [72]. In line with this, analyses of workers involved in the production of TiO_2-NP pigments reported increases in urinary levels of Ti as well as elevations of several oxidatively modified nucleic acids, while noting that the measured concentrations of biomarkers remained either within the range typical of the general population or below the occupational limits [71]. Urinary Ti is generally not regarded as a suitable dose marker in occupational settings, although changes over a work shift may still reflect uptake of the ultrafine fraction of TiO_2 [71].

Literature recognizes a range of factors that could modify individual responses to nanoparticles as potentially relevant [74], although direct evidence for heritable traits that influence susceptibility to nanoparticle exposure remains limited [74]. ENMs are capable of affecting phase I and phase II enzymes, components of DNA repair, and mechanisms of epigenetic control, and these findings raise the possibility that such pathways could provide useful markers of susceptibility, although at present the field of susceptibility biomarkers in nanosafety is better seen as an emerging area of research rather than as a set of tools that can be routinely applied to occupational risk assessment [74].

1.5.2 Omics-based biomarkers and toxicogenomics

Toxicogenomics (TGx) has been employed to measure global responses to ENM exposure and to link certain physicochemical properties of these materials with the induced toxicity [75]. Whole-genome expression analyses are especially powerful, as they allow broad interrogation of pathway-specific effects following exposure, and can reveal dose-dependent relationships between the exposure conditions and specific

toxicological outcomes [76]. The wealth of data generated by omics approaches often includes large lists of deregulated features, and a major analytical challenge is therefore how to separate changes that reflect normal host defense mechanisms from those that may be more useful as markers of progression towards adverse effects [75].

The field of omics-driven biomarker identification is expanding beyond analyses of messenger ribonucleic acid (mRNA), and reviews of the TGx program have highlighted that microRNAs (miRNAs) are among the most stable ribonucleic acid (RNA) species, making them promising in the creation of diagnostic tests based on body fluids for early detection of toxic insults [74, 77]. Responses in terms of DNA methylation have been reported to depend upon factors such as the type of nanomaterial, its surface chemistry, concentrations, the types of cells involved, and the exposure periods, underscoring that epigenetic changes are highly context-dependent [78].

1.5.3 Lung transcriptomic responses to engineered nanomaterials

Analysis of lung transcriptomes following exposure to various forms of TiO_2-NPs and MWCNTs has shown that high-throughput techniques are able to detect both common inflammatory responses and differences that are specific to particular materials in terms of the magnitude of their effects [75, 76, 79]. In mice subjected to inhalational exposure of surface-coated TiO_2-NPs, there was significant induction of genes associated with acute-phase responses and innate defense – including *Saa1, Saa3, C3*, fibrinogen, *S100A8*, and several chemokines – together with prominent neutrophilic inflammation [80]. The same work also reported significant changes in pulmonary miRNA expression (with 55 miRNAs altered in total) and confirmed increased levels of miR-449a, miR-1, and miR-135b, suggesting that the response studied went beyond changes at the level of mRNA alone [80]. In a wider study that involved six different forms of TiO_2-NPs, it was found that all of these particles were capable of inducing acute pulmonary inflammation through broadly similar pathways, although the intensity and the duration of the resulting transcriptional changes were dependent on specific physical properties of the particles [79]. In particular, anatase was found to be less responsive than either rutile or mixtures of rutile/anatase, while among the rutile particles, the response appeared to be influenced by factors such as crystallinity and coatings [79]. Comparisons with gene sets associated with fibrosis induced by the MWCNTs exposure indicated only modest

overlap, and the authors characterized the signals that suggested pro-fibrotic activity in the TiO_2-NPs experiments as subtle [79].

The body of transcriptomic work on MWCNTs increasingly points to common fibrogenic developmental programs among these materials [75, 76]. Comparative analyses guided by AOP across the particles NM-401, NRCWE-026, and Mitsui-7 revealed recurrent disruptions of pathways associated with the acute-phase response, chemokines, NF-kB signaling, dendritic cells, leukocyte extravasation, and TGF-beta, and found overlapping ranges of benchmark-dose confidence intervals for these effects despite marked differences in the physicochemical properties of the particles examined [76]. These data therefore indicate that different MWCNTs may act through broadly similar pathogenic mechanisms in the development of pulmonary fibrosis [76]. A subsequent meta-analysis of transcriptomes from ENMs affecting the lung distinguished early responses characterized by enrichment of genes involved in defense against infection and in wound repair from later responses that more closely matched patterns of Th2-type or adaptive immune biology [75]. Within these later time points, particular genes such as *Arg1, C3,* and *Ccl17* drew attention because they remained upregulated at 7, 28, and 56 days following exposure, and were indicated in that work as biomarkers of alternatively activated macrophages that may play a role in pulmonary fibrosis [75].

Studies attempting to use circulating blood measures as surrogates for changes seen in pulmonary gene expression have yielded inconclusive results: when mice were exposed to the same types of MWCNTs encountered in occupational contexts, their lungs displayed alterations in thousands of genes together with clear signs of inflammatory-fibrotic disease, whereas changes in whole-blood transcription were more modest and showed little overlap with those in the lung [72, 81]. At the level of functional predictions, only one broad disease or function category was common to both tissues, and no close parallels were identified concerning detailed molecular or cellular processes [81]. Nevertheless, the authors detected signals of systemic perturbation in the blood and noted that whole blood may be useful for detecting the systemic effects of inhalation injury, even if it does not faithfully replicate the lung transcriptome on a one-to-one basis. Thus, whole-blood measures could be applicable for biomonitoring purposes, although the available data do not yet support using them as direct markers for patterns of gene expression in lung tissue [72, 81].

1.6 Computational toxicology and machine learning for high-dimensional data

1.6.1 Machine learning in toxicology and nanosafety

Machine learning (ML) techniques are increasingly used in computational toxicology to complement established statistical and mechanistic approaches for analyzing complex toxicological data and developing predictive models [82]. Within the area of nanosafety, interpretable techniques such as decision trees, random forests, and extra-trees have been employed to develop predictive models of lung cytotoxicity, while other regression methods have been used to link physical characteristics such as size, shape, and surface charge with aspects of the assay itself, as key determinants of nanoparticle toxicity [82, 83]. These efforts illustrate a recurrent issue – the trade-off between predictive accuracy and the desire for unstable models. They also show how the prospects for applying ML to toxicology are being supported by hybrid platforms that integrate data on physicochemistry, toxicokinetics, bioassays, and molecular structure, although progress in the field continues to depend on high-quality, harmonized data and transparent procedures for model validation [82–84]. The increasing ubiquity of high-dimensional omics data has made computational methods essential for modern toxicological studies, since analyses of global patterns of transcripts, proteins, or metabolites routinely require preliminary stages of statistical curation [85–87]. In transcriptomics, RNA sequencing (RNA-seq) and DNA microarray provide highly quantitative, reliable measurements of gene expression patterns [87]. Downstream computational work in this field typically involves tasks such as the identification of differentially expressed genes, the mapping of those genes onto larger biological functions or signaling pathways, and comparisons of transcriptional signatures between different experimental sets. In early uses of toxicogenomics, it was found that the gene-expression signatures specific to particular chemicals were distinct enough to allow accurate classification of toxicants in blinded tests [87]. Recent work has extended these concepts to studies of heterogeneous tissues and has highlighted the utility of single-cell RNA sequencing (scRNA-seq) for teasing apart cell-specific effects that remain hidden when data are analyzed in bulk [88]. Current pipelines now routinely incorporate automated methods for cell-type annotation, as well as methods for inferring pseudotemporal trajectories that do not depend on any prior marker information [88].

The conversion of complex, high-dimensional gene-expression datasets into concise, risk-relevant metrics can be approached through several routes, among

which the most common involves fitting dose-response curves to individual genes, summarizing results at the level of groups of genes (defined as pathways), and then determining a transcriptomic point of departure (tPOD) [89, 90]. Recent studies have shown that methods based directly on the distribution of gene-specific benchmark doses can streamline this process, make fuller use of all of the available benchmark-dose data, and retain robustness even when transcriptomic data sets are small [89].

1.6.2 Integrating particle properties, exposure metrics, and transcriptomic responses

A major difficulty in the field of computational nanosafety lies in the development of analytical approaches capable of unifying disparate types of information within a single framework [76, 91, 92]. Transcriptomic work on inhaled nanomaterials illustrates why such integration is needed: analyses of whole-genome expression patterns in mouse lungs following exposure to TiO_2-NPs, MWCNTs, and carbon black have shown that inflammatory signaling pathways may be shared across different classes of engineered nanomaterials, while the degree of perturbed genes and pathways can differ according to the particular type of particle and its properties [75, 76].

The application of AOP-guided modeling to transcriptomic data has helped clarify links between levels of exposure, patterns of molecular signaling, and the emergence of adverse outcomes [76]. Another study, involving a meta-analysis of lung transcriptomes in response to ENM exposure, showed that responses to MWCNTs grouped with those seen in models of lung disease, and could be divided into an early phase that resembled bacterial or bleomycin-like insults and a later phase that was more reminiscent of Th2-type disease, whereas responses to TiO_2-NPs tended to cluster apart [75]. These findings suggest that short-term changes in gene expression might serve as useful initial-tier screens for ranking the risk that particular ENMs might cause lung damage [75].

1.6.3 Unsupervised and supervised learning

Unsupervised learning differs in that it examines data without using a predefined outcome. Principal component analysis (PCA), clustering, and related dimensionality-reduction methods are commonly used to identify patterns, similarities, and latent structure within complex datasets [93, 94]. These approaches are especially valuable in omics research, where the number of measured genes, proteins, or metabolites may greatly exceed the number of experimental observations

[94, 95]. PCA transforms correlated variables into a smaller number of components representing major sources of variation [93, 95–97], while clustering can group substances, experimental conditions, genes, or biological responses according to their similarity. Unsupervised methods consequently support data visualization, detection of unusual observations, hypothesis generation, and the construction of compact representations of high-dimensional biological responses [93, 95–97].

Supervised learning encompasses methods that learn relationships from observations for which the outcome is already known [82, 98]. Regression methods are used when the predicted endpoint is continuous, whereas classification methods assign observations to predefined categories [99, 100]. Regularized linear models, including ridge regression, are useful when descriptors are numerous or strongly correlated [101, 102]. Tree-based approaches, such as random forests, can represent nonlinear relationships and interactions between variables without requiring an explicit mathematical form to be specified beforehand [103]. In chemistry, environmental science, pharmacology, and toxicology, supervised learning is used to predict biological activity, toxicity, dose–response endpoints, biomarker activation, or hazard classes [83, 104, 105].

Quantitative structure–activity relationship (QSAR) modeling is one of the most established applications of computational methods in chemistry and toxicology [92, 106–110]. QSAR models describe mathematical relationships between a set of substance descriptors and a measurable biological or physicochemical endpoint. In conventional QSAR, these descriptors are usually derived from molecular structure. For nanomaterials, this concept has been extended to nano-QSAR, in which particle size, shape, chemical composition, crystal structure, surface area, surface charge, coating, agglomeration state, and exposure-related properties may all contribute to model predictions [111]. Such models are applied to screen insufficiently characterized materials, prioritize substances for experimental testing, support read-across, and identify properties associated with biological activity. According to OECD guidance, a scientifically reliable QSAR model should have a defined endpoint, an unambiguous algorithm, a specified applicability domain, appropriate measures of model performance, and, where possible, a mechanistic interpretation [112].

Simplex Representation of Molecular Structure (SiRMS) descriptors provide a fragment-based representation of chemical structure for QSAR and related machine-learning applications [113–115]. The approach decomposes a structure into simplexes containing between one and six atoms and characterizes them according to atom types, bonding patterns, functional groups, and physicochemical properties such as partial charge, lipophilicity, refractivity, and hydrogen-bonding capacity

[113–115]. This representation captures local chemical environments while retaining sufficient structural detail to support interpretation of relationships between particular fragments and a modeled endpoint.

2. Research Challenge, Hypothesis, Objectives

2.1 Research Challenge and Main Hypothesis

The rapid expansion of ENMs poses significant challenges for the evaluation of their safety, yet metal-and carbon-based nanomaterials are the most commonly employed across a range of industrial, biomedical, catalytic, coating, and advanced materials domains. Due to their nanoscale dimensions, high specific surface area, and diverse morphologies and surface properties, these ENMs interact with biological systems differently from bulk materials. Therefore, pulmonary exposure to ENMs continues to be regarded as a major point of concern, particularly where inhalation may constitute a significant route of entry (for example, in workplaces). Current approaches in nanotoxicology rely heavily on experimental work and on measurements of apical outcomes such as inflammatory responses, fibrosis, signs of organ-specific toxicity, and later adverse events. Although such readouts are essential for the identification of hazards, they usually provide only limited insight into the earlier, molecular events that may lead to tissue injury. Moreover, experimental safety studies are slow and costly, are constrained by the ethical imperative to minimize the use of animals, and are difficult to adapt to the growing range of ENMs, which vary in size, shape, surface chemistry, and composition.

Modern approaches for evaluating nanosafety increasingly rely on NAMs and concepts such as Safe-and-Sustainable-by-Design (SSbD), together with computational tools that promise early prediction of hazardous effects of nanomaterials. Within this shift, transcriptomics plays a valuable role, because changes at the level of gene expression can serve as early biomarkers of nanomaterial-induced perturbations – often before any clear adverse phenotypic effects become apparent. These biomarkers provide detailed, mechanistic insights into how particular properties of nanomaterials may contribute to toxicological outcomes. However, transcriptomic studies often generate large numbers of differentially expressed genes – sometimes on the order of hundreds or thousands. Attempting to model each gene in turn can therefore be computationally demanding, difficult to

interpret, and susceptible to overfitting, while truncating the dataset runs the risk of losing important, coordinated, system-wide signals.

Thus, this dissertation is based on the hypothesis that pulmonary exposure to ENMs induces coordinated transcriptomic changes that reflect perturbations of biological processes underlying their modes of action, and that these responses are related to nanomaterial physicochemical properties and exposure conditions. These changes can be represented by complementary, predictive, and biologically meaningful endpoints to characterize the overall pulmonary response and identify molecular processes relevant to hazard assessment and potency ranking.

To explore this hypothesis, the research was guided by the following key questions:

1. Can the TRI capture biologically meaningful pulmonary responses to MWCNT exposure and support potency ranking from material properties and exposure conditions?
2. Can the TRI approach be extended to TiO_2-NPs while retaining its biological relevance and predictive value?
3. Which transcriptomic biomarkers reveal biological processes and potential modes of action following TiO_2-NPs exposure, and how do their responses relate to material properties and exposure conditions?

By addressing these questions, this dissertation aims to establish a coherent properties-to-omics modeling strategy for nanosafety assessment. The proposed framework aims to bridge materials chemistry, toxicogenomics, machine learning, and regulatory toxicology by transforming complex omics datasets into interpretable, predictive endpoints relevant to pulmonary hazard assessment.

2.2 Detailed Research Hypotheses

The main hypothesis assumes that the physicochemical and exposure-related properties of engineered nanomaterials influence pulmonary toxicity by inducing reproducible transcriptomic perturbations. These perturbations can be quantified either as global transcriptomic-response patterns or as specific biomarker-level responses and modeled using machine learning. This assumption is tested through the following three detailed hypotheses.

Hypothesis 1

I hypothesize that the global transcriptomic response induced by carbon-based nanomaterials can be compressed into a single biologically meaningful TRI without

losing the essential structure of the gene-expression space. Specifically, thousands of MWCNT-responsive differentially expressed genes can be reduced to one integrated transcriptomic endpoint using principal component analysis, and this endpoint can be quantitatively predicted from MWCNT structural and exposure-related descriptors using a single validated nano-QSAR model.

Hypothesis 2

I hypothesize that the TRI is not limited to one exposed nanomaterial type, but represents a transferable framework for modeling nanomaterial-induced transcriptomic perturbation across distinct material classes. However, as different ENMs can induce transcriptomic responses with different variance structures, the optimal TRI composition must be re-optimized for each dataset and material class to balance biological information retention, model robustness, and external predictivity.

Hypothesis 3

I hypothesize that supervised ML modeling can identify transcriptomic biomarkers quantitatively associated with nanomaterial-induced pulmonary pathology and predict their activation using physicochemical properties and exposure-related conditions. In particular, TiO_2-NP properties such as particle size, deposited surface area, surface charge, concentration, and post-exposure time can be linked to gene-expression changes associated with pulmonary inflammation and fibrosis, thereby supporting early hazard prediction and NAM-compatible nanosafety assessment.

2.3 Detailed Research Objectives

The research hypotheses were tested by achieving the following objectives:

Objective 1.

To develop and validate an ML-based framework for compressing the high-dimensional transcriptomic response induced by MWCNTs into a single TRI. This objective aims to determine whether thousands of differentially expressed genes in mouse lung tissue following exposure to a set of MWCNTs differ in their fold-change patterns and whether these differences can be represented by one biologically meaningful transcriptomic endpoint and predicted from MWCNT physicochemical and structural properties using a single nano-QSAR model. The objective includes transcriptomic data preprocessing, differentially expressed gene (DEG) selection,

dimensionality reduction using principal component analysis, TRI calculation, regression model development, rigorous validation, and applicability domain assessment.

Objective 2.

To evaluate whether the TRI framework can be transferred from carbon-based nanomaterials to metal oxide nanoparticles and optimized for TiO_2-NP-induced pulmonary transcriptomic responses. This objective aims to determine whether TRI is a material-specific or broadly applicable omics endpoint by constructing and comparing different TRI variants for TiO_2-NPs. The objective includes identifying the optimal principal component composition of TRI, linking the optimized TRI to TiO_2-NP physicochemical and exposure-related descriptors, and validating whether the resulting model provides a robust, interpretable, and predictive bridge between nanoparticle properties and global transcriptomic response.

Objective 3.

To identify and predict transcriptomic biomarkers of nanomaterial-induced pulmonary pathology using supervised ML. This objective aims to determine whether individual gene-expression responses associated with inflammation, fibrosis, oxidative stress, and related pulmonary outcomes can be quantitatively linked to TiO_2-NP properties and exposure conditions. The objective includes developing and validating biomarker-level ML models, identifying key physicochemical drivers of gene dysregulation, and evaluating whether these models can support early hazard identification, mechanistic interpretation, and SbD prioritization of nanomaterials.

3. Results

3.1 Linking MWCNTs properties to pulmonary transcriptomic responses through the Transcriptomic Response Index

3.1.1. Introduction

The toxicological concerns surrounding MWCNTs are largely due to their resemblance to asbestos fibers as well as their marked persistence within lung tissue [116]. A number of investigations have documented significant hazardous effects of exposure to MWCNTs – most notably within the respiratory tract, where sustained inflammatory responses, progression to fibrosis, and increased incidence of lung cancer have been reported [117, 118]. Beyond pulmonary effects, experimental studies have implicated MWCNT exposure in a range of developmental and reproductive outcomes, including congenital malformations, embryonic loss, behavioral alterations in offspring, and delayed parturition, despite the growing understanding of the toxicological profile of these materials [118]. Specific molecular pathways that give rise to such reactions remain unclear, and toxicogenomics has emerged as an important tool for probing the effects of MWCNTs on the lungs. Recent work has uncovered patterns of gene expression associated with fibrosis and inflammation that are reminiscent of those seen in response to well-known fibrogenic agents as well as to infectious insults [75, 119–122].

The rapid expansion of publicly available transcriptomic datasets has opened new opportunities for assessing nanosafety in a mechanistically grounded manner, and earlier work has indicated that quantitative relationships can be drawn between the physicochemical characteristics of MWCNTs and changes in transcriptomic profiles at the whole-pathway level [120, 123]. Concurrently, the approach of perturbation-based machine learning (PTML) has attracted interest as a complementary method for predicting toxicity [105, 124–126]. One of the main challenges facing predictive toxicology driven by transcriptomic data is the high dimensionality of gene-expression measurements: a single experiment can yield thousands of DEGs, and this makes traditional approaches that model each gene separately both computationally onerous and difficult to interpret.

In order to overcome the limitations noted previously, this work introduced the TRI as a novel, systems-level readout of pulmonary responses to MWCNTs. The aim of TRI was to compress the broad changes seen at the level of the transcriptome into a single biologically interpretable metric, thus preserving the essential structure of the transcriptomic response while rendering it tractable to predictive analyses. The approach combined unsupervised dimensionality reduction via PCA with supervised modeling using ridge regression to relate TRI behavior to the structural descriptors of MWCNTs, as characterized by the SiRMS descriptor set [115, 127–130]. Unlike earlier studies that focused on individual markers or on single transcriptomic pathways [120, 123], this method permits the prediction of the entire space of transcriptomic responses through the use of a single quantitative model. In order to maximize robustness, reproducibility, and regulatory acceptability, the full workflow was carried out within the ChemBioML Platform and was evaluated in accordance with OECD-compatible standards of validation [131]. As a result, the present work represents progress in the integration of transcriptomic data and ML approaches in the field of nanotoxicology and provides a compact, interpretable, and computationally efficient method for predicting transcriptomic responses to inhaled MWCNTs.

3.1.2 Materials and Methods

This work built upon previously published data on 5–7-week-old mice subjected to intratracheal exposures to nine distinct types of MWCNTs [41, 132–136]. These materials varied with respect to length, diameter, aspect ratio, and the nature of their surface functional groups (pristine, –OH-, –COOH-, and –NH2-functionalized types). Each nanotube was examined under two or three different types of exposure, for a total of 20 experimental conditions, and details of the range of materials and conditions evaluated are summarised in Table 1 [41, 132–136]. In order to provide a basis for QSAR analyses, the physicochemical characterization of the set of MWCNTs was carried out with the use of the SiRMS approach [115, 127–130]. In the present work, a concentration-dependent weighting scheme was also applied in order to take account of dose-dependent differences in biological response. The complete set of descriptors was generated with the help of the open-source SiRMS Python toolkit, and this full set of values was then used as the basis, as the space of independent variables, for the construction of nano-QSAR models.

In order to assemble a usable set of transcriptomic data, DEGs drawn from the primary reports were subjected to filtering on the basis of a nominal significance cut-off of $p \leq 0.05$ together with the requirement that their expression levels differed by

at least a factor of 1.5 [41, 80, 133, 137–139] A gene was retained if it met both criteria in at least one tested combination of MWCNT type and dose; it was not required to be differentially expressed under every exposure condition. This filtering produced a curated set of 5,167 DEGs that formed the basis for all subsequent analyses. The criteria for selecting this cut-off were guided by a desire to stay in line with the analytical approaches used in the original studies, whilst at the same time preserving enough biological detail to support modeling efforts [132, 133], and the dimensionality of this large list of DEGs was subsequently reduced by application of PCA to the matrix of 5167 genes [93, 96]. This approach re-expresses correlated measures of gene expression in terms of mutually orthogonal principal components, each of which accounts for some share of the overall variability in the data set. The TRI was then derived as a weighted sum of those principal components that were used, the weights being proportional to the amount of variance explained by each component (formula 1). Thus, TRI provided a single continuous measure that captured the overall transcriptional response to MWCNT exposure [93, 96].

$$TRI = PC1 * Variance_1 + PC2 * Variance_2 + \cdots + PCn * Variance_n \quad (1)$$

A supervised ML approach was employed to correlate the structural attributes of MWCNTs with the TRI endpoint, combining ridge regression with a feature-selection strategy based on a genetic algorithm (GA) [101, 102, 140], The complete set of experimental conditions (20 in total, Table 1) was partitioned for the purposes of model development, with four individual cases randomly allocated to an external test set and the remaining sixteen cases used for model training as well as for internal optimization steps. All stages of this work – preprocessing, feature selection, and tuning of hyperparameters – were carried out within the ChemBioML platform [131], and the resulting models were judged on three criteria – goodness-of-fit, robustness, and the ability to make predictions on unseen data – using standard metrics such as determination coefficient (R^2) and the root-mean-square error of calibration ($RMSE_C$) for goodness-of-fit, and by means of leave-one-out cross-validation measures such as determination coefficient for cross-validation (Q^2_{CV}) and root-mean-square error of calibration ($RMSE_{CV}$) for assessments of robustness, while external predictivity was gauged through use of the external validation coefficient (Q^2) together with root-mean-square error of external validation ($RMSE_{EXT}$). In addition, the Applicability Domain (AD) was explored by means of the leverage method and visualized through a Williams plot [107, 141], which enabled us to pinpoint particular structurally influential outliers and to ensure that the final model could be interpreted reliably within a well-defined and trustworthy space of descriptors.

Table 1. MWCNTs types, their physicochemical characteristics, and exposure conditions

#	MWCNT Symbol	Concentration (μg/mouse)	Length (nm)	Diameter (nm)	Functionalization
NP_1	NM-401	18	4050 (± 2400)	67 (± 26.2)	Pristine
NP_2	NM-401	54	4050 (± 2400)	67 (± 26.2)	Pristine
NP_3	NM-401	162	4050 (± 2400)	67 (± 26.2)	Pristine
NP_4	NRCWE-26	18	850 (± 457)	11 (± 4.5)	Pristine
NP_5	NRCWE-26	54	850 (± 457)	11 (± 4.5)	Pristine
NP_6	NRCWE-26	162	850 (± 457)	11 (± 4.5)	Pristine
NP_7	NRCWE-043	18	771.3 (± 3471)	26.73 (± 6.88)	Pristine
NP_8	NRCWE-043	54	771.3 (± 3471)	26.73 (± 6.88)	Pristine
NP_9	NRCWE-044	18	1330(± 2454)	32.55 (± 14.44)	-OH
NP_10	NRCWE-044	54	1330(± 2454)	32.55 (± 14.44)	-OH
NP_11	NRCWE-045	18	1553(± 2954)	28.07(± 13.85)	-COOH
NP_12	NRCWE-045	54	1553(± 2954)	28.07(± 13.85)	-COOH
NP_13	NRCWE-046	18	717.2 (± 1214)	17.22 (± 5.77)	Pristine
NP_14	NRCWE-046	54	717.2 (± 1214)	17.22 (± 5.77)	Pristine
NP_15	NRCWE-047	18	532.5 (± 591.9)	12.96 (± 4.44)	-OH
NP_16	NRCWE-047	54	532.5 (± 591.9)	12.96 (± 4.44)	-OH
NP_17	NRCWE-048	18	1604(± 5609)	15.08 (± 4.69)	-COOH
NP_18	NRCWE-048	54	1604(± 5609)	15.08 (± 4.69)	-COOH
NP_19	NRCWE-049	18	731.1 (± 1473)	13.85 (± 6.09)	$-NH_2$
NP_20	NRCWE-049	54	731.1 (± 1473)	13.85 (± 6.09)	$-NH_2$

3.1.3 Results and Discussion

The dataset analyzed in this work comprised 5,167 DEGs that showed significant changes under at least one of the conditions of MWCNT exposure examined and was constructed by unifying data drawn from previously published studies of lung transcriptomes [41, 80, 133, 137–139]. As the task of modeling each of the thousands of individual molecular endpoints separately would be computationally expensive, I applied dimensionality reduction by means of PCA. The first nineteen principal components (PC1–PC19) together accounted for roughly 99.9 % of the total variance in the DEG data matrix, showing that these components captured the entirety of the structure of the global response, essentially, whereas a subspace defined only by PC1 and PC2 retained about 43 % of the variance. Together, these results show that it is possible to describe the effects of MWCNT exposure within a reduced latent space without incurring substantial loss of information, provided that a sufficient number of principal components (PCs) are retained. The two-dimensional space defined by PC1 and PC2 showed that most of the different experimental conditions clustered in broadly similar areas – suggestive of a large degree of commonality in the way that the lung responded at the level of gene expression to the range of MWCNTs studied – while the overall spread of points also suggested that transcriptional differences were affected by both the structural properties of the nanotubes as well as by their surface functionalization (figure 1).

The central contribution of this study is the creation of the TRI, a metric capable of summarizing the entirety of transcriptomic perturbations caused by exposure to MWCNTs within a single, biologically interpretable metric, instead of constructing separate predictive models for each of thousands of DEGs. By formulating the TRI as a weighted sum of PCs, the approach was able to retain both the most prominent and more subtle patterns of gene-expression change and simplify the modeling task – an optimization that made it possible to train a single nano-QSAR model, and did so without loss of interpretability.

Thus, TRI does more than simply perform a technical reduction of data size; it provides a biologically meaningful, system-wide endpoint of transcriptomic patterns caused by inhaled MWCNTs. The 20 experimental conditions comprised nine MWCNT types tested at 18 and 54 µg per mouse, with NM-401 and NRCWE-26 additionally tested at 162 µg per mouse. Importantly, the PC1–PC2 distribution shown in Figure 1 and the subsequent TRI calculation were derived solely from the fold-change matrix of 5,167 DEGs across these conditions. In order to numerically characterize the various properties of the nanomaterials, I employed SiRMS descriptors weighted by concentration. As a result, the set of descriptors was able to capture not only the inherent structural features of the nanotubes, but also the quantitative exposure circumstances that are relevant to their biological effects.

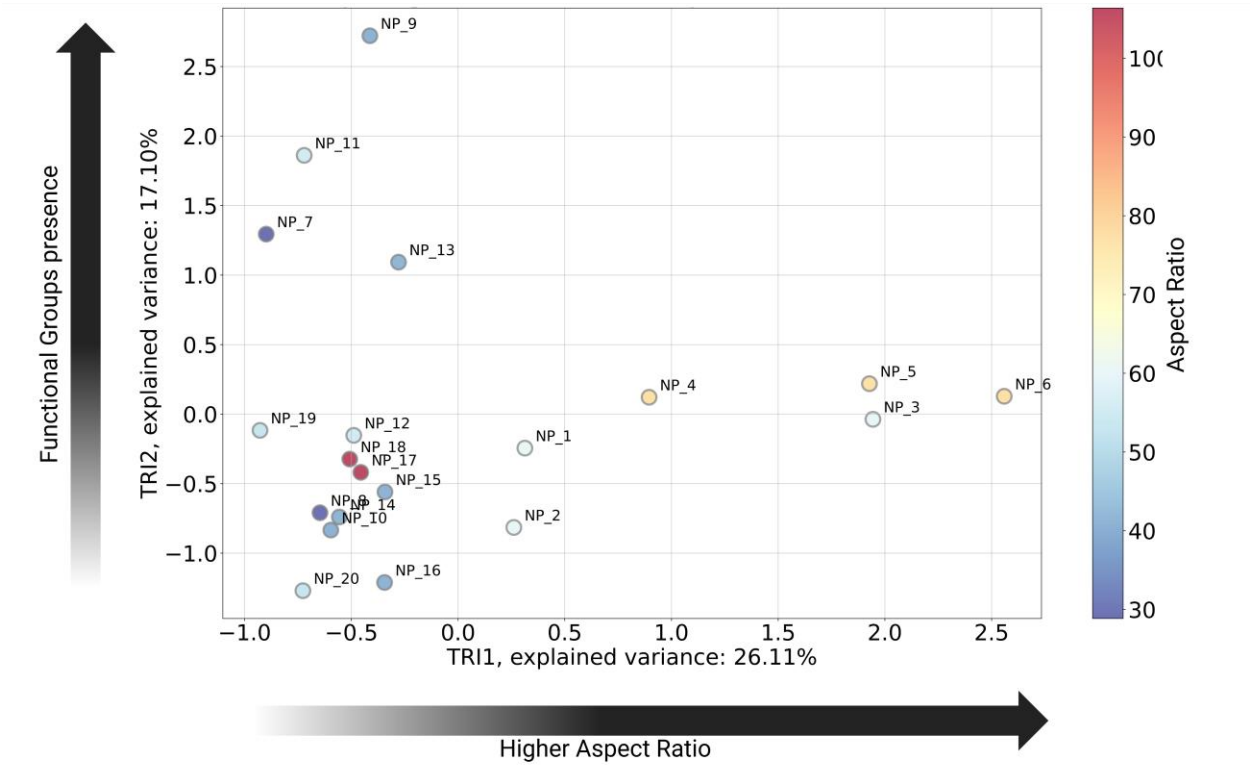


Figure 1. Two-dimensional transcriptomic space, represented as linemap PC1:PC2.

A quantitative relationship linking the physicochemical properties of MWCNT to changes in transcriptomic response was pursued through the construction of a ridge-regression nano-QSAR framework that incorporated a genetic algorithm for feature selection. The model was trained on 16 cases and subsequently tested on an independent set of four cases, with the resulting statistics R^2 = 0.83, $RMSE_C$ = 0.15, Q^2_{CV} = 0.80, $RMSE_{CV}$ = 0.16, and Q^2 = 0.78. The results suggest that the model achieved a favorable balance between goodness-of-fit, robustness, and external predictivity, with only a modest difference between R^2 and Q^2 (indicating a lack of substantial overfitting). Comparison of predicted versus measured values confirmed that the model was able to reproduce the main trends of the data with reasonable accuracy, an important result indicating that the complex, transcriptome-wide response of the lung to MWCNT exposure can be predicted from simple descriptors of the nanomaterials by means of a single quantitative model (Figure 2a). The AD of the final model was explored by means of the leverage method and visualised through a Williams plot, and all cases fell within the generally acceptable range of ±3 for absolute standardized residuals, indicating a lack of serious outliers; one case (NP_3) had comparatively high leverage (h = 0.85), reflecting its marked structural distinctiveness from the other nanotubes – and, although predictions for this point were not inaccurate, its location in the space of leverage suggests that further studies of structurally similar materials could be useful for enlarging the range of descriptors captured by the model (Figure 2b). Taken together, the analyses of the AD support the view that the final model is reliable within the structural diversity represented by the present data set.

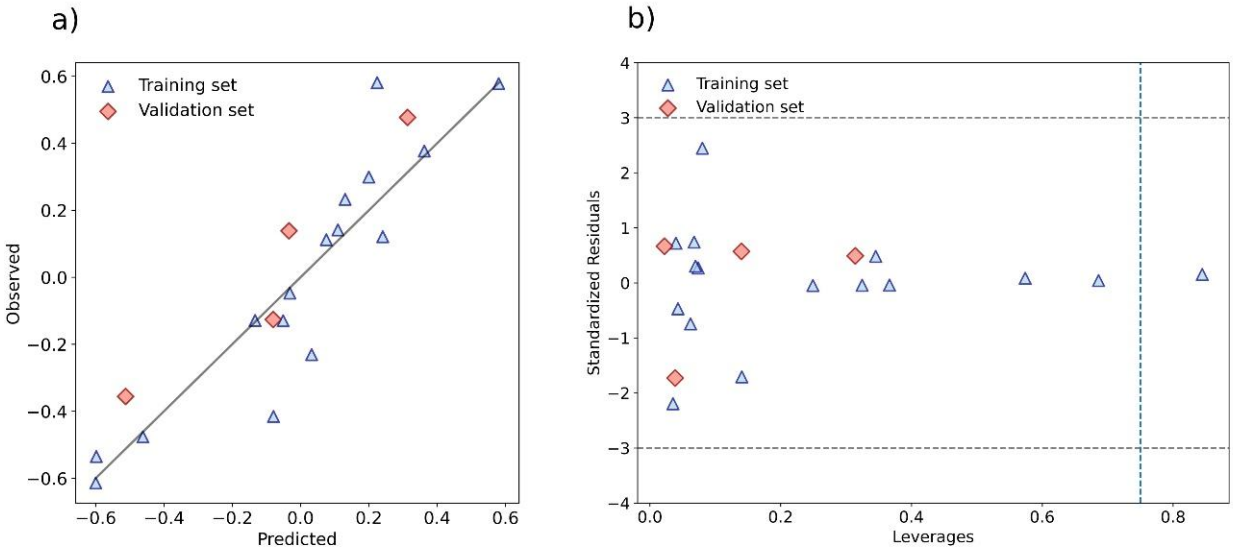


Figure 2. a) Observed-predicted plot presenting predictions' accuracy in two-dimensional space; b) Williams plot presenting the model AD.

The resulting nano-QSAR model highlighted charge-weighted descriptors as making the most substantial contributions to predictions of TRI (formula 2), a finding that aligns with the well-established role of surface charge in modulating the interactions of carbon nanotubes with biological tissues [142–145].

$$TRI = 0.53 * CHARGE_{A-A-B-B-A-A} - 0.57 * CHARGE_{B-B-B-C} + 0.37 * REFRACTIVITY_{B.C} - 0.46 * CHARGE_{A.A} \quad (2)$$

In particular, the data suggest that distinct charged domains within structures of MWCNTs may have opposing effects upon the transcriptomic response – an outcome that is compatible with the known non-uniform electronic charge distribution of these materials – and suggest that the effects of surface functionalization may be largely indirect, acting through changes in local charge patterns and in refractivity rather than through the simple categorical nature of the functional groups themselves. From a toxicological point of view, the importance of charge-related descriptors points to a possible mechanistic link between the electronic properties of MWCNTs and the broad-scale changes in gene expression seen in the lung; charged or polar regions of the nanotubes may therefore have roles in influencing uptake by cells, in their interactions with membranes and with proteins, in their potential to generate oxidative stress, or in the triggering of defensive signaling pathways [142–149]. Taken together, the results support the notion that physicochemical characteristics of nanomaterials can be translated into biologically meaningful changes at the omics level, and therefore strengthen the case for interpreting results from nano-QSAR studies in terms of underlying mechanisms in the field of nanotoxicology.

The results establish TRI as a practical systems-level endpoint for translating complex MWCNT-induced transcriptomic perturbations into a form that is suitable for predictive nanotoxicology. By integrating gene-expression changes across the lung transcriptome into a single interpretable index, the approach reduces the analytical burden associated with thousands of individual DEGs while preserving the biological structure of the response. TRI should therefore be interpreted as a single, dimensionless coordinate summarizing the overall pattern of transcriptomic perturbation under each exposure condition. A higher TRI reflects a coordinated shift toward increased expression of genes with positive TRI coefficients and decreased expression of genes with negative coefficients. This has an important methodological impact, as it enables transcriptomic data to be incorporated into nano-QSAR modeling in a more scalable and interpretable manner, bridging omics-based

toxicology with physicochemical characterization of nanomaterials. The framework also strengthens the mechanistic relevance of predictive modeling by linking the integrated pulmonary response to structural features of MWCNTs, particularly descriptors related to charge distribution and refractivity. In this way, TRI provides not only a compact measure of transcriptomic disturbance but also a basis for comparing nanomaterials, prioritizing further testing, and supporting the development of more efficient, mechanism-informed approaches for nanosafety assessment.

3.2 Transcriptomic Response Index across Titanium-based ENMs

3.2.1. Introduction

The use of TiO_2-NPs is widespread across many commercial and industrial sectors – ranging from coatings and paints to catalysts, cosmetics, and food products – indicating that the risk of inhalation exposure to these particles in real-world settings is non-trivial [41, 79, 150]. Experimental work has shown that when such particles are inhaled, they can trigger acute neutrophilic inflammation, as well as long-term retention of the particles, and, at high concentrations within the lung, may provoke maladaptive reparative responses that are linked to fibrotic changes and, in certain circumstances, to tumor development [41, 79, 150]. These findings have, in turn, stimulated greater interest in approaches that are grounded in mechanistic understanding for the assessment of the respiratory risks posed by TiO_2-NPs as well as other ENMs.

The field of toxicogenomics is actively used for extending traditional toxicological assessments, in part because transcriptomic data are able to record important molecular changes in response to toxic insults and thus serve as highly sensitive indicators of early biological disruption [151]. Within the broader field of NAMs, omics data – especially those derived from studies of gene expression – are increasingly regarded as useful for identifying potential hazards, clarifying mechanisms of action, ranking relative potencies, and establishing transcriptomics-based points of departure (PoDs) [112, 152, 153]. Nonetheless, applying transcriptomic methods to the field of nanosafety is challenging, since the gene expression signatures induced by TiO_2-NPs are affected by a range of factors such as particle size, surface area, the specific properties of their surfaces, as well as by the metrics used for dose and by the timing of measurements after exposure, and technical problems such as batch

effects and the heterogeneity of data sets from different studies can make it difficult to draw firm conclusions from changes at the level of individual genes [154–158]. A particular problem is the high dimensionality of transcriptomic data – a single experiment can generate hundreds or thousands of DEGs in ways that are hard to synthesize into a single, robust, and comprehensible model [95, 120, 123, 159–164] – which represents a major challenge for predictive efforts in this area.

The limitations of previous approaches have prompted the suggestion of the TRI as a systems-level measure of transcriptomic variation, compressible into a single continuous variable [165]. Although its applicability had been demonstrated for carbon-based nanomaterials, a systematic assessment of the most effective ways to construct TRI in the context of TiO_2-NPs had not yet been undertaken. Accordingly, the present work set out to re-optimize the use of TRI with regard to pulmonary reactions to TiO_2-NPs, and to develop a method based on validation for choosing which combinations of PCs should define the index, in order to test whether a concise, understandable, and externally validated form of TRI might be useful as a NAM-ready read-out for global changes in gene expression following exposure to TiO_2-NPs [154, 165].

3.2.2 Materials and Methods

Using publicly available datasets of mouse lung transcription responses following intratracheal delivery of TiO_2-NPs [41, 134–136, 154, 166]. I examined a set of five different TiO_2-NPs, each at three levels of dose (18, 54, and 162 µg per animal) as well as at two time points following exposure (days 1 and 28), thereby generating a total of thirty distinct experimental conditions. The range of nanoparticles considered encompassed several key physicochemical and exposure-related traits – among them surface charge, primary particle size, Brunauer–Emmett–Teller (BET) specific surface area, and deposited surface area – and detailed descriptions of the properties of the particles studied as well as of the experimental set-up are given in Table 2.

Gene expression datasets were derived from lung tissues of 5–7-week-old female C57BL/6 mice under the conditions standardized in the published works [21, 41, 134–136, 154, 166]. Intratracheal instillation was used specifically in order to achieve comparable deposited doses of materials across the different experimental groups. For the purposes of transcriptomic comparison, previously normalized microarray data were analyzed, with each group being benchmarked against its contemporaneous control, and DEGs were called on the basis of a false discovery rate (FDR)-adjusted significance threshold of $p \leq 0.05$ together with an absolute change

in expression of at least 1.5-fold [80, 132, 133, 137–139]. This filtering step produced a list of 621 DEGs that formed the starting point for all further work involving dimensionality reduction and the application of supervised ML methods. In order to reduce dimensionality, I therefore performed PCA on the matrix of the 621 DEGs [93, 95–97], making use of the fact that PCA converts correlated measures of gene expression into mutually orthogonal principal components – each of which can be seen as an independent direction of change within the space of transcriptomic phenotypes – and, on this basis, I defined the TRI as a weighted sum of these PCs, with the weighting of each component being proportional to the amount of variance it explains (formula 1). In order to study the effect of the particular make-up of the index on predictive performance, a range of different versions of TRI were generated, from those based only on PC1 + PC2 (named TRI2) up to those incorporating contributions from PC1 through to PC29 (named TRI29).

In order to relate the TRI to characteristics of TiO_2-NPs as well as to the conditions under which these particles were encountered, supervised ML models were constructed combining ridge regression with a GA-based approach to feature selection [101, 102, 140, 165]. The starting set of predictors comprised a range of experimental and physicochemical parameters measured for the TiO_2-NP samples – such as concentrations, particle sizes, surface charges, BET surface areas, deposited surface areas, and times following exposure – and, in order to allow for possible nonlinear effects among these factors, second-degree polynomial features were generated from this initial set. For the purposes of model building, the 30 cases were ranked according to their TRI values and, beginning with the third case, every fifth entry was allocated to an external test set, resulting in a training set of 24 cases and a held-out set of six cases that were broadly representative of the full range of responses. The test set was fixed prior to optimization of the models and did not influence the choice of predictors, the constitution of the TRI, or the setting of hyperparameters. All work on model building and on subsequent optimization was carried out on the ChemBioML Platform.

Table 2. TiO2-NP types, their physicochemical characteristics, and exposure conditions

TiO_2-NPs_IDs	NPs	Post-exposure duration [days]	Concentration [µg/animal]	Charge*	Size [nm]	BET [m^2/g]	Deposited surface area** [cm^2]
NP_1	NRCWE-001	1	18	0	10	99	17.8
NP_2	NRCWE-001	1	54	0	10	99	53.5
NP_3	NRCWE-001	1	162	0	10	99	160.4
NP_4	NRCWE-002	1	18	1	10	84	15.1
NP_5	NRCWE-002	1	54	1	10	84	45.4
NP_6	NRCWE-002	1	162	1	10	84	136.1
NP_7	NRCWE-001	28	18	0	10	99	17.8
NP_8	NRCWE-001	28	54	0	10	99	53.5
NP_9	NRCWE-001	28	162	0	10	99	160.4
NP_10	NRCWE-002	28	18	1	10	84	15.1
NP_11	NRCWE-002	28	54	1	10	84	45.4
NP_12	NRCWE-002	28	162	1	10	84	136.1
NP_13	NRCWE-030	1	18	0	10.5	139.1	25.0
NP_14	NRCWE-025	1	18	0	38	28.2	5.1
NP_15	NRCWE-030	1	54	0	10.5	139.1	75.1
NP_16	NRCWE-025	1	54	0	38	28.2	15.2
NP_17	NRCWE-030	1	162	0	10.5	139.1	225.3
NP_18	NRCWE-025	1	162	0	38	28.2	45.7
NP_19	NRCWE-030	28	18	0	10.5	139.1	25.0
NP_20	NRCWE-025	28	18	0	38	28.2	5.1
NP_21	NRCWE-030	28	54	0	10.5	139.1	75.1
NP_22	NRCWE-025	28	54	0	38	28.2	15.2
NP_23	NRCWE-030	28	162	0	10.5	139.1	225.3
NP_24	NRCWE-025	28	162	0	38	28.2	45.7
NP_25	UV-Titan L181	1	18	0	20.6	107.7	19.4
NP_26	UV-Titan L181	1	54	0	20.6	107.7	58.2
NP_27	UV-Titan L181	1	162	0	20.6	107.7	174.5
NP_28	UV-Titan L181	28	18	0	20.6	107.7	19.4
NP_29	UV-Titan L181	28	54	0	20.6	107.7	58.2
NP_30	UV-Titan L181	28	162	0	20.6	107.7	174.5

* [0 - no charge; 1 - positive charge(contains positively charged amino groups on the surface2)]; ** calculated based on specific surface area, size and concentration

The quality of the resulting models was judged in terms of goodness-of-fit, robustness, and the ability to make predictions on new data: goodness-of-fit was appraised through use of the R^2 as well as the $RMSE_C$; robustness was appraised by means of leave-one-out cross-validation, through the use of the Q^2_{CV} and of the $RMSE_{CV}$; the quality of external prediction was judged through use of the external validation Q^2 as well as the $RMSE_{EXT}$. The AD was explored by means of the leverage method and illustrated with a Williams plot, which permitted the detection of points of structural importance as well as of possible outliers [107, 130, 141]. The same validation protocol was applied to all versions of the TRI in order to identify the configuration that gave the best trade-off between retained transcriptomic signal and predictive accuracy.

3.2.3 Results and Discussion

The goal of the work was to refine the use of the TRI as a concise measure of pulmonary effects resulting from exposure to TiO_2-NPs, and to do so on the basis of previously experimentally measured transcriptomic datasets from mouse lungs subjected to TiO_2-NP instillations. By applying cut-offs of FDR-adjusted p-values of ≤0.05 together with an absolute fold change of at least 1.5, a total of 621 DEGs were selected for further work. This group of significantly affected DEGs was selected to represent the core set of transcriptomic changes triggered by TiO_2-NP exposure and served as the basis for subsequent dimensionality-reduction procedures, the construction of the TRI, and the training of supervised machine-learning models. Due to high dimensionality inherent to transcriptomic data – where each DEG might be regarded as an independent endpoint, a situation that would entail the need for many (on the order of hundreds) separate predictive models and would tend to fragment the biological interpretation of the data – I reduced the set of 621 DEGs into a single composite measure, the TRI, defined through PCA. This approach allowed the whole of the global transcriptomic response to be summarised as a single continuous variable whilst still allowing individual genes to be traced back to their influence on the composite through the use of PC loadings. Several versions of TRI were produced, incorporating progressively larger numbers of principal components (from TRI2 up through TRI29). For example, TRI2, based on the first two principal components (PC1 and PC2, figure 3), accounted for roughly 44 % of the overall variance in the transcriptomes, whereas TRI29 was able to reconstruct virtually all (99.9 %) of the full range of transcriptomic variation. Consequently, the complete set of models up to TRI29 was able to give a near-total reconstruction of the matrix of differentially expressed genes, whereas the much smaller TRI2 retained a highly compact view of the most important axes of variation. Importantly, the authors did not presume that the form of the TRI that explained the largest amount of variance would automatically be the best predictor of interest, and therefore optimized the make-up of the various TRI variants by balancing the amount of variance explained against the performance of the supervised predictive models.

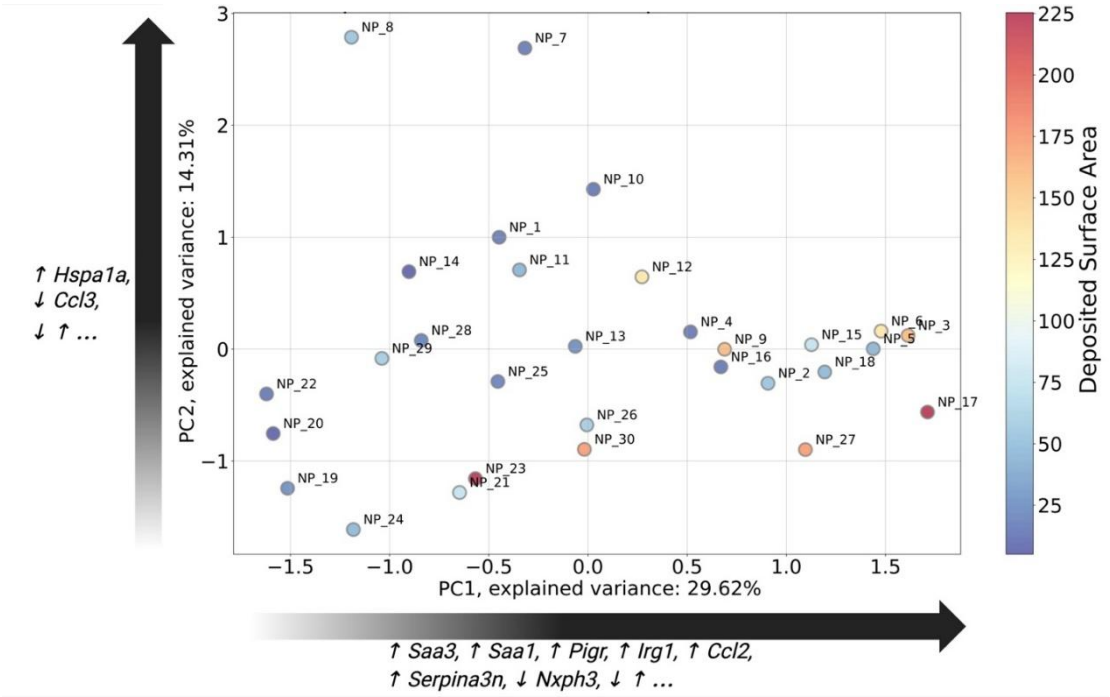


Figure 3. PC1-PC2 scatter plot, representing TRI2, Dose expressed as Deposited Surface Area.

Analyses of the projection of the experimental conditions of TiO_2-NPs exposures onto the first two PCs suggested that the resulting transcriptomic data occupy a broadly continuous space rather than being neatly segregated into discrete clusters: PC1 accounted for 29.62 % of the total variance and PC2 for 14.31 %. Along the axis of PC1 there was evidence of a gradient that correlated partially with deposited surface area – a finding that underscores the importance of dose in shaping the global transcriptional response to the particles – although a number of high-dose conditions did not conform to a simple dose-dependent trajectory, suggesting that additional factors such as particle size, charge and the length of time after exposure may also play a role in modulating the observed effects. Taken together, these data support the view that changes in gene expression induced by TiO_2-NPs cannot be attributed to dose alone. Examination of the loadings for PC1 and PC2 provided insights into the nature of the transcriptomic responses to the TiO_2-NPs instillation: among the genes contributing to PC1 (such as *Saa3, Saa1* and *Serpina3n*) many have known links to the acute phase response and to inflammation [75, 80, 167], and the activity of these genes appears to be closely tied to systemic inflammatory signaling – of note for risks to the lungs and to the cardiovascular system [168, 169]. Other contributors to PC1, such as *Pigr, Irg1* and *Ccl2*, also support the idea that the dominant axis of variation reflected inflammatory and immune responses to exposure to TiO_2-NPs, whereas PC2 appeared to capture a separate biological signal (involving, for example, genes such

as *Ccl3*, which is linked to lung fibrosis, and *Hspa1a*, which is associated with stress and with cardiovascular pathology [170, 171]).

In order to identify the most suitable formulation of the TRI for characterizing lung reactions to TiO_2-NPs instillation, I trained predictive models for each of the TRI variants ranging from TRI2 through to TRI29, using cross-validation as the primary benchmark for model usefulness – quantified in terms of the cross-validated predictivity metric Q^2_{CV} – while also considering cumulative explained variance as a means of gauging how much of the original transcriptomic signal was retained. Such a validation-driven approach was required because, although the incorporation of additional PCs can broaden the scope of transcriptomic capture, it can also add noise, reduce the generalisability of the findings, and increase the risk of overfitting; therefore, it was important to guard against these problems in our analyses. Comparison of the different TRI versions showed that TRI2 offered the best compromise between predictive ability and informational richness, while noting that, despite the fact that TRI29 was able to reconstruct almost the entirety of the transcriptomic space, its greater dimensionality did not confer any advantage in terms of model robustness – in fact, I found that, as additional components beyond the first two (i.e. beyond PC1 and PC2) were added, explained variance tended to rise but at the cost of reduced performance on cross-validated prediction (Figure 4). These findings show that the PCs that explain the greatest amounts of variance are not necessarily the same ones that are most useful for predicting outcomes of interest, an observation that is noteworthy given that, in a previous study on MWCNTs, TRI19 had been singled out as the most predictive index [165]. The fact that I found differences between TRI2 (for TiO_2-NPs) and TRI19 (for MWCNTs) suggests that the choice of TRI – and its optimization – is dependent on both the particular material under study and on the characteristics of the available data. Exposure to TiO_2-NPs appears to provoke a relatively focused, compact transcriptional response dominated by pathways of immune activation, acute phase signaling, and oxidative stress, in contrast to the effects of MWCNTs, which appear to drive a wider, longer-lasting program of inflammation and fibrosis [172–176], and suggest that the concept of TRI as a sort of fixed endpoint with a fixed number of principal components is inappropriate: rather, the constituent PCs of TRI should be re-optimized whenever a new class of nanomaterials, or new data sets, are examined.

Following the identification of TRI2 as the most suitable endpoint, we constructed a ridge-regression model that incorporated a GA-driven feature selection, in order to relate this index to physicochemical properties of TiO_2-NPs as well as to contextual factors of exposure. The resulting predictor demonstrated strong fit,

robustness, and good performance both internally and on the external validation set, achieving values of R^2 = 0.79, $RMSE_C$ = 0.15, Q^2_{CV} = 0.70, $RMSE_{CV}$ = 0.18, Q^2 = 0.79, and RMSEext = 0.15. Because R^2 and Q^2 were equal – and their difference did not exceed the generally accepted tolerance of 0.3 [111, 177] – this equality is taken as evidence that the model was not overfit, and the comparison of experimental versus predicted values on the observed–predicted plot showed a good degree of agreement, supporting the conclusion that the optimized TRI2 can be predicted quantitatively from descriptions of TiO_2-NPs together with relevant exposure circumstances (Figure 5a). The AD of the final model was examined with a leverage-based method and illustrated graphically with a Williams plot, and all cases fell within the accepted range of ±3 for standardized residuals, indicating a lack of problematic outliers, whilst none of the points lay outside the domain defined by leverage criteria – suggesting that predictions were made within the range of physicochemical and exposure conditions represented in the training set, and giving support to the view that the model is internally consistent (Figure 5b).

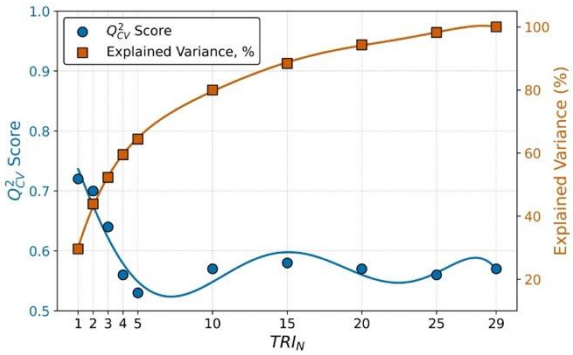


Figure 4. Plot presenting the dependency of Q^2_{CV} (higher is better) and explained variance (higher is better) of TRI based on the number of included PCs.

The final predictive model produced by the ChemBioML Platform (formula 3) identified that the optimized response of TRI2 was largely dependent on BET-specific surface area, primary particle size, the concentration of nanoparticles to which mice were exposed, and the time elapsed following removal of those particles. These factors are all mechanistically plausible and broadly concordant with already known mechanisms of pulmonary toxicity induced by TiO_2-NPs.

$$TRI_2 = 0.13 \times Concentration - 0.195 \times BET^2 - 0.071 \times Time \times Concentration - 0.167 \times Time \times Size - 0.201 \times Size \quad (3)$$

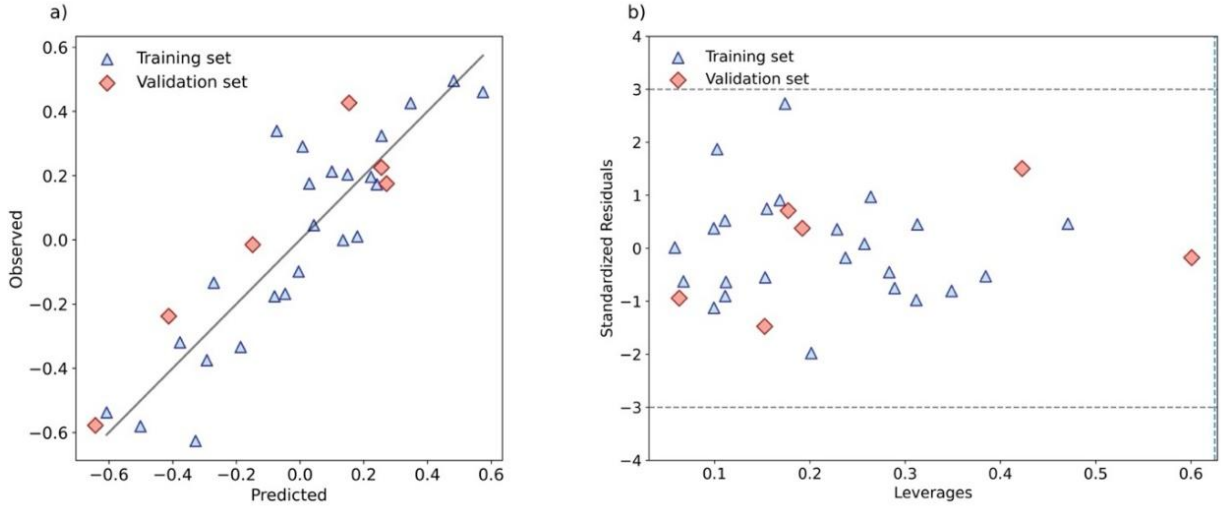


Figure 5. (a) Observed–predicted plot illustrating the accuracy of predictions in two-dimensional space; (b) Williams plot depicting the model's AD.

In particular, specific surface area, as a surrogate for the quantity of reactive surface available to cells, has been linked to oxidative stress as well as to inflammatory signaling pathways that involve *NRF2, MAPK* and *NF-κB* [178, 179], while the size of primary particles is biologically relevant – smaller particles typically have greater surface-area-to-mass ratios, may exhibit higher curvature and defect densities, and can behave differently in terms of cellular uptake and intracellular trafficking [179–183]. Consistent with findings in the field, concentration of the nanoparticles continued to emerge as a major determinant of changes at the transcriptional level in response to TiO_2-NPs, in line with the dose–response relationships documented in a number of studies [184–186]. At higher levels of exposure, cellular responses shift from patterns of adaptive stress signaling to ones characterized by more overtly harmful processes such as inflammation, apoptosis, necroinflammation, and disruptions of the cell cycle [184–187], while the contribution of post-exposure time to predictions of TRI2 reflected the inherently dynamic character of transcriptomic readouts. Early time points are usually dominated by signals of oxidative stress, innate immunity, and cytokine activity, whereas later time points may instead be more reflective of processes of resolution, of tissue repair and remodeling, of extracellular-matrix assembly, or of persistent inflammatory activity [136, 176, 186, 188]. Interactions that involved terms combining time with concentration and with size suggest that the effect of TiO_2-NP exposure on gene expression is dependent not only on the physicochemical properties of the particles themselves but also on the way in which biological responses evolve after exposure –

of particular note given that pulmonary transcriptomic responses are dynamic and may differ markedly between measurements taken shortly after exposure and those taken at later times – and the final model therefore supplies a quantitative linkage between the properties of materials, the conditions of exposure, and the resulting large-scale molecular changes.

The findings show that the concept of TRI, although originally developed in the context of MWCNT studies, can be transferred to investigations of other metal oxide nanoparticles, including TiO_2-NPs-induced pulmonary transcriptomic responses. This transfer, however, concerns the overall workflow rather than the establishment of a single, universal definition of TRI. Because the numerical value of TRI depends on the particular set of genes analyzed, the specific experimental set-up, the methods of pre-processing applied, and the structure of the PCs used in a given study, values of TRI are only directly comparable within the context of a harmonized dataset from which they have been derived. For this reason, the application of TRI to new nanomaterial classes or exposure paradigms requires recalculation and optimization of the index rather than direct reuse of previously established values. A key contribution of this work is the formalization of a data-driven rule for the choice of the most appropriate index of TRI – namely, that the choice of the 'optimal' index should take into account, in tandem, measures of cross-validated predictive performance, of the amount of transcriptomic variance explained, of results from external validation, and of assessments of the AD. By doing so, I have taken steps to move TRI from a proof-of-concept measure into something that is more standardized and compatible with NAM approaches. In the case of TiO_2-NPs, the optimized TRI also provides a quantitative bridge between physicochemical and exposure-related nanoparticle descriptors and the global pulmonary transcriptomic response. This suggests that in regulatory or screening settings, TRI might be useful as a continuously valued measure of transcriptomic bioactivity – for purposes such as early prioritization of test items, ranking of potency, grouping, providing support for read-across arguments, and for extrapolation within a given study.

The present work demonstrates that the global changes in gene expression triggered by TiO_2-NPs can be represented within a single, optimally-tuned index and that the endpoint designated as TRI2 proved particularly effective at capturing the major biologically relevant dimensions of the transcriptomic response, avoiding the loss of predictive power that can accompany the addition of too many components. The resulting ML model was shown to perform well on external validation, to be interpretable, and to be driven by variables of clear mechanistic relevance – among them BET surface area, primary particle size, concentration, and time elapsed after

exposure – and the results support the view that TRI constitutes a transferable, though at least somewhat dataset-dependent, approach for reducing complex transcriptomic outputs to endpoints that are suitable for use in a NAM context for assessing the hazards of nanomaterials.

3.3 Titanium-based ENMs exposure biomarkers

3.3.1. Introduction

Biomarkers represent quantifiable signs of perturbation of biological systems, disease pathology, or specific responses to foreign substances [189]. In the fields of toxicology and risk assessment, those biomarkers that are derived from omics data are considered highly informative because such measures can detect early, subclinical molecular changes that occur before visible damage to tissues and can thereby illuminate the mode of action (MoA) of stressors in disease states [190–192]. This importance is underscored by the growing focus on NAMs, within which high-throughput and detailed, high-content methods are increasingly being recognized as useful for evaluating the safety of chemicals and ENMs [153].

The practical application of omics technologies to assess the hazards of nanomaterials is still limited, in part because traditional statistical methods – while capable of detecting changes in gene expression in response to TiO_2-NPs – often lack the power to predictively integrate a range of ENM physicochemical traits together with varied experimental circumstances; they are not always able to predict how groups of such factors jointly influence biological readouts. ML methods offer a way forward, since they are well-suited to analyses of high-dimensional toxicogenomic data, to the optimization of which features are used as predictors, and to the construction of models that link properties of the material with consequent biological effects [159, 193]. This is particularly important for studies of lung biology, where responses may depend on many intertwined variables such as particle size, the surface area of deposited particles, charge, dose, and time elapsed after exposure.

The aim of the work was to use supervised ML techniques to detect transcriptomic signatures that signal pulmonary damage caused by TiO_2-NPs, as well as to predict such changes based on the physicochemical traits of the particles. By modeling the behavior of over 600 differentially expressed genes, I aimed to pinpoint those genes whose expression patterns are particularly responsive to characteristics of TiO_2-NPs and that have close links to processes of inflammation and fibrotic

pathology [41, 166]. In doing so, the study seeks to overcome a major drawback of traditional transcriptomics – namely, the difficulty of turning measurements of gene expression into actionable predictive models of biomarkers – and therefore contributes to efforts at early hazard detection, to the development of SSbD approaches, and to wider adoption of omics-driven NAMs within regulatory toxicology.

3.3.2 Materials and Methods

The present investigation relied on transcriptomic datasets that had been generated in earlier studies concerning exposure of mouse lungs to TiO_2-NPs possessing a range of physicochemical traits [41, 134–136, 166]. In those original works, detailed descriptions were given of both the *in vivo* experiments and of the resulting global patterns of transcription in lung tissue following treatment with TiO_2-NPs. Gene-expression measurements were taken from lung tissue of female C57BL/6 mice aged 5-7 weeks after a single intratracheal instillation of five different TiO_2-NPs, at two time points following exposure (days 1 and 28), and included groups of 6–8 female mice treated with the nanoparticles, with vehicle-only controls, and with a reference group exposed to the well-known carbon black nanoparticle Printex90 [41, 134, 135]. The dosages of TiO_2-NP used in the original work – 18, 54, and 162 μg per mouse – had been chosen on the basis of earlier toxicological work as well as of estimates of typical occupational exposures, and correspond to roughly 1.5, 5, and 15 eight-hour workdays at the maximum level of permissible exposure for TiO_2 in Denmark [41, 134, 135]. The set of nanoparticle properties that were analyzed – which included variables such as time after exposure, dose, primary particle size, BET surface area, charge, and deposited surface area – was used as input for the construction of predictive models based on machine learning, and key details regarding the physicochemical properties of the TiO_2-NPs studied, together with their experimental contexts, are collated in Table 2.

Gene-expression data were subjected to filtering in order to isolate those transcripts that showed a significant response to at least one condition of TiO_2-NP treatment. A gene was therefore deemed differentially expressed when it satisfied the criteria $p \leq 0.05$ together with $|\text{fold change}| \geq 1.5$ – a cut-off that is in line with previous work in this area and that is thought to retain sufficient detail for subsequent modeling efforts [41, 80, 133, 137–139]. For each combination of gene and condition, responses were then recoded as binary variables: genes that met the criteria for differential expression were placed in a ‘responsive’ group (coded as "1"),

and those that fell below the cut-offs were placed in a 'non-responsive' group (coded as "0").

Random Forest (RF) classifiers were constructed in order to predict changes in gene expression in response to various physicochemical and exposure-related characteristics of TiO_2-NPs, and hyperparameter optimization was carried out via grid search with cross-validation (GridSearchCV). Other tree-based methods, including Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), were also tested in order to benchmark different approaches to prediction, although none of these alternatives led to improvements sufficient to justify their use; therefore, the analyses proceeded with the use of RF models alone. In addition, a distinct classifier was built for each different DEG, giving rise to a set of more than 600 individual models that linked properties of TiO_2-NPs, together with details of their use, to specific classes of gene expression change.

Due to frequent imbalances between gene-response classes, I employed the Synthetic Minority Oversampling Technique (SMOTE) in order to even out the composition of our training sets; by generating synthetic examples of the minority responses, this approach helped to better represent under-sampled phenotypes and to mitigate bias in favor of the majority class. In order to evaluate the predictive performance of our models, 6 observations were randomly withheld from the training process and subsequently used as independent test cases – these points were never part of the calibration procedure, thereby permitting assessment of the models' capacity to generalize to new data – and, in parallel, I carried out 5-fold cross-validation in order to probe the robustness of the methods.[98] The quality of the various classifiers was appraised on the basis of a range of metrics, including Accuracy, Precision, Recall, the F1 score, area under the receiver operating characteristic curve (ROC-AUC), and confusion matrices, with the aim of determining, for each method, how well they could discriminate between responsive and non-responsive patterns of gene expression. These measures were then used together to evaluate overall performance. For an additional test of generalisability, the predictor of the response of Saa3 was applied to quantitative reverse-transcription polymerase chain reaction (qRT-PCR) data arising from separate studies of TiO_2-NP exposure of the lung that had been conducted under the same experimental [194, 195]. Those data comprised measurements at post-exposure times ranging from 1 to 180 days, as well as deposited exposure totals in the range 13.2–481.3 cm^2, and I calculated fold changes both from mean values and from the associated standard deviations taken across the different groups of animals.

AD was delineated by first normalizing the input variables used to train the classifiers and then detecting outliers (denoted as X-outliers) within this normalized set of training cases; the AD was consequently characterized as the space of responses, together with the associated physicochemical properties, represented in the training material. Because cases that differed markedly from the distribution of the training set were regarded as potentially unreliable (given the possibility that the learning algorithm may not have captured the behavior of sparsely populated regions of feature space), predictions for such atypical test examples were treated with caution. In order to identify robust biomarker classifiers, I applied cuts on overall performance, retaining only those models that scored at least 0.7 on measures of ROC-AUC, Accuracy, Precision, Recall, and F1-score [196–199]. In addition, any models that contained genes without informative symbols were removed from consideration in subsequent work on biological significance, so that the end result was a collection of transcriptomic biomarkers whose behavior could be reliably predicted from the properties of TiO_2-NPs, and which were also biologically interpretable.

3.3.3. Results and Discussion

The primary aim of the work was to determine whether machine-learning methods could be harnessed to identify and predict transcriptomic changes accompanying exposure of mouse lungs to TiO_2-NPs. By applying thresholds of statistical significance ($p \leq 0.05$) together with a criterion on the magnitude of change (absolute fold change ≥ 1.5), I selected 621 DEGs for subsequent modeling. These data encompassed 30 different experimental conditions, which included five distinct types of TiO_2-NPs characterized by differing physicochemical properties, three different concentrations of exposure, and two time points following exposure. Each gene was recoded as a binary outcome – 'responsive' when its expression was significantly altered, and 'non-responsive' when it was not – allowing the construction of a separate supervised classifier for each DEG. Altogether, 621 Random Forest classifiers were trained, linking information about the physicochemical characteristics of the particles as well as various measures of exposure (such as deposited surface area, particle size, concentration, surface charge) and the post-exposure time, to the responsive/non-responsive status of each gene. These classifiers were assessed through cross-validation as well as on independent test sets, with a range of standard classification metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC) together with confusion matrices, and alternative tree-based approaches

such as XGBoost and LightGBM were tried, although none of these were found to outperform the Random Forest approach; consequently, Random Forest was adopted as the final method of analysis. An exemplar case of model performance was provided by the model for *Saa3*, which showed strong discriminative power both in cross-validation and on external data, and the *Saa3* model was further tested on independent quantitative polymerase chain reaction (qPCR) measurements from earlier studies of TiO_2-NP exposure, where it yielded high balanced accuracy and thus evidence for robust predictive power beyond the original training set [194, 195].

In order to isolate the most dependable and biologically interpretable classifiers, I applied cut-offs on overall predictive performance – retaining only those models that scored at least 0.7 on measures of ROC-AUC as well as on Accuracy, Precision, Recall and F1-score – which resulted in a set of 22 high-quality models. Among these, five that corresponded to genes without informative symbols were excluded from subsequent biological analyses, leaving 17 property-driven transcriptomic biomarkers (namely, *Saa3, Saa1, Pigr, Irg1, Hspa1a, Ccl2, Il1b, Serpina3n, Ccl3, Bcl6b, Il18bp, Nxph3, Hmox1, Mcam, H2-K1, Serpine1*, and *Malt1*). These genes were chosen in part on the grounds that changes in their expression could be predicted from the physicochemical properties of TiO_2-NPs as well as from details of the exposure, and that they are of known biological relevance to lung disease – many of them are linked to inflammatory pathways, acute phase responses, immune activation, oxidative stress, and to fibrotic processes (for example, *Saa3, Saa1, Il1b* and *Serpina3n* are closely tied to acute inflammation and to the acute phase response, whereas *Ccl3* and *Malt1* are involved in inflammatory signalling and in pathways associated with lung injury and fibrosis [75, 80, 168–170, 200, 201]). The identification of these 17 genes therefore shows that the transcriptomic changes induced by TiO_2-NPs are not random but are systematically linked to particular properties of the particles as well as to specific conditions of exposure; notably, several of the genes identified here have also been implicated at the protein level in inflammatory responses following inhalation of particles – for instance, *Ccl2, Il1b* and *Ccl3* have been found in lung tissue after exposure to UV-Titan L181, and levels of *Ccl2* in bronchoalveolar lavage have correlated with neutrophil influx in the context of pulmonary exposure to diesel exhaust [166, 202]. Likewise, circulating levels of *Saa3* have shown correlations with neutrophil activity after exposure to metal oxides [195] – findings that reinforce the biological relevance of the biomarkers that I have selected.

Through a series of feature-importance analyses aimed at elucidating which physicochemical properties of TiO_2-NPs most strongly affected predictions for 17

chosen biomarkers, it emerged that deposited surface area was by far the most influential factor on average, followed in importance by post-exposure time and by surface charge – findings that suggest that changes in transcriptomic responses to TiO_2-NPs arise from a combination of dosage-related, temporal, and surface chemistry factors rather than from any single characteristic of the particles themselves (Figure 6). In particular, the finding that deposited surface area was the dominant predictor is in line with previous work showing the importance of retained particle surface area as a driver of pulmonary inflammation following exposure to nanomaterials [41, 45, 203], and this property, which combines information about dose, particle size and surface properties, may therefore be seen as a more biologically relevant measure of exposure than simple mass concentration. As smaller particles tend to have higher surface area per unit mass, this may account for increased contact with lung tissue, greater rates of cellular uptake, and more pronounced effects on oxidative stress and inflammatory signaling in these contexts [41, 203–205]. The prominent role of deposited surface area therefore supports the view that interactions mediated at the surface of the particles are central to the perturbation of gene expression induced by TiO_2-NPs. Surface charge also made a substantial contribution to predictions of gene expression changes – a result that is biologically plausible given the known effects of nanoparticle charge on interactions with cell membranes, on the formation of protein coronas, on cellular internalisation, and on intracellular trafficking [206] – and the fact that positively charged particles may show enhanced uptake and longer retention, with consequent amplification of transcriptional changes associated with inflammation, oxidative stress and immune signalling, suggests that differences in charge may partly underlie why otherwise comparable TiO_2-NPs (in terms of size or dose) can give rise to different patterns of gene expression. Finally, post-exposure time was found to be an important predictor as well, signalling that the transcriptomic responses to TiO_2-NPs are dynamic over time, with short-term changes tending to be associated with acute inflammation, innate immunity and acute-phase signalling, whereas later time points may reflect more persistent inflammation, immune modulation, tissue repair or recovery processes [207–210]; thus, the timing of observations is key to making sense of the effects of TiO_2-NPs on the lung, and for discriminating between transient changes at the molecular level and potentially persistent pathological states.

This investigation has defined a panel of seventeen genes whose expression patterns are directly relevant to assessments of pulmonary risk, many of which participate in acute inflammatory signalling, chemokine-mediated communication, recruitment of immune cells, oxidative stress responses, and regulation of the

extracellular matrix – all of which are mechanistically tied to adverse outcomes such as pulmonary inflammation, fibrotic changes and long-term alterations in tissue architecture (for instance, genes such as *Ccl2* and *Il1b* play roles in monocyte/macrophage activation and inflammatory signalling, whereas *Serpine1* and *Malt1* may be linked to pathways involved in tissue remodeling and lung [170, 200, 201, 211]). The use of biomarker-driven predictive models such as those presented here therefore underpins the wider shift towards omics-informed NAMs, and, by linking physical–chemical characteristics of TiO_2-NPs with transcriptomic biomarkers, provides a mechanistically interpretable means of flagging potential hazards at early stages – information that is of particular value for programmes aimed at designing safe and sustainable products, where groups of nanomaterials might be ranked or altered prior to synthesis or large-scale deployment on the basis of predicted biological responses. In addition, the models described here may prove useful to the field of regulatory toxicology, by offering quantitative data on the way that factors such as particle size, surface area, charge and length of exposure can affect early molecular events.

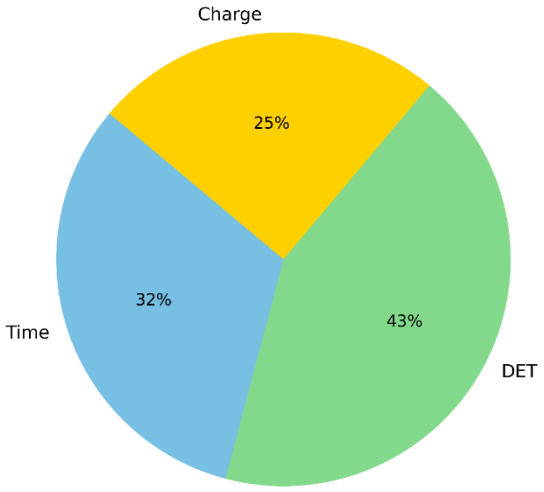


Figure 6. Average Feature Importances for the models predicting the selected 17 biomarkers.

The work shows that it is possible to use ML approaches to predict transcriptomic signatures of lung damage induced by TiO_2-NPs on the basis of the physicochemical characteristics of those particles together with details of their exposure circumstances. The resulting set of 17 biomarkers appears to have both diagnostic and explanatory value, encompassing key molecular pathways involved in inflammation, acute-phase responses, immune activation, oxidative stress, and fibrotic signaling, and analyses of feature importance showed that deposited surface area, surface charge, and the post-exposure time are key factors in determining these

effects. Taken together, the models developed here offer a workable method for flagging potential hazards at an early stage, aid the wider adoption of omics-based NAMs, and may help to make the use of nanomaterials safer by allowing predictions – on the basis of particular physical properties – of how such materials might affect pulmonary molecular biology.

4. Conclusion

This dissertation addressed a central problem in modern nanosafety assessment: how to convert complex, high-dimensional pulmonary transcriptomic responses induced by ENMs into predictive and biologically meaningful endpoints. The work was built around the hypothesis that transcriptomic perturbations caused by nanomaterial exposure are not random or purely descriptive signals, but structured biological responses that can be quantified, interpreted, and predicted from physicochemical and exposure-related properties using validated machine-learning approaches.

The results support this hypothesis. Across two distinct classes of engineered nanomaterials (MWCNTs and TiO_2-NPs), the studies demonstrated that omics data can be transformed into endpoints suitable for mechanistic hazard assessment and potency ranking. For MWCNTs, thousands of differentially expressed genes were compressed into a TRI that retained the essential structure of the global pulmonary transcriptomic response. This showed that the broad biological effect of carbon-based nanomaterials can be represented by a compact systems-level endpoint and predicted from structural descriptors using a validated nano-QSAR model. The model, therefore, provided a practical alternative to modeling thousands of genes individually, while still preserving the broader transcriptomic signal relevant to pulmonary toxicity.

The extension of the TRI concept to TiO_2-NPs further demonstrated that the framework is transferable, but not mechanically universal. The TiO_2-NPs study showed that the optimal TRI formulation must be adapted to the variance structure and biological characteristics of each dataset. While the MWCNT response required a broader principal-component representation, the TiO_2-NPs response was best captured by a more compact TRI based on the first two principal components. This finding is important because it shows that TRI is not simply a fixed mathematical index, but a flexible modeling strategy for transforming omics complexity into a predictive endpoint. The optimized TiO_2-NPs model linked global transcriptomic activity to biologically plausible descriptors, including surface area, particle size, concentration, and post-exposure time, confirming that nanomaterial-induced transcriptomic responses can be quantitatively connected to measurable material and exposure properties.

In parallel, the biomarker-level analysis provided a more mechanistic layer of interpretation. The supervised machine-learning models identified a panel of TiO_2-

NP-responsive genes associated with acute-phase responses, inflammatory signaling, immune activation, oxidative stress, and tissue-remodeling processes. Genes such as *Saa1, Saa3, Ccl2, Il1b, Hmox1, Serpine1*, and others represent biologically interpretable markers of pulmonary perturbation and possible progression toward adverse outcomes. The feature-importance analysis further showed that deposited surface area, post-exposure time, and surface charge were among the dominant drivers of predicted gene activation. This supports the broader toxicological interpretation that nanomaterial hazard cannot be explained by nominal concentration alone. Instead, the pulmonary response reflects an interaction between biologically relevant dose, surface-mediated activity, particle-cell interactions, and time-dependent tissue response.

Together, the global TRI models and biomarker-level models provide complementary views of nanomaterial-induced toxicity. The TRI captures the integrated magnitude and direction of transcriptomic perturbation at the systems level, whereas the biomarker models identify specific molecular signals connected to inflammation, oxidative stress, and fibrotic mechanisms. This combination strengthens the overall framework because it connects prediction with interpretation: the models not only estimate whether a transcriptomic response is likely to occur, but also help explain which material properties may drive that response and which biological pathways may be involved.

The dissertation therefore establishes a coherent properties-to-omics predictive framework for nanosafety assessment. It shows that physicochemical and exposure-related properties of engineered nanomaterials can be linked to both global and gene-specific transcriptomic outcomes through validated ML workflows. This directly addresses the need for NAMs and IATA that can reduce dependence on extensive animal testing while preserving mechanistic relevance. The developed approach can support early hazard identification, potency ranking, prioritization of materials for further testing, and SSbD strategies by identifying potentially hazardous properties before later-stage adverse outcomes are observed.

5. References

[1] Fu R, Fu G-D, 2011. Polymeric nanomaterials from combined click chemistry and controlled radical polymerization. Polym Chem 2, 465–475. https://doi.org/10.1039/C0PY00174K

[2] International Organization for Standardization, 2023. Nanotechnologies – Vocabulary — Part 1: Core vocabulary

[3] Dolez PI, 2015. Nanomaterials Definitions, Classifications, and Applications. In: Nanoengineering. Elsevier, pp 3–40

[4] Bafna A, Beaucage G, Mirabella F, Mehta S, 2003. 3D Hierarchical orientation in polymer–clay nanocomposite films. Polymer (Guildf) 44, 1103–1115. https://doi.org/10.1016/S0032-3861(02)00833-9

[5] Nanomaterials, Nanotechnologies and Design, 2009. Elsevier

[6] Handbook of Nanosafety, 2014. Elsevier

[7] Raj S, Jose S, Sumod US, Sabitha M, 2012. Nanotechnology in cosmetics: Opportunities and challenges. J Pharm Bioallied Sci 4, 186–93. https://doi.org/10.4103/0975-7406.99016

[8] Cheng Z, Li M, Dey R, Chen Y, 2021. Nanomaterials for cancer therapy: current progress and perspectives. J Hematol Oncol 14, 85. https://doi.org/10.1186/s13045-021-01096-0

[9] Nguyen TMT, Wang P-W, Hsu H-M, Cheng F-Y, Shieh D-B, Wong T-Y, Chang H-J, 2019. Dental cement's biological and mechanical properties improved by ZnO nanospheres. Materials Science and Engineering: C 97, 116–123. https://doi.org/10.1016/j.msec.2018.12.007

[10] Som C, Wick P, Krug H, Nowack B, 2011. Environmental and health effects of nanomaterials in nanotextiles and façade coatings. Environ Int 37, 1131–1142. https://doi.org/10.1016/j.envint.2011.02.013

[11] Khan A, Wen Y, Huq T, Ni Y, 2018. Cellulosic Nanomaterials in Food and Nutraceutical Applications: A Review. J Agric Food Chem 66, 8–19. https://doi.org/10.1021/acs.jafc.7b04204

[12] Li J, Xin M, Ma Z, Shi Y, Pan L, 2021. Nanomaterials and their applications on bio-inspired wearable electronics. Nanotechnology 32, 472002. https://doi.org/10.1088/1361-6528/abe6c7

[13] Grand View Research, 2026. Nanomaterials Market (2026 - 2033) Size, Share & Trends Analysis Report By Material (Titanium, Carbon Nanotubes, Silver, Aluminum Oxide, Gold, Platinum, Nickel), By Application (Medical, Electronics, Automotive, Aerospace), By Region, And Segment Forecasts

[14] Younis AB, Haddad Y, Kosaristanova L, Smerkova K, 2023. Titanium dioxide nanoparticles: Recent progress in antimicrobial applications. WIREs Nanomedicine and Nanobiotechnology 15. https://doi.org/10.1002/wnan.1860

[15] Mohammadparast V, Mallard BL, 2023. The effect and underlying mechanisms of titanium dioxide nanoparticles on glucose homeostasis: A literature review. Journal of Applied Toxicology 43, 22–31. https://doi.org/10.1002/jat.4318

[16] Margarucci LM, Romano Spica V, Gianfranceschi G, Valeriani F, 2019. Untouchability of natural spa waters: Perspectives for treatments within a personalized water safety plan. Environ Int 133, 105095. https://doi.org/10.1016/j.envint.2019.105095

[17] Thevenot P, Cho J, Wavhal D, Timmons RB, Tang L, 2008. Surface chemistry influences cancer killing effect of TiO2 nanoparticles. Nanomedicine 4, 226–236. https://doi.org/10.1016/j.nano.2008.04.001

[18] Khosroyar S, Arastehnodeh A, 2018. Improving hydrophilic and antimicrobial properties of membrane by adding nanoparticles of titanium dioxide and copper oxide. Membrane Water Treatment 9, 481–487. https://doi.org/10.12989/mwt.2018.9.6.481
[19] Sun C-F, Meany B, Wang Y, 2014. Characteristics and Applications of Carbon Nanotubes with Different Numbers of Walls. In: Carbon Nanotubes and Graphene. Elsevier, pp 313–339
[20] Manzetti S, Andersen O, 2012. Toxicological aspects of nanomaterials used in energy harvesting consumer electronics. Renewable and Sustainable Energy Reviews 16, 2102–2110. https://doi.org/10.1016/j.rser.2012.01.037
[21] Wang X, Li B, Jing H, Dong X, Leng X, 2020. MWCNT-mediated combinatorial photothermal ablation and chemo-immunotherapy strategy for the treatment of melanoma. J Mater Chem B 8, 4245–4258. https://doi.org/10.1039/C9TB02238D
[22] Saranya J, Sreeja BS, Padmalaya G, Radha S, Senthil Kumar P, 2020. CdO nanoparticles, c-MWCNT nanoparticles and CdO nanoparticles/c-MWCNT nanocomposite fibres: in vitro assessment of anti-proliferative and apoptotic studies in HeLa cancer cell line. IET Nanobiotechnol 14, 695–700. https://doi.org/10.1049/iet-nbt.2020.0020
[23] Dizaji BF, Farboudi A, Rahbar A, Azarbaijan MH, Asgary MR, 2020. The Role of single- and multi-walled Carbon Nanotube in Breast Cancer Treatment. Ther Deliv 11, 653–672. https://doi.org/10.4155/tde-2020-0019
[24] Şen M, Oğuz M, Avcı İ, 2024. Non-toxic flexible screen-printed MWCNT-based electrodes for non-invasive biomedical applications. Talanta 268, 125341. https://doi.org/10.1016/j.talanta.2023.125341
[25] Shah A, Senapati S, Murthy HCA, Singh LR, Mahato M, 2023. Supercapacitor Performance of NiO, NiO-MWCNT, and NiO–Fe-MWCNT Composites. ACS Omega 8, 33380–33391. https://doi.org/10.1021/acsomega.3c03044
[26] Staudinger U, Janke A, Simon F, Jakisch L, Bittrich E, Formanek P, Mielke L, Schlicke H, Li Q, Eckstein K, Pospiech D, 2025. MWCNT Localization and Electrical Percolation in Thin Films of Semifluorinated PMMA Block Copolymers. Polymers (Basel) 17, 1271. https://doi.org/10.3390/polym17091271
[27] Dou H, Liu P, Li X-Y, He C-H, 2020. Effect of MWCNT on the Structure and Property of Nanofibrous Bundles by Blown Bubble Spinning. Recent Pat Nanotechnol 13, 171–180. https://doi.org/10.2174/1872210513666190426143559
[28] Shi H, Magaye R, Castranova V, Zhao J, 2013. Titanium dioxide nanoparticles: a review of current toxicological data. Part Fibre Toxicol 10, 15. https://doi.org/10.1186/1743-8977-10-15
[29] Kuijpers E, Bekker C, Fransman W, Brouwer D, Tromp P, Vlaanderen J, Godderis L, Hoet P, Lan Q, Silverman D, Vermeulen R, Pronk A, 2016. Occupational Exposure to Multi-Walled Carbon Nanotubes During Commercial Production Synthesis and Handling. Ann Occup Hyg 60, 305–17. https://doi.org/10.1093/annhyg/mev082
[30] Bourdon JA, Saber AT, Jacobsen NR, Jensen KA, Madsen AM, Lamson JS, Wallin H, Møller P, Loft S, Yauk CL, Vogel UB, 2012. Carbon black nanoparticle instillation induces sustained inflammation and genotoxicity in mouse lung and liver. Part Fibre Toxicol 9, 5. https://doi.org/10.1186/1743-8977-9-5
[31] Oberdörster G, Castranova V, Asgharian B, Sayre P, 2015. Inhalation Exposure to Carbon Nanotubes (CNT) and Carbon Nanofibers (CNF): Methodology and Dosimetry. Journal of Toxicology and Environmental Health, Part B 18, 121–212. https://doi.org/10.1080/10937404.2015.1051611

[32] Sturm R, Hofmann W, 2009. A theoretical approach to the deposition and clearance of fibers with variable size in the human respiratory tract. J Hazard Mater 170, 210–218. https://doi.org/10.1016/j.jhazmat.2009.04.107
[33] Yang W, Wang L, Mettenbrink EM, DeAngelis PL, Wilhelm S, 2021. Nanoparticle Toxicology. Annu Rev Pharmacol Toxicol 61, 269–289. https://doi.org/10.1146/annurev-pharmtox-032320-110338
[34] Liu H, Ma L, Zhao J, Liu J, Yan J, Ruan J, Hong F, 2009. Biochemical Toxicity of Nano-anatase TiO2 Particles in Mice. Biol Trace Elem Res 129, 170–180. https://doi.org/10.1007/s12011-008-8285-6
[35] El-Kalliny AS, Abdel-Wahed MS, El-Zahhar AA, Hamza IA, Gad-Allah TA, 2023. Nanomaterials: a review of emerging contaminants with potential health or environmental impact. Discover Nano 18, 68. https://doi.org/10.1186/s11671-023-03787-8
[36] Armstead AL, Li B, 2016. Nanotoxicity: emerging concerns regarding nanomaterial safety and occupational hard metal (WC-Co) nanoparticle exposure. Int J Nanomedicine Volume 11, 6421–6433. https://doi.org/10.2147/IJN.S121238
[37] Wang J, Fan Y, 2014. Lung Injury Induced by TiO2 Nanoparticles Depends on Their Structural Features: Size, Shape, Crystal Phases, and Surface Coating. Int J Mol Sci 15, 22258–22278. https://doi.org/10.3390/ijms151222258
[38] Oberdörster G, Oberdörster E, Oberdörster J, 2005. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environ Health Perspect 113, 823–39. https://doi.org/10.1289/ehp.7339
[39] Geiser M, Kreyling WG, 2010. Deposition and biokinetics of inhaled nanoparticles. Part Fibre Toxicol 7, 2. https://doi.org/10.1186/1743-8977-7-2
[40] Kreyling WG, Holzwarth U, Schleh C, Hirn S, Wenk A, Schäffler M, Haberl N, Semmler-Behnke M, Gibson N, 2019. Quantitative biokinetics over a 28 day period of freshly generated, pristine, 20 nm titanium dioxide nanoparticle aerosols in healthy adult rats after a single two-hour inhalation exposure. Part Fibre Toxicol 16, 29. https://doi.org/10.1186/s12989-019-0303-7
[41] Halappanavar S, Saber AT, Decan N, Jensen KA, Wu D, Jacobsen NR, Guo C, Rogowski J, Koponen IK, Levin M, Madsen AM, Atluri R, Snitka V, Birkedal RK, Rickerby D, Williams A, Wallin H, Yauk CL, Vogel U, 2015. Transcriptional profiling identifies physicochemical properties of nanomaterials that are determinants of the in vivo pulmonary response. Environ Mol Mutagen 56, 245–264. https://doi.org/10.1002/em.21936
[42] Asgharian B, Owen TP, Kuempel ED, Jarabek AM, 2018. Dosimetry of inhaled elongate mineral particles in the respiratory tract: The impact of shape factor. Toxicol Appl Pharmacol 361, 27–35. https://doi.org/10.1016/j.taap.2018.05.001
[43] Braakhuis HM, Park MV, Gosens I, De Jong WH, Cassee FR, 2014. Physicochemical characteristics of nanomaterials that affect pulmonary inflammation. Part Fibre Toxicol 11, 18. https://doi.org/10.1186/1743-8977-11-18
[44] Lee D-K, Kim G, Maruthupandy M, Lee K, Cho W-S, 2024. Multimodal pulmonary clearance kinetics of carbon black nanoparticles deposited in the lungs of rats: the role of alveolar macrophages. Part Fibre Toxicol 21, 32. https://doi.org/10.1186/s12989-024-00591-9
[45] Cosnier F, Seidel C, Valentino S, Schmid O, Bau S, Vogel U, Devoy J, Gaté L, 2021. Retained particle surface area dose drives inflammation in rat lungs following acute, subacute, and subchronic inhalation of nanomaterials. Part Fibre Toxicol 18, 29. https://doi.org/10.1186/s12989-021-00419-w
[46] Sanchez VC, Pietruska JR, Miselis NR, Hurt RH, Kane AB, 2009. Biopersistence and potential adverse health impacts of fibrous nanomaterials:

what have we learned from asbestos? WIREs Nanomedicine and Nanobiotechnology 1, 511–529. https://doi.org/10.1002/wnan.41
[47] Buckley A, Warren J, Hodgson A, Marczylo T, Ignatyev K, Guo C, Smith R, 2017. Slow lung clearance and limited translocation of four sizes of inhaled iridium nanoparticles. Part Fibre Toxicol 14, 5. https://doi.org/10.1186/s12989-017-0185-5
[48] Creutzenberg O, Pohlmann G, Schaudien D, Kock H, 2022. Toxicokinetics of Nanoparticles Deposited in Lungs Using Occupational Exposure Scenarios. Front Public Health 10. https://doi.org/10.3389/fpubh.2022.909247
[49] Marui T, Tomonaga T, Izumi H, Yoshiura Y, Nishida C, Higashi H, Wang K-Y, Shijo M, Kubo M, Shimada M, Morimoto Y, 2022. Pulmonary toxicity of tungsten trioxide nanoparticles in an inhalation study and an intratracheal instillation study. J Occup Health 64. https://doi.org/10.1002/1348-9585.12367
[50] Braakhuis HM, Gosens I, Krystek P, Boere JA, Cassee FR, Fokkens PH, Post JA, van Loveren H, Park MV, 2014. Particle size dependent deposition and pulmonary inflammation after short-term inhalation of silver nanoparticles. Part Fibre Toxicol 11, 49. https://doi.org/10.1186/s12989-014-0049-1
[51] Noël A, Charbonneau M, Cloutier Y, Tardif R, Truchon G, 2013. Rat pulmonary responses to inhaled nano-TiO2: effect of primary particle size and agglomeration state. Part Fibre Toxicol 10, 48. https://doi.org/10.1186/1743-8977-10-48
[52] Vietti G, Lison D, van den Brule S, 2015. Mechanisms of lung fibrosis induced by carbon nanotubes: towards an Adverse Outcome Pathway (AOP). Part Fibre Toxicol 13, 11. https://doi.org/10.1186/s12989-016-0123-y
[53] Ryman-Rasmussen JP, Cesta MF, Brody AR, Shipley-Phillips JK, Everitt JI, Tewksbury EW, Moss OR, Wong BA, Dodd DE, Andersen ME, Bonner JC, 2009. Inhaled carbon nanotubes reach the subpleural tissue in mice. Nat Nanotechnol 4, 747–751. https://doi.org/10.1038/nnano.2009.305
[54] Halappanavar S, van den Brule S, Nymark P, Gaté L, Seidel C, Valentino S, Zhernovkov V, Høgh Danielsen P, De Vizcaya A, Wolff H, Stöger T, Boyadziev A, Poulsen SS, Sørli JB, Vogel U, 2020. Adverse outcome pathways as a tool for the design of testing strategies to support the safety assessment of emerging advanced materials at the nanoscale. Part Fibre Toxicol 17, 16. https://doi.org/10.1186/s12989-020-00344-4
[55] Clippinger AJ, Ahluwalia A, Allen D, Bonner JC, Casey W, Castranova V, David RM, Halappanavar S, Hotchkiss JA, Jarabek AM, Maier M, Polk W, Rothen-Rutishauser B, Sayes CM, Sayre P, Sharma M, Stone V, 2016. Expert consensus on an in vitro approach to assess pulmonary fibrogenic potential of aerosolized nanomaterials. Arch Toxicol 90, 1769–1783. https://doi.org/10.1007/s00204-016-1717-8
[56] Oomen A, Bleeker E, Bos P, Van Broekhuizen F, Gottardo S, Groenewold M, Hristozov D, Hund-Rinke K, Irfan M-A, Marcomini A, Peijnenburg W, Rasmussen K, Jiménez A, Scott-Fordsmand J, Van Tongeren M, Wiench K, Wohlleben W, Landsiedel R, 2015. Grouping and Read-Across Approaches for Risk Assessment of Nanomaterials. Int J Environ Res Public Health 12, 13415–13434. https://doi.org/10.3390/ijerph121013415
[57] More S, Bampidis V, Benford D, Bragard C, Halldorsson T, Hernández-Jerez A, Hougaard Bennekou S, Koutsoumanis K, Lambré C, Machera K, Naegeli H, Nielsen S, Schlatter J, Schrenk D, Silano (deceased) V, Turck D, Younes M, Castenmiller J, Chaudhry Q, Cubadda F, Franz R, Gott D, Mast J, Mortensen A, Oomen AG, Weigel S, Barthelemy E, Rincon A, Tarazona J, Schoonjans R, 2021. Guidance on risk assessment of nanomaterials to be applied in the food and feed chain: human and animal health. EFSA Journal 19. https://doi.org/10.2903/j.efsa.2021.6768

[58] Landsiedel R, Ma-Hock L, Wiench K, Wohlleben W, Sauer UG, 2017. Safety assessment of nanomaterials using an advanced decision-making framework, the DF4nanoGrouping. Journal of Nanoparticle Research 19, 171. https://doi.org/10.1007/s11051-017-3850-6
[59] Afantitis A, Melagraki G, Isigonis P, Tsoumanis A, Varsou DD, Valsami-Jones E, Papadiamantis A, Ellis L-JA, Sarimveis H, Doganis P, Karatzas P, Tsiros P, Liampa I, Lobaskin V, Greco D, Serra A, Kinaret PAS, Saarimäki LA, Grafström R, Kohonen P, Nymark P, Willighagen E, Puzyn T, Rybinska-Fryca A, Lyubartsev A, Alstrup Jensen K, Brandenburg JG, Lofts S, Svendsen C, Harrison S, Maier D, Tamm K, Jänes J, Sikk L, Dusinska M, Longhin E, Rundén-Pran E, Mariussen E, El Yamani N, Unger W, Radnik J, Tropsha A, Cohen Y, Leszczynski J, Ogilvie Hendren C, Wiesner M, Winkler D, Suzuki N, Yoon TH, Choi J-S, Sanabria N, Gulumian M, Lynch I, 2020. NanoSolveIT Project: Driving nanoinformatics research to develop innovative and integrated tools for in silico nanosafety assessment. Comput Struct Biotechnol J 18, 583–602. https://doi.org/10.1016/j.csbj.2020.02.023
[60] Forest V, 2022. Experimental and Computational Nanotoxicology—Complementary Approaches for Nanomaterial Hazard Assessment. Nanomaterials 12, 1346. https://doi.org/10.3390/nano12081346
[61] Di Cristo L, Oomen AG, Dekkers S, Moore C, Rocchia W, Murphy F, Johnston HJ, Janer G, Haase A, Stone V, Sabella S, 2021. Grouping Hypotheses and an Integrated Approach to Testing and Assessment of Nanomaterials Following Oral Ingestion. Nanomaterials 11, 2623. https://doi.org/10.3390/nano11102623
[62] Braakhuis HM, Murphy F, Ma-Hock L, Dekkers S, Keller J, Oomen AG, Stone V, 2021. An Integrated Approach to Testing and Assessment to Support Grouping and Read-Across of Nanomaterials After Inhalation Exposure. Appl In Vitro Toxicol 7, 112–128. https://doi.org/10.1089/aivt.2021.0009
[63] Giubilato E, Cazzagon V, Amorim MJB, Blosi M, Bouillard J, Bouwmeester H, Costa AL, Fadeel B, Fernandes TF, Fito C, Hauser M, Marcomini A, Nowack B, Pizzol L, Powell L, Prina-Mello A, Sarimveis H, Scott-Fordsmand JJ, Semenzin E, Stahlmecke B, Stone V, Vignes A, Wilkins T, Zabeo A, Tran L, Hristozov D, 2020. Risk Management Framework for Nano-Biomaterials Used in Medical Devices and Advanced Therapy Medicinal Products. Materials 13, 4532. https://doi.org/10.3390/ma13204532
[64] Scott-Fordsmand J, Peijnenburg W, Semenzin E, Nowack B, Hunt N, Hristozov D, Marcomini A, Irfan M, Jiménez AS, Landsiedel R, Tran L, Oomen A, Bos P, Hund-Rinke K, 2017. Environmental Risk Assessment Strategy for Nanomaterials. Int J Environ Res Public Health 14, 1251. https://doi.org/10.3390/ijerph14101251
[65] Shao G, Beronius A, Nymark P, 2023. SciRAPnano: a pragmatic and harmonized approach for quality evaluation of in vitro toxicity data to support risk assessment of nanomaterials. Frontiers in Toxicology 5. https://doi.org/10.3389/ftox.2023.1319985
[66] Furxhi I, Perucca M, Blosi M, Lopez de Ipiña J, Oliveira J, Murphy F, Costa AL, 2022. ASINA Project: Towards a Methodological Data-Driven Sustainable and Safe-by-Design Approach for the Development of Nanomaterials. Front Bioeng Biotechnol 9. https://doi.org/10.3389/fbioe.2021.805096
[67] Schmutz M, Borges O, Jesus S, Borchard G, Perale G, Zinn M, Sips ÄAJAM, Soeteman-Hernandez LG, Wick P, Som C, 2020. A Methodological Safe-by-Design Approach for the Development of Nanomedicines. Front Bioeng Biotechnol 8. https://doi.org/10.3389/fbioe.2020.00258
[68] Varsou D-D, Afantitis A, Tsoumanis A, Melagraki G, Sarimveis H, Valsami-Jones E, Lynch I, 2019. A safe-by-design tool for functionalised

nanomaterials through the Enalos Nanoinformatics Cloud platform. Nanoscale Adv 1, 706–718. https://doi.org/10.1039/C8NA00142A
[69] Ruijter N, Soeteman-Hernández LG, Carrière M, Boyles M, McLean P, Catalán J, Katsumiti A, Cabellos J, Delpivo C, Sánchez Jiménez A, Candalija A, Rodríguez-Llopis I, Vázquez-Campos S, Cassee FR, Braakhuis H, 2023. The State of the Art and Challenges of In Vitro Methods for Human Hazard Assessment of Nanomaterials in the Context of Safe-by-Design. Nanomaterials 13, 472. https://doi.org/10.3390/nano13030472
[70] Fatkhutdinova LM, Khaliullin TO, Vasil'yeva OL, Zalyalov RR, Mustafin IG, Kisin ER, Birch ME, Yanamala N, Shvedova AA, 2016. Fibrosis biomarkers in workers exposed to MWCNTs. Toxicol Appl Pharmacol 299, 125–131. https://doi.org/10.1016/j.taap.2016.02.016
[71] Buonaurio F, Astolfi ML, Canepari S, Di Basilio M, Gibilras R, Mecchia M, Papacchini M, Paci E, Pigini D, Tranfo G, 2020. Urinary Oxidative Stress Biomarkers in Workers of a Titanium Dioxide Based Pigment Production Plant. Int J Environ Res Public Health 17, 9085. https://doi.org/10.3390/ijerph17239085
[72] Schulte PA, Leso V, Niang M, Iavicoli I, 2019. Current state of knowledge on the health effects of engineered nanomaterials in workers: a systematic review of human studies and epidemiological investigations. Scand J Work Environ Health 45, 217–238. https://doi.org/10.5271/sjweh.3800
[73] Beard JD, Erdely A, Dahm MM, de Perio MA, Birch ME, Evans DE, Fernback JE, Eye T, Kodali V, Mercer RR, Bertke SJ, Schubauer-Berigan MK, 2018. Carbon nanotube and nanofiber exposure and sputum and blood biomarkers of early effect among U.S. workers. Environ Int 116, 214–228. https://doi.org/10.1016/j.envint.2018.04.004
[74] Iavicoli I, Leso V, Schulte PA, 2016. Biomarkers of susceptibility: State of the art and implications for occupational exposure to engineered nanomaterials. Toxicol Appl Pharmacol 299, 112–124. https://doi.org/10.1016/j.taap.2015.12.018
[75] Nikota J, Williams A, Yauk CL, Wallin H, Vogel U, Halappanavar S, 2016. Meta-analysis of transcriptomic responses as a means to identify pulmonary disease outcomes for engineered nanomaterials. Part Fibre Toxicol 13, 25. https://doi.org/10.1186/s12989-016-0137-5
[76] Labib S, Williams A, Yauk CL, Nikota JK, Wallin H, Vogel U, Halappanavar S, 2015. Nano-risk Science: application of toxicogenomics in an adverse outcome pathway framework for risk assessment of multi-walled carbon nanotubes. Part Fibre Toxicol 13, 15. https://doi.org/10.1186/s12989-016-0125-9
[77] Schulte PA, Kuempel ED, Drew NM, 2018. Characterizing risk assessments for the development of occupational exposure limits for engineered nanomaterials. Regulatory Toxicology and Pharmacology 95, 207–219. https://doi.org/10.1016/j.yrtph.2018.03.018
[78] Valente A, Vieira L, Silva MJ, Ventura C, 2023. The Effect of Nanomaterials on DNA Methylation: A Review. Nanomaterials 13, 1880. https://doi.org/10.3390/nano13121880
[79] Rahman L, Wu D, Johnston M, William A, Halappanavar S, 2017. Toxicogenomics analysis of mouse lung responses following exposure to titanium dioxide nanomaterials reveal their disease potential at high doses. Mutagenesis 32, 59–76. https://doi.org/10.1093/mutage/gew048
[80] Halappanavar S, Jackson P, Williams A, Jensen KA, Hougaard KS, Vogel U, Yauk CL, Wallin H, 2011. Pulmonary response to surface-coated nanotitanium dioxide particles includes induction of acute phase response genes, inflammatory cascades, and changes in microRNAs: A toxicogenomic

study. Environ Mol Mutagen 52, 425–439. https://doi.org/10.1002/em.20639
[81] Khaliullin TO, Yanamala N, Newman MS, Kisin ER, Fatkhutdinova LM, Shvedova AA, 2020. Comparative analysis of lung and blood transcriptomes in mice exposed to multi-walled carbon nanotubes. Toxicol Appl Pharmacol 390, 114898. https://doi.org/10.1016/j.taap.2020.114898
[82] Tonoyan L, Siraki AG, 2024. Machine learning in toxicological sciences: opportunities for assessing drug toxicity. Frontiers in Drug Discovery 4. https://doi.org/10.3389/fddsv.2024.1336025
[83] Singh AV, Varma M, Laux P, Choudhary S, Datusalia AK, Gupta N, Luch A, Gandhi A, Kulkarni P, Nath B, 2023. Artificial intelligence and machine learning disciplines with the potential to improve the nanotoxicology and nanomedicine fields: a comprehensive review. Arch Toxicol 97, 963–979. https://doi.org/10.1007/s00204-023-03471-x
[84] Bolt HM, Hengstler JG, 2020. The rapid development of computational toxicology. Arch Toxicol 94, 1371–1372. https://doi.org/10.1007/s00204-020-02768-5
[85] Campagnolo L, Lacconi V, Filippi J, Martinelli E, 2024. Twenty years of in vitro nanotoxicology: how AI could make the difference. Frontiers in Toxicology 6. https://doi.org/10.3389/ftox.2024.1470439
[86] Ramirez-Hincapie S, Birk B, Ternes P, Giri V, Zickgraf FM, Haake V, Herold M, Kamp H, Driemert P, Landsiedel R, Richling E, Funk-Weyer D, van Ravenzwaay B, 2023. Application of high throughput in vitro metabolomics for hepatotoxicity mode of action characterization and mechanistic-anchored point of departure derivation: a case study with nitrofurantoin. Arch Toxicol 97, 2903–2917. https://doi.org/10.1007/s00204-023-03572-7
[87] Schirmer K, Fischer BB, Madureira DJ, Pillai S, 2010. Transcriptomics in ecotoxicology. Anal Bioanal Chem 397, 917–923. https://doi.org/10.1007/s00216-010-3662-3
[88] Kim SH, Cho SY, 2023. Single-cell transcriptomics to understand the cellular heterogeneity in toxicology. Mol Cell Toxicol 19, 223–228. https://doi.org/10.1007/s13273-022-00304-3
[89] Costa E, Johnson KJ, Walker CA, O'Brien JM, 2024. Transcriptomic point of departure determination: a comparison of distribution-based and gene set-based approaches. Front Genet 15. https://doi.org/10.3389/fgene.2024.1374791
[90] Farmahin R, Williams A, Kuo B, Chepelev NL, Thomas RS, Barton-Maclaren TS, Curran IH, Nong A, Wade MG, Yauk CL, 2017. Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment. Arch Toxicol 91, 2045–2065. https://doi.org/10.1007/s00204-016-1886-5
[91] Meneses J, González-Durruthy M, Fernandez-de-Gortari E, Toropova AP, Toropov AA, Alfaro-Moreno E, 2023. A Nano-QSTR model to predict nano-cytotoxicity: an approach using human lung cells data. Part Fibre Toxicol 20, 21. https://doi.org/10.1186/s12989-023-00530-0
[92] Miszczak M, Khan K, Danielsen PH, Jensen KA, Vogel U, Grafström R, Gajewicz-Skretna A, 2025. Dynamic QSAR modeling for predicting in vivo genotoxicity and inflammation induced by nanoparticles and advanced materials: a time-dose-property/response approach. J Nanobiotechnology 23, 420. https://doi.org/10.1186/s12951-025-03510-y
[93] I. T. Jolliffe, 2002. Principal Component Analysis. Springer-Verlag, New York
[94] Granato D, Santos JS, Escher GB, Ferreira BL, Maggio RM, 2018. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional

properties in foods: A critical perspective. Trends Food Sci Technol 72, 83–90. https://doi.org/10.1016/j.tifs.2017.12.006
[95] Jolliffe IT, Cadima J, 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, 20150202. https://doi.org/10.1098/rsta.2015.0202
[96] Lever J, Krzywinski M, Altman N, 2017. Principal component analysis. Nat Methods 14, 641–642. https://doi.org/10.1038/nmeth.4346
[97] Saha D, Manickavasagan A, 2021. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. Curr Res Food Sci 4, 28–44. https://doi.org/10.1016/j.crfs.2021.01.002
[98] Pedregosa F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Pedregosa F, Varoquaux G, Gramfort A, Thirion B, Grisel O, Dubourg V, Passos A, Brucher M, Perrot andÉdouardand M, Duchesnay andÉdouard, Duchesnay EDOUARDDUCHESNAY Fré, 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830
[99] Cawley GC, Talbot NLC, 2008. Efficient approximate leave-one-out cross-validation for kernel logistic regression. Mach Learn 71, 243–264. https://doi.org/10.1007/s10994-008-5055-9
[100] Matošević G, Dobša J, Mladenić D, 2021. Using Machine Learning for Web Page Classification in Search Engine Optimization. Future Internet 13, 9. https://doi.org/10.3390/fi13010009
[101] Dai D, Javed F, Karlsson P, Månsson K, 2025. Nonlinear forecasting with many predictors using mixed data sampling kernel ridge regression models. Ann Oper Res. https://doi.org/10.1007/s10479-025-06486-y
[102] Mohammed MS, Paraman N, Rahman AA-HA, Ghaleb FA, Al-Dhamari A, Marsono MN, 2021. PEW: Prediction-Based Early Dark Cores Wake-up Using Online Ridge Regression for Many-Core Systems. IEEE Access 9, 124087–124099. https://doi.org/10.1109/ACCESS.2021.3109717
[103] Teixeira AL, Leal JP, Falcao AO, 2013. Random forests for feature selection in QSPR models - An application for predicting standard enthalpy of formation of hydrocarbons. J Cheminform 5. https://doi.org/10.1186/1758-2946-5-9
[104] Lewis JE, Kemp ML, 2021. Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. Nat Commun 12, 2700. https://doi.org/10.1038/s41467-021-22989-1
[105] Concu R, Kleandrova V V., Speck-Planche A, Cordeiro MNDS, 2017. Probing the toxicity of nanoparticles: a unified *in silico* machine learning model based on perturbation theory. Nanotoxicology 11, 891–906. https://doi.org/10.1080/17435390.2017.1379567
[106] Polishchuk P, Tinkov O, Khristova T, Ognichenko L, Kosinskaya A, Varnek A, Kuz'min V, 2016. Structural and Physico-Chemical Interpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis. J Chem Inf Model 56, 1455–1469. https://doi.org/10.1021/acs.jcim.6b00371
[107] Aptula AO, Jeliazkova NG, Schultz TW, Cronin MTD, 2005. The Better Predictive Model: High q^2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set? QSAR Comb Sci 24, 385–396. https://doi.org/10.1002/qsar.200430909
[108] Ognichenko LN, Kuz'min VE, Artemenko AG, 2009. New Structural Descriptors of Molecules on the Basis of Symbiosis of the Informational Field Model and Simplex Representation of Molecular Structure. QSAR Comb Sci 28, 939–945. https://doi.org/10.1002/qsar.200860073
[109] Jillella GK, Ojha PK, Roy K, 2021. Application of QSAR for the identification of key molecular fragments and reliable predictions of effects of textile dyes on

growth rate and biomass values of Raphidocelis subcapitata. Aquatic Toxicology 238, 105925. https://doi.org/10.1016/j.aquatox.2021.105925
[110] Masand VH, Mahajan DT, Nazeruddin GM, Hadda T Ben, Rastija V, Alfeefy AM, 2015. Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model. Medicinal Chemistry Research 24, 1241–1264. https://doi.org/10.1007/s00044-014-1193-8
[111] Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari TP, Michalkova A, Hwang H-M, Toropov A, Leszczynska D, Leszczynski J, 2011. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nat Nanotechnol 6, 175–178. https://doi.org/10.1038/nnano.2011.10
[112] OECD Omics Reporting Framework (OORF): Guidance on reporting elements for the regulatory use of omics data from laboratory-based toxicology studies, 2023. OECD
[113] Kuz'min VE, Artemenko AG, Muratov EN, 2008. Hierarchical QSAR technology based on the Simplex representation of molecular structure. J Comput Aided Mol Des 22, 403–421. https://doi.org/10.1007/s10822-008-9179-6
[114] Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy A V., Andronati SA, Makan SYu, 2005. Hierarchic system of QSAR models (1D–4D) on the base of simplex representation of molecular structure. J Mol Model 11, 457–467. https://doi.org/10.1007/s00894-005-0237-x
[115] Polishchuk P, Mokshyna E, Kosinskaya A, Muats A, Kulinsky M, Tinkov O, Ognichenko L, Khristova T, Artemenko A, Kuz'min V, 2017. Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure. pp 107–147
[116] Gupta SS, Singh KP, Gupta S, Dusinska M, Rahman Q, 2022. Do Carbon Nanotubes and Asbestos Fibers Exhibit Common Toxicity Mechanisms? Nanomaterials (Basel) 12. https://doi.org/10.3390/nano12101708
[117] Vietti G, Lison D, van den Brule S, 2016. Mechanisms of lung fibrosis induced by carbon nanotubes: towards an Adverse Outcome Pathway (AOP). Part Fibre Toxicol 13, 11. https://doi.org/10.1186/s12989-016-0123-y
[118] Kobayashi N, Izumi H, Morimoto Y, 2017. Review of toxicity studies of carbon nanotubes. J Occup Health 59, 394–407. https://doi.org/10.1539/joh.17-0089-RA
[119] Nishida C, Izumi H, Tomonaga T, Takeshita J, Wang K-Y, Yamasaki K, Yatera K, Morimoto Y, 2020. Predictive Biomarkers for the Ranking of Pulmonary Toxicity of Nanomaterials. Nanomaterials 10, 2032. https://doi.org/10.3390/nano10102032
[120] Jagiello K, Halappanavar S, Rybińska-Fryca A, Willliams A, Vogel U, Puzyn T, 2021. Transcriptomics-Based and AOP-Informed Structure–Activity Relationships to Predict Pulmonary Pathology Induced by Multiwalled Carbon Nanotubes. Small 17. https://doi.org/10.1002/smll.202003465
[121] Merugu S, Jagiello K, Gajewicz-Skretna A, Halappanavar S, Willliams A, Vogel U, Puzyn T. The Impact of Carbon Nanotube Properties on Lung Pathologies and Atherosclerosis Through Acute Inflammation: a New AOP-Anchored in Silico NAM. Small n/a, 2501185. https://doi.org/https://doi.org/10.1002/smll.202501185
[122] Bornholdt J, Saber AT, Lilje B, Boyd M, Jørgensen M, Chen Y, Vitezic M, Jacobsen NR, Poulsen SS, Berthing T, Bressendorff S, Vitting-Seerup K, Andersson R, Hougaard KS, Yauk CL, Halappanavar S, Wallin H, Vogel U, Sandelin A, 2017. Identification of Gene Transcription Start Sites and Enhancers Responding to Pulmonary Carbon Nanotube Exposure *in Vivo*. ACS Nano 11, 3597–3613. https://doi.org/10.1021/acsnano.6b07533

[123] Merugu S, Jagiello K, Gajewicz-Skretna A, Halappanavar S, Willliams A, Vogel U, Puzyn T, 2025. The Impact of Carbon Nanotube Properties on Lung Pathologies and Atherosclerosis Through Acute Inflammation: a New AOP-Anchored *in Silico* NAM. Small 21. https://doi.org/10.1002/smll.202501185
[124] Luan F, Kleandrova V V., González-Díaz H, Ruso JM, Melo A, Speck-Planche A, Cordeiro MNDS, 2014. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. Nanoscale 6, 10623. https://doi.org/10.1039/C4NR01285B
[125] Kleandrova V V., Luan F, González-Díaz H, Ruso JM, Melo A, Speck-Planche A, Cordeiro MNDS, 2014. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. Environ Int 73, 288–294. https://doi.org/10.1016/j.envint.2014.08.009
[126] Kleandrova V V., Luan F, González-Díaz H, Ruso JM, Speck-Planche A, Cordeiro MNDS, 2014. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. Environ Sci Technol 48, 14686–14694. https://doi.org/10.1021/es503861x
[127] Alves VM, Bobrowski T, Melo-Filho CC, Korn D, Auerbach S, Schmitt C, Muratov EN, Tropsha A, 2021. QSAR Modeling of SARS-CoV M ^{pro} Inhibitors Identifies Sufugolix, Cenicriviroc, Proglumetacin, and other Drugs as Candidates for Repurposing against SARS-CoV-2. Mol Inform 40. https://doi.org/10.1002/minf.202000113
[128] Khan K, Kumar V, Colombo E, Lombardo A, Benfenati E, Roy K, 2022. Intelligent consensus predictions of bioconcentration factor of pharmaceuticals using 2D and fragment-based descriptors. Environ Int 170, 107625. https://doi.org/10.1016/j.envint.2022.107625
[129] Kuz'min V, Artemenko A, Ognichenko L, Hromov A, Kosinskaya A, Stelmakh S, Sessions ZL, Muratov EN, 2021. Simplex representation of molecular structure as universal QSAR/QSPR tool. Struct Chem 32, 1365–1392. https://doi.org/10.1007/s11224-021-01793-z
[130] Nesterkina M, Muratov V, Ognichenko L, Kravchenko I, Kuz'min V, 2021. Quantitative structure–activity relationship study on prolonged anticonvulsant activity of terpene derivatives in pentylenetetrazole test. Open Chem 19, 1184–1192. https://doi.org/10.1515/chem-2021-0108
[131] (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure − Activity Relationship models, predictions, and results based on multiple predictions, 2023
[132] Poulsen SS, Saber AT, Williams A, Andersen O, Købler C, Atluri R, Pozzebon ME, Mucelli SP, Simion M, Rickerby D, Mortensen A, Jackson P, Kyjovska ZO, Mølhave K, Jacobsen NR, Jensen KA, Yauk CL, Wallin H, Halappanavar S, Vogel U, 2015. MWCNTs of different physicochemical properties cause similar inflammatory responses, but differences in transcriptional and histological markers of fibrosis in mouse lungs. Toxicol Appl Pharmacol 284, 16–32. https://doi.org/10.1016/j.taap.2014.12.011
[133] Halappanavar S, Rahman L, Nikota J, Poulsen SS, Ding Y, Jackson P, Wallin H, Schmid O, Vogel U, Williams A, 2019. Ranking of nanomaterial potency to induce pathway perturbations associated with lung responses. NanoImpact 14, 100158. https://doi.org/10.1016/j.impact.2019.100158
[134] Saber A, Jacobsen N, Mortensen A, Szarek J, Jackson P, Madsen A, Jensen K, Koponen IK, Brunborg G, Gützkow K, Vogel U, Wallin H, 2012. Nanotitanium dioxide toxicity in mouse lung is reduced in sanding dust from paint. Part Fibre Toxicol 9, 4. https://doi.org/10.1186/1743-8977-9-4

[135] Saber A, Mortensen A, Szarek J, Jacobsen N, Levin M, Koponen I, Jensen K, Vogel U, Wallin H, 2019. Toxicity of pristine and paint-embedded TiO_2 nanomaterials. Hum Exp Toxicol 38, 11–24. https://doi.org/10.1177/0960327118774910
[136] Wallin H, Kyjovska ZO, Poulsen SS, Jacobsen NR, Saber AT, Bengtson S, Jackson P, Vogel U, 2017. Surface modification does not influence the genotoxic and inflammatory effects of TiO_2 nanoparticles after pulmonary exposure by instillation in mice. Mutagenesis 32, 47–57. https://doi.org/10.1093/mutage/gew046
[137] McCarthy DJ, Smyth GK, 2009. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics 25, 765–771. https://doi.org/10.1093/bioinformatics/btp053
[138] Peart MJ, Smyth GK, van Laar RK, Bowtell DD, Richon VM, Marks PA, Holloway AJ, Johnstone RW, 2005. Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. Proceedings of the National Academy of Sciences 102, 3697–3702. https://doi.org/10.1073/pnas.0500369102
[139] Raouf A, Zhao Y, To K, Stingl J, Delaney A, Barbara M, Iscove N, Jones S, McKinney S, Emerman J, Aparicio S, Marra M, Eaves C, 2008. Transcriptome Analysis of the Normal Human Mammary Cell Commitment and Differentiation Process. Cell Stem Cell 3, 109–118. https://doi.org/10.1016/j.stem.2008.05.018
[140] Fortin F-A, Marc-André Gardner U, Parizeau M, Gagné C, 2012. DEAP: Evolutionary Algorithms Made Easy François-Michel De Rainville
[141] Sengottiyan S, Mikolajczyk A, Jagiełło K, Swirog M, Puzyn T, 2023. Core, Coating, or Corona? The Importance of Considering Protein Coronas in nano-QSPR Modeling of Zeta Potential. ACS Nano 17, 1989–1997. https://doi.org/10.1021/acsnano.2c06977
[142] Li R, Wang X, Ji Z, Sun B, Zhang H, Chang CH, Lin S, Meng H, Liao Y-P, Wang M, Li Z, Hwang AA, Song T-B, Xu R, Yang Y, Zink JI, Nel AE, Xia T, 2013. Surface Charge and Cellular Processing of Covalently Functionalized Multiwall Carbon Nanotubes Determine Pulmonary Toxicity. ACS Nano 7, 2352–2368. https://doi.org/10.1021/nn305567s
[143] Jang M-H, Hwang YS, 2018. Effects of functionalized multi-walled carbon nanotubes on toxicity and bioaccumulation of lead in Daphnia magna. PLoS One 13, e0194935. https://doi.org/10.1371/journal.pone.0194935
[144] Zeinabad HA, Zarrabian A, Saboury AA, Alizadeh AM, Falahati M, 2016. Interaction of single and multi wall carbon nanotubes with the biological systems: tau protein and PC12 cells as targets. Sci Rep 6, 26508. https://doi.org/10.1038/srep26508
[145] Lönnecke K, Eberhardt O, Wallmersperger T, 2023. Electrostatic charge distribution in armchair and zigzag carbon nanotubes: a numerical comparison of CNT charge models. Acta Mech 234, 1–16. https://doi.org/10.1007/s00707-021-03085-3
[146] Cai Z, Zafferani M, Akande OM, Hargrove AE, 2022. Quantitative Structure-Activity Relationship (QSAR) Study Predicts Small-Molecule Binding to RNA Structure. J Med Chem 65, 7262–7277. https://doi.org/10.1021/acs.jmedchem.2c00254
[147] Huchthausen J, Braasch J, Escher BI, König M, Henneberger L, 2024. Effects of Chemicals in Reporter Gene Bioassays with Different Metabolic Activities Compared to Baseline Toxicity. Chem Res Toxicol 37, 744–756. https://doi.org/10.1021/acs.chemrestox.4c00017
[148] Sasaki H, Sato H, Kuriyama-Matsumura K, Sato K, Maebara K, Wang H, Tamba M, Itoh K, Yamamoto M, Bannai S, 2002. Electrophile Response Element-mediated Induction of the Cystine/Glutamate Exchange Transporter

Gene Expression. Journal of Biological Chemistry 277, 44765–44771. https://doi.org/10.1074/jbc.M208704200
[149] Muench M, Hsin C-H, Ferber E, Berger S, Mueller MJ, 2016. Reactive electrophilic oxylipins trigger a heat stress-like response through HSFA1 transcription factors. J Exp Bot 67, 6139–6148. https://doi.org/10.1093/jxb/erw376
[150] Rossi EM, Pylkkänen L, Koivisto AJ, Vippola M, Jensen KA, Miettinen M, Sirola K, Nykäsenoja H, Karisola P, Stjernvall T, Vanhala E, Kiilunen M, Pasanen P, Mäkinen M, Hämeri K, Joutsensaari J, Tuomi T, Jokiniemi J, Wolff H, Savolainen K, Matikainen S, Alenius H, 2010. Airway Exposure to Silica-Coated TiO2 Nanoparticles Induces Pulmonary Neutrophilia in Mice. Toxicological Sciences 113, 422–433. https://doi.org/10.1093/toxsci/kfp254
[151] Lee S, Weiss T, Bühler M, Mena J, Lottenbach Z, Wegmann R, Sun M, Bihl M, Augustynek B, Baumann SP, Goetze S, van Drogen A, Pedrioli PGA, Penton D, Festl Y, Buck A, Kirschenbaum D, Zeitlberger AM, Neidert MC, Vasella F, Rushing EJ, Wollscheid B, Hediger MA, Weller M, Snijder B, 2024. High-throughput identification of repurposable neuroactive drugs with potent anti-glioblastoma activity. Nat Med 30, 3196–3208. https://doi.org/10.1038/s41591-024-03224-y
[152] Paul Friedman K, Thomas RS, Wambaugh JF, Harrill JA, Judson RS, Shafer TJ, Williams AJ, Lee J-YJ, Loo L-H, Gagné M, Long AS, Barton-Maclaren TS, Whelan M, Bouhifd M, Rasenberg M, Simanainen U, Sobanski T, 2025. Integration of new approach methods for the assessment of data-poor chemicals. Toxicological Sciences 205, 74–105. https://doi.org/10.1093/toxsci/kfaf019
[153] European Chemicals Agency., 2016. New approach methodologies in regulatory science – Proceedings of a scientific workshop – Helsinki,19–20 April 2016. ECHA
[154] Muratov V, Jagiello K, Mikolajczyk A, Danielsen PH, Halappanavar S, Vogel U, Puzyn T, 2025. The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers. J Hazard Mater 493, 138240. https://doi.org/10.1016/j.jhazmat.2025.138240
[155] Kumar S, Mickael C, Kumar R, Prasad RR, Campbell N V, Zhang H, Li M, McKeon BA, Allen TE, Graham BB, Yu Y-RA, Stenmark KR, 2024. Single cell transcriptomic analyses reveal diverse and dynamic changes of distinct populations of lung interstitial macrophages in hypoxia-induced pulmonary hypertension. Front Immunol 15, 1372959. https://doi.org/10.3389/fimmu.2024.1372959
[156] Perez L, Ambroise J, Bearzatto B, Yakoub Y, Palmai-Pallag M, Bouzin C, Ryelandt L, Pavan C, Huaux F, Lison D, 2025. Transcriptomic profile of TiO2 particle overload in rat alveolar macrophages: examining responses between P25 and two pigment grade particles in vitro. Arch Toxicol. https://doi.org/10.1007/s00204-025-04152-7
[157] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA, 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11, 733–9. https://doi.org/10.1038/nrg2825
[158] Yu Y, Mai Y, Zheng Y, Shi L, 2024. Assessing and mitigating batch effects in large-scale omics studies. Genome Biol 25, 254. https://doi.org/10.1186/s13059-024-03401-9
[159] Serra A, Fratello M, Cattelani L, Liampa I, Melagraki G, Kohonen P, Nymark P, Federico A, Kinaret PAS, Jagiello K, Ha MK, Choi J-S, Sanabria N, Gulumian M, Puzyn T, Yoon T-H, Sarimveis H, Grafström R, Afantitis A, Greco D, 2020. Transcriptomics in Toxicogenomics, Part III: Data Modelling

for Risk Assessment. Nanomaterials 10, 708. https://doi.org/10.3390/nano10040708
[160] Langfelder P, Horvath S, 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. https://doi.org/10.1186/1471-2105-9-559
[161] Bhuva DD, Cursons J, Davis MJ, 2020. Stable gene expression for normalisation and single-sample scoring. Nucleic Acids Res 48, e113–e113. https://doi.org/10.1093/nar/gkaa802
[162] Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ, 2018. Single sample scoring of molecular phenotypes. BMC Bioinformatics 19, 404. https://doi.org/10.1186/s12859-018-2435-4
[163] Hänzelmann S, Castelo R, Guinney J, 2013. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 14, 7. https://doi.org/10.1186/1471-2105-14-7
[164] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP, 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102
[165] Muratov V, Jagiełło K, Puzyn T, 2025. TRIumph in Nanotoxicology: Simplifying Transcriptomics into a Single Predictive Variable. Nanoscale Horiz. https://doi.org/10.1039/D5NH00330J
[166] Husain M, Saber AT, Guo C, Jacobsen NR, Jensen KA, Yauk CL, Williams A, Vogel U, Wallin H, Halappanavar S, 2013. Pulmonary instillation of low doses of titanium dioxide nanoparticles in mice leads to particle retention and gene expression changes in the absence of inflammation. Toxicol Appl Pharmacol 269, 250–262. https://doi.org/10.1016/j.taap.2013.03.018
[167] Danielsen PH, Poulsen SS, Knudsen KB, Clausen PA, Jensen KA, Wallin H, Vogel U, 2024. Physicochemical properties of 26 carbon nanotubes as predictors for pulmonary inflammation and acute phase response in mice following intratracheal lung exposure. Environ Toxicol Pharmacol 107, 104413. https://doi.org/10.1016/j.etap.2024.104413
[168] Hadrup N, Zhernovkov V, Jacobsen NR, Voss C, Strunz M, Ansari M, Schiller HB, Halappanavar S, Poulsen SS, Kholodenko B, Stoeger T, Saber AT, Vogel U, 2020. Acute Phase Response as a Biological Mechanism-of-Action of (Nano)particle-Induced Cardiovascular Disease. Small 16. https://doi.org/10.1002/smll.201907476
[169] Saber AT, Jacobsen NR, Jackson P, Poulsen SS, Kyjovska ZO, Halappanavar S, Yauk CL, Wallin H, Vogel U, 2014. Particle-induced pulmonary acute phase response may be the causal link between particle inhalation and cardiovascular disease. WIREs Nanomedicine and Nanobiotechnology 6, 517–531. https://doi.org/10.1002/wnan.1279
[170] Ishida Y, Kimura A, Kondo T, Hayashi T, Ueno M, Takakura N, Matsushima K, Mukaida N, 2007. Essential Roles of the CC Chemokine Ligand 3-CC Chemokine Receptor 5 Axis in Bleomycin-Induced Pulmonary Fibrosis through Regulation of Macrophage and Fibrocyte Infiltration. Am J Pathol 170, 843–854. https://doi.org/10.2353/ajpath.2007.051213
[171] Li S, Yang P, 2021. Relationship between HSPA1A-regulated gene expression and alternative splicing in mouse cardiomyocytes and cardiac hypertrophy. J Thorac Dis 13, 5517–5533. https://doi.org/10.21037/jtd-21-1222
[172] Perez L, Ambroise J, Bearzatto B, Froidure A, Pilette C, Yakoub Y, Palmai-Pallag M, Bouzin C, Ryelandt L, Pavan C, Huaux F, Lison D, 2025. Unique transcriptomic responses of rat and human alveolar macrophages in an in vitro model of overload with TiO2 and carbon black. Part Fibre Toxicol 22, 8. https://doi.org/10.1186/s12989-025-00624-x

[173] Chézeau L, Sébillaud S, Safar R, Seidel C, Dembélé D, Lorcin M, Langlais C, Grossmann S, Nunge H, Michaux S, Dubois-Pot-Schneider H, Rihn B, Joubert O, Binet S, Cosnier F, Gaté L, 2018. Short- and long-term gene expression profiles induced by inhaled TiO2 nanostructured aerosol in rat lung. Toxicol Appl Pharmacol 356, 54–64. https://doi.org/10.1016/j.taap.2018.07.013
[174] Sager TM, Umbright CM, Mustafa GM, Roberts JR, Orandle MS, Cumpston JL, McKinney WG, Boots T, Kashon ML, Joseph P, 2022. Pulmonary toxicity and gene expression changes in response to whole-body inhalation exposure to multi-walled carbon nanotubes in rats. Inhal Toxicol 34, 200–218. https://doi.org/10.1080/08958378.2022.2081386
[175] Guo C, Wright MD, Buckley A, Laycock A, Berthing T, Vogel U, Cosnier F, Gaté L, Leonard MO, Smith R, 2025. Pulmonary Toxicity of Long, Thick MWCNT and Very Long, Thin Carboxylated MWCNT Aerosols Following 28 Days Whole-Body Exposure. Toxics 13, 401. https://doi.org/10.3390/toxics13050401
[176] Valentino SA, Chézeau L, Seidel C, Sébillaud S, Lorcin M, Chalansonnet M, Cosnier F, Gaté L, 2021. Exposure to TiO2 Nanostructured Aerosol Induces Specific Gene Expression Profile Modifications in the Lungs of Young and Elderly Rats. Nanomaterials (Basel) 11. https://doi.org/10.3390/nano11061466
[177] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P, 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111, 1361–1375. https://doi.org/10.1289/ehp.5758
[178] Mikolajczyk A, Sizochenko N, Mulkiewicz E, Malankowska A, Nischk M, Jurczak P, Hirano S, Nowaczyk G, Zaleska-Medynska A, Leszczynski J, Gajewicz A, Puzyn T, 2017. Evaluating the toxicity of TiO_2 -based nanoparticles to Chinese hamster ovary cells and *Escherichia coli:* a complementary experimental and computational approach. Beilstein Journal of Nanotechnology 8, 2171–2180. https://doi.org/10.3762/bjnano.8.216
[179] Fresegna AM, Ursini CL, Ciervo A, Maiello R, Casciardi S, Iavicoli S, Cavallo D, 2021. Assessment of the Influence of Crystalline Form on Cyto-Genotoxic and Inflammatory Effects Induced by TiO2 Nanoparticles on Human Bronchial and Alveolar Cells. Nanomaterials (Basel) 11. https://doi.org/10.3390/nano11010253
[180] Okuda-Shimazaki J, Takaku S, Kanehira K, Sonezaki S, Taniguchi A, 2010. Effects of Titanium Dioxide Nanoparticle Aggregate Size on Gene Expression. Int J Mol Sci 11, 2383–2392. https://doi.org/10.3390/ijms11062383
[181] Suthar JK, Rakesh B, Vaidya A, Ravindran S, 2023. Comprehensive Analysis of Titanium Oxide Nanoparticle Size and Surface Properties on Neuronal PC-12 Cells: Unraveling Cytotoxicity, Dopaminergic Gene Expression, and Acetylcholinesterase Inhibition. J Xenobiot 13, 662–684. https://doi.org/10.3390/jox13040043
[182] Mano SS, Kanehira K, Sonezaki S, Taniguchi A, 2012. Effect of polyethylene glycol modification of TiO_2nanoparticles on cytotoxicity and gene expressions in human cell lines. Int J Mol Sci 13, 3703–3717. https://doi.org/10.3390/ijms13033703
[183] El-Said KS, Ali EM, Kanehira K, Taniguchi A, 2014. Molecular mechanism of DNA damage induced by titanium dioxide nanoparticles in toll-like receptor 3 or 4 expressing human hepatocarcinoma cell lines. J Nanobiotechnology 12, 48. https://doi.org/10.1186/s12951-014-0048-2
[184] Li W, Jia MX, Deng J, Wang JH, Zuberi Z, Yang S, Ba J, Chen Z, 2020. MicroRNA Response and Toxicity of Potential Pathways in Human Colon

Cancer Cells Exposed to Titanium Dioxide Nanoparticles. Cancers (Basel) 12, 1236. https://doi.org/10.3390/cancers12051236
[185] Fuster E, Candela H, Estévez J, Vilanova E, Sogorb MA, 2021. Titanium Dioxide, but Not Zinc Oxide, Nanoparticles Cause Severe Transcriptomic Alterations in T98G Human Glioblastoma Cells. Int J Mol Sci 22. https://doi.org/10.3390/ijms22042084
[186] Jayaram DT, Kumar A, Kippner LE, Ho P-Y, Kemp ML, Fan Y, Payne CK, 2019. TiO2 nanoparticles generate superoxide and alter gene expression in human lung cells. RSC Adv 9, 25039–25047. https://doi.org/10.1039/c9ra04037d
[187] Alswady-Hoff M, Erdem JS, Phuyal S, Knittelfelder O, Sharma A, Fonseca D de M, Skare Ø, Slupphaug G, Zienolddiny S, 2021. Long-Term Exposure to Nanosized TiO2 Triggers Stress Responses and Cell Death Pathways in Pulmonary Epithelial Cells. Int J Mol Sci 22. https://doi.org/10.3390/ijms22105349
[188] Domingo MG, Kurtz M, Maglione G, Martin M, Brites F, Tasat DR, Olmedo DG, 2024. Chronic exposure to <scp> TiO $_2$ </scp> micro- and nano particles: A biochemical and histopathological experimental study. J Biomed Mater Res B Appl Biomater 112. https://doi.org/10.1002/jbm.b.35443
[189] Califf RM, 2018. Biomarker definitions and their applications. Exp Biol Med 243, 213–221. https://doi.org/10.1177/1535370217750088
[190] Ebner JN, 2021. Trends in the Application of “Omics” to Ecotoxicology and Stress Ecology. Genes (Basel) 12, 1481. https://doi.org/10.3390/genes12101481
[191] Fortino V, Kinaret PAS, Fratello M, Serra A, Saarimäki LA, Gallud A, Gupta G, Vales G, Correia M, Rasool O, Ytterberg J, Monopoli M, Skoog T, Ritchie P, Moya S, Vázquez-Campos S, Handy R, Grafström R, Tran L, Zubarev R, Lahesmaa R, Dawson K, Loeschner K, Larsen EH, Krombach F, Norppa H, Kere J, Savolainen K, Alenius H, Fadeel B, Greco D, 2022. Biomarkers of nanomaterials hazard from multi-layer data. Nat Commun 13, 3798. https://doi.org/10.1038/s41467-022-31609-5
[192] Xu H, Lam SH, Shen Y, Gong Z, 2013. Genome-Wide Identification of Molecular Pathways and Biomarkers in Response to Arsenic Exposure in Zebrafish Liver. PLoS One 8, e68737. https://doi.org/10.1371/journal.pone.0068737
[193] Li Y, Huang H, Li Y, Ye Z, Li X, Liu K, Liu M, Liu L, Jiang J, 2025. Characterizing soil COPs eco-risk in China. J Hazard Mater 489, 137588. https://doi.org/10.1016/j.jhazmat.2025.137588
[194] Danielsen PH, Knudsen KB, Štrancar J, Umek P, Koklič T, Garvas M, Vanhala E, Savukoski S, Ding Y, Madsen AM, Jacobsen NR, Weydahl IK, Berthing T, Poulsen SS, Schmid O, Wolff H, Vogel U, 2020. Effects of physicochemical properties of TiO2 nanomaterials for pulmonary inflammation, acute phase response and alveolar proteinosis in intratracheally exposed mice. Toxicol Appl Pharmacol 386, 114830. https://doi.org/10.1016/j.taap.2019.114830
[195] Gutierrez CT, Loizides C, Hafez I, Brostrøm A, Wolff H, Szarek J, Berthing T, Mortensen A, Jensen KA, Roursgaard M, Saber AT, Møller P, Biskos G, Vogel U, 2023. Acute phase response following pulmonary exposure to soluble and insoluble metal oxide nanomaterials in mice. Part Fibre Toxicol 20, 4. https://doi.org/10.1186/s12989-023-00514-0
[196] Franklin RJ, Mohana, Dabbagol V, 2020. Anomaly Detection in Videos for Video Surveillance Applications using Neural Networks. In: 2020 Fourth International Conference on Inventive Systems and Control (ICISC). IEEE, pp 632–637
[197] Guéniche N, Huguet A, Bruyere A, Habauzit D, Le Hégarat L, Fardel O, 2021. Comparative in silico prediction of P-glycoprotein-mediated transport for

2010–2020 US FDA-approved drugs using six Web-tools. Biopharm Drug Dispos 42, 393–398. https://doi.org/10.1002/bdd.2299
[198] Hosseini M, Ramazani J, 2016. Evaluation of Acute Physiology and Chronic Health Evaluation II and sequential organ failure assessment scoring systems for prognostication of outcomes among Intensive Care Unit's patients. Saudi J Anaesth 10, 168. https://doi.org/10.4103/1658-354X.168817
[199] Ling Z, Liu X, Luo Y, Wu X, Yuan L, Tong X, Li L, Xiang C, 2013. Associations between Vaginal Pathogenic Community and Bacterial Vaginosis in Chinese Reproductive-Age Women. PLoS One 8, e76589. https://doi.org/10.1371/journal.pone.0076589
[200] Fusco R, Siracusa R, D'Amico R, Cordaro M, Genovese T, Gugliandolo E, Peritore AF, Crupi R, Di Paola R, Cuzzocrea S, Impellizzeri D, 2020. Mucosa-Associated Lymphoid Tissue Lymphoma Translocation 1 Inhibitor as a Novel Therapeutic Tool for Lung Injury. Int J Mol Sci 21, 7761. https://doi.org/10.3390/ijms21207761
[201] Groves DT, Jiang Y, 1995. Chemokines, a Family of Chemotactic Cytokines. Critical Reviews in Oral Biology & Medicine 6, 109–118. https://doi.org/10.1177/10454411950060020101
[202] McCarrick S, Malmborg V, Gren L, Danielsen PH, Tunér M, Palmberg L, Broberg K, Pagels J, Vogel U, Gliga AR, 2025. Pulmonary exposure to renewable diesel exhaust particles alters protein expression and toxicity profiles in bronchoalveolar lavage fluid and plasma of mice. Arch Toxicol 99, 797–814. https://doi.org/10.1007/s00204-024-03915-y
[203] Schmid O, Stoeger T, 2016. Surface area is the biologically most effective dose metric for acute nanoparticle toxicity in the lung. J Aerosol Sci 99, 133–143. https://doi.org/10.1016/j.jaerosci.2015.12.006
[204] Hoshyar N, Gray S, Han H, Bao G, 2016. The Effect of Nanoparticle Size on *In Vivo* Pharmacokinetics and Cellular Interaction. Nanomedicine 11, 673–692. https://doi.org/10.2217/nnm.16.5
[205] Wang M, Li J, Dong S, Cai X, Simaiti A, Yang X, Zhu X, Luo J, Jiang L-H, Du B, Yu P, Yang W, 2020. Silica nanoparticles induce lung inflammation in mice via ROS/PARP/TRPM2 signaling-mediated lysosome impairment and autophagy dysfunction. Part Fibre Toxicol 17, 23. https://doi.org/10.1186/s12989-020-00353-3
[206] Forest V, Pourchez J, 2017. Preferential binding of positive nanoparticles on cell membranes is due to electrostatic interactions: A too simplistic explanation that does not take into account the nanoparticle protein corona. Materials Science and Engineering: C 70, 889–896. https://doi.org/10.1016/j.msec.2016.09.016
[207] Bonner JC, 2010. Nanoparticles as a Potential Cause of Pleural and Interstitial Lung Disease. Proc Am Thorac Soc 7, 138–141. https://doi.org/10.1513/pats.200907-061RM
[208] Dhruba SR, Rahman A, Rahman R, Ghosh S, Pal R, 2019. Recursive model for dose-time responses in pharmacological studies. BMC Bioinformatics 20, 317. https://doi.org/10.1186/s12859-019-2831-4
[209] Pandey A, Mishra AK, 2022. Immunomodulation, Toxicity, and Therapeutic Potential of Nanoparticles. BioTech 11, 42. https://doi.org/10.3390/biotech11030042
[210] Shvedova AA, Yanamala N, Kisin ER, Tkach A V., Murray AR, Hubbs A, Chirila MM, Keohavong P, Sycheva LP, Kagan VE, Castranova V, 2014. Long-term effects of carbon containing engineered nanomaterials and asbestos in the lung: one year postexposure comparisons. American Journal of Physiology-Lung Cellular and Molecular Physiology 306, L170–L182. https://doi.org/10.1152/ajplung.00167.2013

[211] Gutierrez C, Poulsen S, Sørli J, Wallin H, Halappanavar S, Yauk C, Vogel U, 2025. Substance interaction with lung resident cell membrane components leading to atherosclerosis. AOP-Wiki. aopwiki.org/aops/237

6. Co-author declarations

University of Gdańsk

Doktorant
Pracownia Chemoinformatyki Środowiska
Wydział Chemii
Mgr Viacheslav Muratov

Gdańsk, 30.07.2026 r.

DECLARATION

I declare that I am a co-author of the following publication included in my doctoral dissertation:

1. **TRIumph in nanotoxicology: simplifying transcriptomics into a single predictive variable**
   Authors: Viacheslav Muratov, Karolina Jagiello, Tomasz Puzyn
   Published: Nanoscale Horizons, Volume 10, 3116-3126, 2025, MniSW points – 200, IF = 6,4
   My contributions to the above article were writing – original draft, validation, methodology, investigation, formal analysis, data curation, and conceptualization.

2. **An integrative omics-machine learning framework for predicting pulmonary responses to titanium dioxide nanoparticles**
   Authors: Viacheslav Muratov, Karolina Jagiello, Tomasz Puzyn
   Published: Nanoscale, Volume 18, 7021-7033, 2026, MniSW points – 140, IF = 5,2
   My contributions to the above article were writing – original draft, validation, methodology, investigation, formal analysis, data curation, and conceptualization.

3. **The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers**
   Authors: Viacheslav Muratov, Karolina Jagiello, Alicja Mikolajczyk, Pernille Høgh Danielsen, Sabina Halappanavar, Ulla Vogel, Tomasz Puzyn
   Published: Journal of Hazardous Materials, Volume 493, 5 August 2025, 138240, MniSW points – 200, IF = 11,3
   My contributions to the above article were writing – original draft, validation, methodology, investigation, formal analysis, data curation, conceptualization.

Viacheslav Muratov

Faculty of Chemistry
Department of Environmental Chemistry and Radiochemistry

ul. Wita Stwosza 63
80-308 Gdańsk, Poland

tel. +48 58 523 52 50,
e-mail:
viacheslav.muratov@phdstud.ug.edu.pl

www.en.ug.edu.pl
www.qsar.eu.org

Fahrenheit Universities | sea·EU

University of Gdańsk

Kierownik Pracowni
Pracownia Chemoinformatyki Środowiska
Wydział Chemii
Professor dr hab. Tomasz Puzyn

Gdańsk, 24.06.2026 r.

DECLARATION

I declare that I am a co-author of the following publication included in the doctoral dissertation of Mgr Viacheslav Muratov:

1. **TRIumph in nanotoxicology: simplifying transcriptomics into a single predictive variable**
   Authors: Viacheslav Muratov, Karolina Jagiello, Tomasz Puzyn
   Published: Nanoscale Horizons, Volume 10, 3116-3126, 2025, MniSW points – 200, IF = 6,4
   My contribution to the above article was writing – review & editing, funding acquisition, and project administration.

2. **An integrative omics-machine learning framework for predicting pulmonary responses to titanium dioxide nanoparticles**
   Authors: Viacheslav Muratov, Karolina Jagiello, Tomasz Puzyn
   Published: Nanoscale, Volume 18, 7021-7033, 2026, MniSW points – 140, IF = 5,2
   My contribution to the above article was writing – review & editing, funding acquisition, and project administration.

3. **The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers**
   Authors: Viacheslav Muratov, Karolina Jagiello, Alicja Mikolajczyk, Pernille Høgh Danielsen, Sabina Halappanavar, Ulla Vogel, Tomasz Puzyn
   Published: Journal of Hazardous Materials, Volume 493, 5 August 2025, 138240, MniSW points – 200, IF = 11,3
   My contributions to the above article were writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, and Conceptualization.

KIEROWNIK
Katedra Chemii i Radiochemii
Środowiska
prof. dr hab. Tomasz Puzyn

Faculty of Chemistry
Department of Environmental
Chemistry and Radiochemistry

ul. Wita Stwosza 63
80-308 Gdańsk, Poland

tel. +48 58 523 52 48,
e-mail:
tomasz.puzyn@ug.edu.pl

www.en.ug.edu.pl
www.qsar.eu.org

Fahrenheit Universities | sea·EU

University of Gdańsk

Adiunkt
Pracownia Chemoinformatyki Środowiska
Wydział Chemii
dr inż. Karolina Jagiełło

Gdańsk, 24.06.2026 r.

DECLARATION

I declare that I am a co-author of the following publication included in the doctoral dissertation of Mgr Viacheslav Muratov:

1. **TRIumph in nanotoxicology: simplifying transcriptomics into a single predictive variable**
   Authors: Viacheslav Muratov, Karolina Jagiello, Tomasz Puzyn
   Published: Nanoscale Horizons, Volume 10, 3116-3126, 2025, MniSW points – 200, IF = 6,4
   My contribution to the above article was conceptualization, writing – review & editing, funding acquisition, and project administration.

2. **An integrative omics-machine learning framework for predicting pulmonary responses to titanium dioxide nanoparticles**
   Authors: Viacheslav Muratov, Karolina Jagiello, Tomasz Puzyn
   Published: Nanoscale, Volume 18, 7021-7033, 2026, MniSW points – 140, IF = 5,2
   My contribution to the above article was conceptualization, writing – review & editing, funding acquisition, and project administration.

3. **The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers**
   Authors: Viacheslav Muratov, Karolina Jagiello, Alicja Mikolajczyk, Pernille Høgh Danielsen, Sabina Halappanavar, Ulla Vogel, Tomasz Puzyn
   Published: Journal of Hazardous Materials, Volume 493, 5 August 2025, 138240, MniSW points – 200, IF = 11,3
   My contributions to the above article were Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, and Conceptualization.

Karolina Jagiełło

Faculty of Chemistry
Department of Environmental Chemistry and Radiochemistry

ul. Wita Stwosza 63
80-308 Gdańsk, Poland

tel. +48 58 523 52 47,
e-mail:
karolina.jagiello@ug.edu.pl

www.en.ug.edu.pl
www.qsar.eu.org

Fahrenheit Universities | sea-EU

University of Gdańsk

Adiunkt
Pracownia Chemoinformatyki Środowiska
Wydział Chemii
dr. Alicja Mikołajczyk

Gdańsk, 24.06.2026 r.

DECLARATION

I declare that I am a co-author of the following publication included in the doctoral dissertation of Mgr Viacheslav Muratov:

The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers

Authors: Viacheslav Muratov, Karolina Jagiello, Alicja Mikolajczyk, Pernille Høgh Danielsen, Sabina Halappanavar, Ulla Vogel, Tomasz Puzyn

Published: Journal of Hazardous Materials, Volume 493, 5 August 2025, 138240, MniSW points – 200, IF = 11,3

My contribution to the above article was writing – review & editing, validation.

Faculty of Chemistry
Department of Environmental Chemistry and Radiochemistry

ul. Wita Stwosza 63
80-308 Gdańsk, Poland

tel. +48 58 523 52 47,
e-mail:
alicja.mikolajczyk@ug.edu.pl

www.en.ug.edu.pl
www.qsar.eu.org

Fahrenheit Universities | sea·EU

Dr. Danielsen Pernille Høgh
The National Research Centre for the Working Environment
Copenhagen DK-2100, Denmark
ped@nfa.dk

Copenhagen, Denmark, date 12. June 2026

DECLARATION

I declare that I am a co-author of the following publication included in the doctoral dissertation of Mgr Viacheslav Muratov:

The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers

Authors: Viacheslav Muratov, Karolina Jagiello, Alicja Mikolajczyk, Pernille Høgh Danielsen, Sabina Halappanavar, Ulla Vogel, Tomasz Puzyn

Published: Journal of Hazardous Materials, Volume 493, 5 August 2025, 138240, MniSW points – 200, IF = 11,3

My contribution to the above article was Writing – review & editing.

Pernille Danielsen

Copenhagen, Denmark, June 12, 2026

Professor Ulla Vogel
The National Research Centre for the Working Environment
Copenhagen DK-2100, Denmark
ubv@nfa.dk

DECLARATION

I declare that I am a co-author of the following publication included in the doctoral dissertation of Mgr Viacheslav Muratov:

The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers

Authors: Viacheslav Muratov, Karolina Jagiello, Alicja Mikolajczyk, Pernille Høgh Danielsen, Sabina Halappanavar, Ulla Vogel, Tomasz Puzyn

Published: Journal of Hazardous Materials, Volume 493, 5 August 2025, 138240, MniSW points – 200, IF = 11,3

My contribution to the above article was Writing – review & editing, Methodology, Data curation.

Kind regards,



Ulla Vogel
Professor, National Research Centre for the Working Environment

June 12, 2026

Sabina Halappanavar, Ph.D.
Environmental Health Science and Research Bureau
Health Canada
Ottawa, Ontario K1A 0K9, Canada
sabina.halappanavar@hc-sc.gc.ca

DECLARATION

I declare that I am a co-author of the following publication included in the doctoral dissertation of Mgr Viacheslav Muratov:

The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers

Authors: Viacheslav Muratov, Karolina Jagiello, Alicja Mikolajczyk, Pernille Høgh Danielsen, Sabina Halappanavar, Ulla Vogel, Tomasz Puzyn

Published: Journal of Hazardous Materials, Volume 493, 5 August 2025, 138240, MniSW points – 200, IF = 11,3

My contribution to the above article was Writing – review & editing, Methodology, Investigation.

halappanavar, sabina
Digitally signed by halappanavar, sabina
Date: 2026.06.12 08:11:28 -04'00'

7. Publications included in the dissertation

Nanoscale Horizons

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal

Check for updates

TRIumph in nanotoxicology: simplifying transcriptomics into a single predictive variable

Viacheslav Muratov, [a] Karolina Jagiello [*ab] and Tomasz Puzyn*[ab]

Cite this: DOI: 10.1039/d5nh00330j

Received 12th May 2025,
Accepted 26th August 2025

DOI: 10.1039/d5nh00330j

rsc.li/nanoscale-horizons

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

The primary aim of our study was to address the problem of transcriptomic data complexity by introducing a novel transcriptomic response index (TRI), compressing the entire transcriptomic space into a single variable, and linking it with the inhaled multiwalled carbon nanotubes (MWCNTs) properties. This methodology allows us to predict fold change values of thousands of differentially expressed genes (DEGs) using a single variable and a single quantitative structure–activity relationship (QSAR) model. In the context of this work, TRI compressed 5167 DEGs into a single variable, explaining 99.9% of the entire transcriptomic space. Further TRI was linked to the properties of inhaled MWCNTs using a nano-QSAR model with statistics R^2 = 0.83, Q_{CV}^2 = 0.8, and Q^2 = 0.78, which show a high level of goodness-of-fit, robustness, and predictability of the obtained model. By training a nano-QSAR model on fold changes of thousands of DEGs using a single variable, our study significantly contributes not only to new approach methodologies (NAMs) focused on reducing animal testing but also decreases the amount of computational resources needed for work with complex transcriptomic data. Developed during this work, the software called ChemBioML Platform (https://chembioml.com) offers researchers a powerful free-to-use tool for training regulatory acceptable machine learning (ML) models without a strong background in programming. The ChemBioML Platform integrates the ML capabilities of Python with the advanced graphical interface of unreal engine 5, creating a bridge between scientific research and the game development industry.

New concepts

We introduce the transcriptomic response index (TRI) – a novel, biologically meaningful metric that compresses complex, high-dimensional gene expression data into a single predictive variable. This approach represents a conceptual breakthrough in the field of nanotoxicogenomics, enabling a direct, quantitative link between the structural properties of nanomaterials and global transcriptomic effects in exposed organisms. In contrast to existing models that either rely on isolated biomarkers or attempt to predict thousands of individual gene responses separately, TRI captures the entire transcriptomic variance (99.9%) in one interpretable parameter. This allows us to construct a robust, regulatorily compliant nano-QSAR model based on standard molecular descriptors (SiRMS), making transcriptomics scalable, reproducible, and predictive. This work uniquely bridges systems biology, chemoinformatics, and machine learning, offering a new paradigm for evaluating nanomaterial-induced biological perturbations. It further advances the principles of new approach methodologies (NAMs) by enabling mechanism-based hazard prediction without the need for animal testing or deep computational resources. The TRI concept opens new directions in nanoinformatics, particularly for safe-by-design strategies, high-throughput screening, and regulatory decision-making in nanotechnology.

Introduction

Given the structural similarities between multiwalled carbon nanotubes (MWCNTs) and asbestos fibers, coupled with their bio-persistence in lung tissue, concerns have been raised about the potential health effects of MWCNT exposure.[1] MWCNTs have been demonstrated to exhibit significant toxicity, with numerous studies indicating adverse health outcomes, particularly in the pulmonary system.[2,3] Inhalation of MWCNTs has been shown to cause sustained inflammation, fibrosis, and lung cancer.[2,3] The fibrogenic activity of MWCNTs has been linked to direct effects on fibroblasts, promoting their proliferation, differentiation, and collagen production.[2] Additionally, in animal studies, inhalation of MWCNTs has led to fetal malformations, fetal loss, behavioral changes in offspring, and delays in litter delivery.[3]

Although the adverse effects of MWCNTs are well-documented, the molecular mechanisms driving these effects remain largely unknown. In recent studies, researchers have started applying toxicogenomic approaches to close this gap and investigate the pulmonary toxicity of MWCNTs.[4–8] It was shown that MWCNTs can induce gene expression changes associated with lung fibrosis and inflammation, comparable to those caused by known fibrogenic agents like bleomycin and bacterial infections.[4] Recent research highlights the potential of gene expression profiling in

[a] University of Gdansk, Faculty of Chemistry, Laboratory of Environmental Chemoinformatics, Wita Stwosza 63, 80-308 Gdansk, Poland. E-mail: karolina.jagiello@ug.edu.pl, tomasz.puzyn@ug.edu.pl
[b] QSAR Lab Ltd., Trzy lipy 3, 80-172 Gdansk, Poland

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

predicting the pulmonary toxicity of new nanomaterials,[4,5] offering a promising methodology for future research and risk assessment. Building on this genomic insight, advanced computational methods have emerged as powerful tools for correlating material properties with biological outcomes. For instance, Jagiello *et al.* established a methodology for constructing quantitative structure–activity relationship (QSAR) models based on canonical pathways. The authors proposed a grouping strategy for MWCNTs based on the aspect ratio affecting genes associated with specific pathways.[6] Following this, Merugu *et al.* developed a QSAR model to predict canonical pathway induced by MWCNTs and triggering lung fibrosis and atherosclerosis, both initiated through acute immune response.[7] Fortino *et al.* highlight the importance of these findings because MWCNTs cause pulmonary damage in the body after intratracheal instillation.[8]

In parallel, perturbation-based machine learning (PTML) modeling has gained traction as a complementary strategy for toxicity prediction. PTML offers a high degree of interpretability and supports the simultaneous modeling of multiple toxicological endpoints across diverse biological systems, including cells, microbes, and animal models. This flexibility is especially advantageous in nanotoxicology, where experimental conditions often vary and biological responses are heterogeneous.[9–12] By integrating diverse datasets, PTML enables multi-endpoint predictions that improve both the efficiency of hazard assessments and the prioritization of materials for further testing. Moreover, its potential to support regulatory decision-making through transparent and mechanistically informed predictions positions PTML as a valuable asset in the next generation of nanotoxicological research.

The objective of our study is to investigate the potential of combining supervised learning methods, such as ridge regression, with unsupervised methods, like principal component analysis (PCA), to generate comprehensive and accurate predictions for the thousands of differentially expressed genes (DEGs) affected by MWCNT exposure in mouse lungs. In contrast to previous approaches, such as the work by Jagiello *et al.* and Merugu *et al.*, which linked MWCNT properties with the benchmark dose lower confidence limit (BMDL) of a single transcriptomic pathway,[6,7] here, we introduce a novel transcriptomic response index (TRI) derived from the fold change (FC) values of thousands of DEGs. By compressing the high-dimensional transcriptomic space into a single, informative variable, TRI measures all gene perturbations induced by MWCNT exposure. Biologically, a higher TRI value reflects a more pronounced deviation from baseline gene expression, which may correlate with the activation of key toxicological processes such as inflammation, fibrosis, and oxidative stress. We introduce a two-tiered analytical strategy to calculate TRI and link it to the exposure conditions. First, unsupervised PCA compresses the complex transcriptomic data into a comprehensive index (TRI) that describes overall gene expression changes. Subsequently, a supervised ridge regression model is used to correlate the structural properties of the MWCNTs, quantified using simplex representation of molecular structure (SiRMS) descriptors, with the TRI. Although SiRMS descriptors are traditionally applied to small molecules,[13–17] they have also been successfully adapted to predict metal oxide toxicity.[18] Our study is the first to leverage the full version of SiRMS descriptors to predict the transcriptomic response to MWCNT exposure in mouse lungs.

To ensure the reliability and robustness of our predictive models, we developed the ChemBioML Platform – an integrated machine learning software that features robust data preprocessing, a self-optimizing algorithm for selecting optimal descriptors and hyperparameters, and OECD-compatible validation techniques.[19] This platform is designed to generate accurate, reproducible results that satisfy regulatory and academic standards for QSAR and other ML model evaluations. Thus, our work advances scientific knowledge on the molecular mechanisms underlying MWCNT-induced lung toxicity and offers a practical tool for the academic community. This manuscript and the accompanying GitHub repository (**https://chembioml.com**, **https://github.com/chembiodev/ChemBioML-Platform**) provide a robust platform for training ML models that can elucidate the complex molecular interactions following MWCNT exposure.

Experimental

Data collection

The lung transcriptomic data from mice exposed to MWCNTs with varying properties (*e.g.*, size, functionalization) have been previously published.[20,21] As this study uses an already published dataset, we recommend referring to the original publications for detailed experimental procedures. Here, only a brief summary is provided. Gene expression profiles were obtained from lung tissues of 5–7-week-old female C57BL/6 mice exposed intratracheally to nine different MWCNTs with varying properties. Samples were collected on one post-exposure day. Each MWCNT was administered at doses of 18 μg per mouse and 54 μg per mouse; additionally, NM-401 and NRCWE-26 were tested at a higher dose of 162 μg per mouse. Table 1 summarizes the names and key characteristics of the MWCNTs studied (*e.g.*, size, concentration). We have used structurally modified MWCNTs with three different groups ($-NH_2$, –COOH, –OH), in addition to pristine MWCNTs. However, it is important to note that no additional external coatings or surface layers were applied to the MWCNTs in this work; thus, they should be regarded as uncoated nanomaterials. Comprehensive details on sample collection, animal exposure, and transcriptomic analysis are available in the original studies.[20–25] The exposure levels were selected based on the Danish occupational exposure limit of 3.5 mg m^{-3} for Printex90 carbon black particles.[26]

To create a dataset for our study, we have filtered differentially expressed genes (DEGs) from two studies[20,21] with the criteria of p-value ≤ 0.05 and an absolute FC ≥ 1.5.[21,22,27–30] DEGs that changed significantly in response to at least one MWCNT were filtered. Although a stricter threshold, such as $|FC| \geq 2$, has been used in other studies,[31,32] we opted for $|FC| \geq 1.5$ to remain consistent with the analysis approach used in the original datasets.[20,21] This ensured methodological alignment with the initial design of the transcriptomic data. This process

 This journal is © The Royal Society of Chemistry 2025

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

Table 1 MWCNT types, their physicochemical characteristics, and exposure conditions

#	MWCNT name	Concentration (μg per mouse)	Length (nm)	Diameter (nm)	Aspect ratio	Functionalization
NP_1	NM-401	18	4050 (±2400)	67 (±26.2)	61	Pristine
NP_2	NM-401	54	4050 (±2400)	67 (±26.2)	61	Pristine
NP_3	NM-401	162	4050 (±2400)	67 (±26.2)	61	Pristine
NP_4	NRCWE-26	18	850 (±457)	11 (±4.5)	77	Pristine
NP_5	NRCWE-26	54	850 (±457)	11 (±4.5)	77	Pristine
NP_6	NRCWE-26	162	850 (±457)	11 (±4.5)	77	Pristine
NP_7	NRCWE-043	18	771.3 (±3471)	26.73 (±6.88)	29	Pristine
NP_8	NRCWE-043	54	771.3 (±3471)	26.73 (±6.88)	29	Pristine
NP_9	NRCWE-044	18	1330 (±2454)	32.55 (±14.44)	41	–OH
NP_10	NRCWE-044	54	1330 (±2454)	32.55 (±14.44)	41	–OH
NP_11	NRCWE-045	18	1553 (±2954)	28.07 (±13.85)	55	–COOH
NP_12	NRCWE-045	54	1553 (±2954)	28.07 (±13.85)	55	–COOH
NP_13	NRCWE-046	18	717.2 (±1214)	17.22 (±5.77)	42	Pristine
NP_14	NRCWE-046	54	717.2 (±1214)	17.22 (±5.77)	42	Pristine
NP_15	NRCWE-047	18	532.5 (±591.9)	12.96 (±4.44)	41	–OH
NP_16	NRCWE-047	54	532.5 (±591.9)	12.96 (±4.44)	41	–OH
NP_17	NRCWE-048	18	1604 (±5609)	15.08 (±4.69)	106	–COOH
NP_18	NRCWE-048	54	1604 (±5609)	15.08 (±4.69)	106	–COOH
NP_19	NRCWE-049	18	731.1 (±1473)	13.85 (±6.09)	53	$-NH_2$
NP_20	NRCWE-049	54	731.1 (±1473)	13.85 (±6.09)	53	$-NH_2$

resulted in a curated dataset containing 5167 DEGs with post-exposure expression data. For each of the 9 MWCNTs, gene expression was measured under 2 or 3 dosing conditions – 18 μg per mouse and 54 μg per mouse for all MWCNTs, with NM-401 and NRCWE-26 additionally tested at 162 μg per mouse – yielding a total of 20 distinct experimental conditions (see Table S1).

Unsupervised machine learning methods applied to TRI calculation

To reduce the dimensionality of the dataset containing all DEGs, we applied principal component analysis (PCA).[33,34] This method transforms correlated genes into a set of orthogonal variables, known as principal components (PCs), which are ranked by the amount of variance they capture from the data. The first principal component (PC1) captures the greatest variance in the data, with each subsequent component capturing progressively less. This process simplifies high-dimensional data, making it easier to visualize and analyze.[35] After performing PCA on the transcriptomic dataset, we obtained a set of PCs that capture independent dimensions of variability in the gene expression data. Each PC represents a distinct pattern of transcriptomic response, with the first few PCs explaining the majority of variance and subsequent PCs capturing more subtle yet potentially informative variations. By weighting each PC by its explained variance and summing over all PCs, the TRI integrates both dominant and nuanced expression patterns into a single composite metric. The calculation follows eqn (1), where n represents the total number of PCs. This weighted approach provides insight into the significance of each PC in determining the overall transcriptomic response.

$$\mathrm{TRI} = \mathrm{PC1} \times \mathrm{Variance}_1 + \mathrm{PC2} \times \mathrm{Variance}_2 + \cdots + \mathrm{PC}n \times \mathrm{Variance}_n \quad (1)$$

Descriptors for carbon-based nanomaterials

In this study, we employed the SiRMS method to numerically characterize the investigated MWCNTs. SiRMS is a chemoinformatic tool widely applied in QSAR modeling to describe and analyze molecular structures, focusing strongly on stereochemistry. It represents molecules' assets of simplex descriptors – tetratomic fragments defined by fixed composition, structure, chirality, and symmetry – effectively capturing the chemical properties of atoms and their interactions.[13–18] Integrating diverse physicochemical properties of atoms, such as electronic charge and lipophilicity, enhances the interpretability of QSAR models and provides comprehensive insights into molecular interactions after MWCNTs exposure.[36–38]

For our calculations, we applied the open-source SiRMS Python package, available on GitHub [**https://github.com/DrrDom/sirms**]. This package allows weighting descriptors by various physicochemical properties, including electronic charge, lipophilicity (LogP), refractivity, and hydrogen bonding.[36–38] A concentration-dependent weighting of the simplex descriptors was applied to account for the varying concentrations of MWCNTs in the experiments (18 μg, 54 μg, and 162 μg per mouse). Specifically, we used the 18 μg dose as a baseline for concentration-dependent weighting. Since 54 μg is three times 18 μg and 162 μg is nine times 18 μg, the corresponding descriptors were multiplied by factors of 3 and 9, respectively. This concentration-weighting approach integrates the concentration of inhaled MWCNTs into the molecular descriptors, ensuring that they reflect the dose-dependent variations in the organism's reaction.

Supervised machine learning methods for linking TRI with MWCNTs' structural properties

To ensure robust model performance, we have split the set of 20 data points from Table 1 into training and external validation sets. Four data points were randomly selected and excluded from the training set. These four data points represent the whole dataset, ensuring a reliable evaluation procedure. Further, these data points were used only to evaluate the external predictive performance of the model.

Ridge regressor was used with the genetic algorithm (GA) feature selection procedure to train the models predicting TRI.

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

Ridge is a linear regression method that applies L2 regularization during the training of the model, allowing us to use cross-correlated descriptors[39] and to use a highly dimensional space of independent variables[40] for modeling. Furthermore, the GA is a feature selection method that allows us to identify the best set of descriptors for training the nano-QSAR model.[41]

A comprehensive validation was performed to address possible overfitting, which can occur when training the model on such a small dataset. The goodness-of-fit of the models was evaluated using the coefficient of determination (R^2) and the root mean square error of calibration ($RMSE_C$). The robustness of the model was assessed through leave-one-out cross-validation (CV) using the coefficient of determination for CV (Q_{CV}^2) and root mean square error for CV ($RMSE_{CV}$). Predictive performance was validated through the external validation coefficient and the root mean square error of external validation ($RMSE_{EXT}$), together with the coefficient of determination for the external predictions (Q^2). The applicability domain of the models, defined by the range of responses and structural feature descriptors, was analyzed using the leverage approach and visualized with a Williams plot. This method is widely used in the evaluation of AD in QSAR modeling with well-defined, previously described algorithms.[42,43]

Results & discussion

Transcriptomic-based data

Our study mainly aimed to explore the possibilities of applying unsupervised and supervised ML methods to compress and predict expression changes for thousands of DEGs using a single nano-QSAR model. We started the analysis by combining datasets from previously published studies.[20,21] First, the datasets were merged, and the DEGs with gaps for some MWCNTs were removed. Subsequently, we selected gene expression profiles from lung tissues of female C57BL/6 mice (aged 5–7 weeks) that were intratracheally exposed to MWCNTs, retaining only those genes with a p-value ≤ 0.05 and an absolute FC ≥ 1.5 in at least one MWCNT condition.[21,22,27–30] By applying this filtering, we got a set of 5167 DEGs significantly affected by the MWCNTs exposure in the lungs (Table S1).

DEGs compression into a single TRI

To address the challenges posed by highly multidimensional and complex transcriptomic data comprising 5167 DEGs as in our study (Table S1), we developed the TRI and then linked this index with the properties of MWCNTs *via* a single nanoQSAR model. Traditional computational techniques typically train separate models for each gene, treating each DEG as a separate endpoint.[44] While feasible for smaller gene sets, this approach becomes computationally expensive when applied to datasets involving thousands of genes. In this context, TRI provides a single endpoint containing all the extractable information about the contributing DEGs.

The dimensionality reduction achieved by consolidating 5167 gene-specific endpoints into a single TRI variable substantially streamlines the modeling process. Rather than training thousands of individual nano-QSAR models, one for each DEG, the TRI approach enables the training of a single, integrated model. Beyond computational efficiency, the TRI approach enhances interpretability and generalizability. By summarizing the global transcriptomic response into a compact form, the TRI is better suited for predictive modeling and can be readily integrated with other omics data or phenotypic outcomes. This systems-level representation is especially advantageous in contexts where understanding coordinated gene activity is more informative than isolated gene-level effects.

To compute TRI, first, the PCA was applied to the dataset of 5167 DEGs. The calculated values of all principal components and the corresponding TRI scores are provided in Table S2. The first 19 PCs (PC1 through PC19) were selected for TRI calculation, as they collectively explain ~99.9% of the total variance in the over 5000 gene expression matrix. This indicates that they capture nearly the entire structure of the dataset, making them a robust foundation for dimensionality reduction and feature integration.

In contrast, retaining only the first two PCs captures just ~43% of the total variance, which would lead to substantial information loss and potentially weaken model performance. Nevertheless, PC1 and PC2 were utilized to construct a simplified two-dimensional "transcriptomic space" to visualize and explore similarities between different MWCNTs based on DEG fold-change patterns (Fig. 1). This 2D representation provides an intuitive overview of the underlying transcriptomic relationships, while the full set of 19 PCs ensures high-fidelity data representation for TRI construction.

Examination of Fig. 1 reveals that most inhaled MWCNTs cluster in the negative regions of both PCs, indicating a high degree of similarity. However, differences are evident; PC2 predominantly reflects the variance associated with DEGs influenced by pristine MWCNTs. In contrast, PC1 exhibits a cluster comprising multiple MWCNT variants with low aspect ratios, including those with –OH and –COOH functionalization and pristine forms. This distribution indicates that using either PC1 or PC2 alone as the endpoint for the nano-QSAR model is insufficient, as neither fully captures the transcriptomic response of all studied MWCNTs. To overcome this limitation, we computed the TRI, which integrates 19 principal components weighted by their respective explained variance. This approach ensures that TRI captures the multidimensional structure of the data, providing a more complete and biologically meaningful representation of global transcriptomic shifts across all studied MWCNTs.

TRI is a highly interpretable and reversible PCA-based index. As a linear combination of gene expression values, TRI enables the prediction of global transcriptomic shifts while maintaining the ability to trace individual gene contributions. Specifically, genes with positive loadings (Table S3) will be upregulated as TRI increases, whereas those with negative loadings will be downregulated. This interpretability makes TRI not only transparent but also biologically informative.

In addition to offering high interpretability through its linear structure and transparent formulation, the TRI is also straightforward to validate. Although unsupervised methods

 This journal is © The Royal Society of Chemistry 2025

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

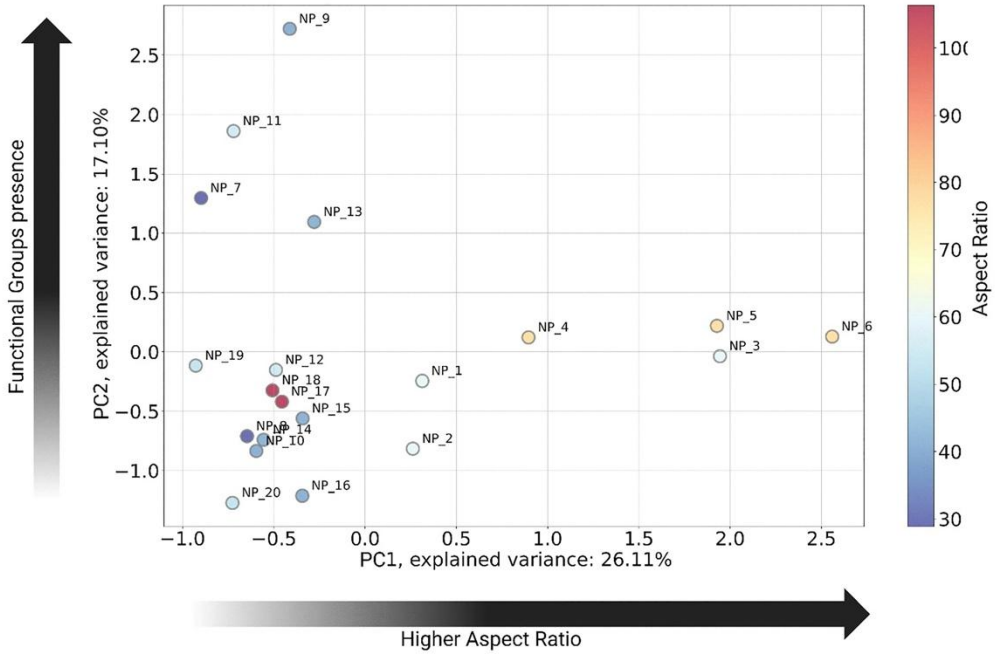


Fig. 1 Two-dimensional transcriptomic space, represented as a line map PC1:PC2.

based on principal component analysis (PCA) inherently lack explicit quantitative performance metrics such as R^2 or RMSE, TRI benefits from robust qualitative validation. In our study, this included careful preprocessing of the transcriptomic dataset by filtering genes based on stringent significance thresholds (p-value ≤ 0.05 and |fold change| ≥ 1.5), ensuring the absence of outliers *via* inspection of the PC1–PC2 projection (Fig. 1), and confirming that TRI effectively captures the transcriptomic variance observed across all relevant differentially expressed genes (Table S1). These steps collectively reinforce confidence that TRI accurately represents the underlying biological signal.

Nevertheless, we recognize that the ideal number and configuration of principal components may differ depending on biological context, dataset heterogeneity, and modeling goals. The fixed use of 19 components in this study was tailored to balance comprehensiveness with interpretability. However, for broader applications, such as large-scale, multi-condition experiments or cross-species comparisons, further refinement of the TRI framework may enhance its performance and generalizability. Potential improvements could include exploring alternative dimensionality reduction approaches (*e.g.*, independent component analysis, factor analysis), integrating non-linear manifold learning techniques such as t-SNE or UMAP, or employing data-driven component selection based on cross-validated model performance. Such refinements would not only optimize model efficiency but also strengthen TRI's utility as a generalizable tool for high-dimensional transcriptomic analysis.

Besides recognized limitations, we introduced the TRI approach as a novel framework for integrating chemical structure descriptors with transcriptomic data. It is important to contextualize how this approach compares with existing methods in the field. While traditional QSAR models can be effective for specific endpoints (Lethal Dose 50, *e.g.*), they often struggle to capture complex biological effects. They can suffer from well-known QSAR limitations (*e.g.*, limited applicability domain and the "activity cliff" problem).[45] On the other end of the spectrum, recent advanced methods leverage high-dimensional transcriptomic data directly. For example, deep learning frameworks have been developed to predict entire gene expression profiles from chemical structures.[46,47] Such methods (*e.g.*, CIGER and DeepCOP) use graph neural networks or other architectures to forecast genome-wide expression changes and have shown impressive performance on large datasets.[46,47] Our TRI approach offers a complementary strategy: instead of predicting thousands of individual gene responses, we condense the transcriptomic effect into an interpretable index and model that index using molecular descriptors for the inhaled MWCNTs.

Properties of inhaled MWCNTs and SiRMS descriptors for the carbon-based nanomaterials

Motivated by extensive research underscoring the impact of MWCNTs' physicochemical properties on toxicity and the need for further investigation into their mode of action (MoA), we focused on calculating comprehensive descriptors for the

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

experimentally investigated nanotubes. We combined nine different MWCNTs, each tested at various concentrations, to generate a total of 20 data points (Table 1). The size of these MWCNTs varies from 532.5 to 4050 nm (length) and 11 to 67 nm (width). The studied MWCNTs have four different functionalization options – pristine (unmodified), –OH, –COOH, and $–NH_2$. Testing MWCNTs at concentrations from 18 μg to 162 μg per mouse increased the training set size, which is crucial for QSAR models, and enabled dose–response analysis. Dose–response modeling is essential in nanotoxicity studies,[48] and by integrating these data with the MWCNTs' properties, our approach enables quantification of the dose-dependent impact on the transcriptomic response. This combination promises a more comprehensive understanding of the MoA of pulmonary toxicity induced by inhaled MWCNTs compared to analyzing them separately.

SiRMS descriptors were calculated for the studied MWCNTs. This method is very commonly used for small molecules due to its ability to balance between the accuracy of predictions and the very comprehensive physicochemical interpretation of the obtained QSAR models.[13–17] The SiRMS methodology partitions the molecular structure into simplexes of 1 to 6 atoms, effectively capturing the molecule's fundamental structural elements. These simplexes represent essential features such as functional groups (*e.g.*, hydroxyl, carboxyl groups) and bonding environments (*e.g.*, single, double, and triple bonds). By decomposing the molecule in this manner, SiRMS enables identifying and quantifying specific chemical environments critical to the material's physico-chemical behavior. This detailed representation enhances the model's ability to predict molecular interactions with biological systems, thereby providing valuable insights into the MoA of pulmonary toxicity induced by inhaled MWCNTs.

Previously, SiRMS descriptors were calculated for metal-based engineered nanomaterials (ENMs) in a simplified form, utilizing a one-dimensional method that provided only a general overview of the structures.[18] The simplification in this case was necessitated by the uncertainty surrounding the crystal structure of the ENMs used.[18] In contrast, for MWCNTs, we do not face such structural limitations. This has allowed us to compute two-dimensional SiRMS descriptors, which offer significantly more profound insights into the studied structures.

Moreover, we leveraged the ability of the SiRMS methodology to weight descriptors by specific properties, including charge, LogP, refractivity, and hydrogen bonding, as detailed in the study published by the script's author.[36–38] Additionally, to account for the varying concentrations of MWCNTs used in our study, we implemented a concentration-dependent weighting of the simplex descriptors. Precisely, the number of simplexes represented in each descriptor was scaled proportionally to the amount of inhaled MWCNTs: for a concentration of 18 μg per mouse, the initial set of simplexes was used; for 54 μg per mouse, the descriptors were multiplied by a factor of three; and for 162 μg per mouse, they were scaled by a factor of nine. This weighting approach assumed that it reflects the proportional increase in exposure and thus enables the model to better account for dose-dependent biological responses. As a result, the complete set of approximately 3500 SiRMS descriptors calculated is provided in Table S4. These descriptors were subsequently utilized for training and interpreting the nano-QSAR model, which predicts changes in thousands of DEGs compressed in the TRI.

In this study, descriptor values for MWCNT exposures were linearly scaled based on concentration levels (18, 54, and 162 μg per mouse), under the assumption of a linear dose–response relationship. This approach was intended to approximate proportional biological effects corresponding to increasing concentrations. However, biological responses to nanomaterials can be nonlinear or exhibit threshold behaviors, and therefore, this linear scaling assumption requires future experimental validation. Further research should investigate nonlinear modeling approaches to assess the validity and limitations of this assumption.

Applying ML algorithms to link TRI to the MWCNTs' structural properties

Prediction of TRI using the structural and dose differences is one of the main steps of this work, allowing for a better understanding and describing the MoA of pulmonary pathologies induced by MWCNTs inhalation. QSAR is a well-established and regulatory recognized technique that enables the prediction of biological outcomes from the interaction of a body or a cell with certain chemicals, as well as with ENMs.[13,14,16,19,36–38,42,45,49–54] With proper validation and interpretation, QSAR can provide powerful insights into the MoA of different biological processes after interaction with chemicals or ENMs, which led us to use this method in our study. To obtain such models, we have integrated a multi-step modeling procedure into our developed software, ChemBioML Platform (**https://github.com/chembiodev/ChemBioML-Platform**). First, the data set of descriptors was pre-processed – we have removed all the descriptors with constant values. Conventional machine learning approaches, such as multiple linear regression (MLR), often force researchers to discard valuable descriptors when their cross-correlation exceeds a certain threshold (*e.g.*, $R > 0.8$). This can result in the loss of important predictive information. To overcome this limitation, we employed ridge regression, which leverages L2 regularization to effectively manage multicollinearity. By penalizing large coefficients rather than removing correlated descriptors entirely, ridge regression preserves more information, improving both the robustness and predictive power of the resulting model.[39] Due to this upgrade, we have left the descriptors with high statistical correlation values.

To prove the obtained model's proper predictability, we randomly selected four data points and moved them to an external validation set. These data points represent the whole set and are not used in model training or any other model optimization procedure. The rest of the 16 data points were used for training and internal validation of the model. Predictions of the training compounds were evaluated with R^2 and RMSEc metrics, and the model showed a good goodness-of-fit. The robustness of the model was evaluated with the same

 This journal is © The Royal Society of Chemistry 2025

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

Table 2 Statistics of the trained nano-QSAR model predicting TRI

Statistic	Value
R^2	0.83
$RMSE_C$	0.15
Q_{CV}^2	0.8
$RMSE_{CV}$	0.16
Q^2	0.78

metrics for CV and showed promising results. The Q^2 metric, an analog of R^2 for the training set, and $RMSE_{EXT}$ were calculated for the predictability evaluation. These statistics can be seen in Table 2. Overall assessment of them shows a good level of goodness-of-fit, robustness, and external predictability. It is important to note that not only does the absolute value of these metrics matter, but the difference between R^2 and Q^2 also mustn't exceed 0.3.[49,52] In the case of our nano-QSAR model, the difference is less than 0.1, which indicates that the model is not overfitted. A plot showing the distribution of predicted and observed values is shown in Fig. 2a. Although the model evaluation employed both external validation and CV, we acknowledge that the relatively small sample size (20 data points) may limit the generalizability of the findings. Further validation using larger, independent datasets is necessary to fully establish the model's predictive performance and mitigate overfitting risks.

Completing the pre-processing steps, we have applied a combination of GridSearch and genetic algorithm (GA) to search for the best ridge hyperparameters and descriptors to train the nano-QSAR model. As a result, we got the ridge regression model eqn (2), a linear regression equation linking the TRI to MWCNT structural properties with good goodness-of-fit, robustness, and predictability.

$$TRI = 0.53 \times CHARGE_{A\text{–}A\text{–}B\text{–}B\text{–}A\text{–}A} - 0.57 \times CHARGE_{B\text{–}B\text{–}B\text{–}C} + 0.37 \times REFRACTIVITY_{B.C} - 0.46 \times CHARGE_{A.A} \quad (2)$$

An essential criterion in QSAR model evaluation is the applicability domain (AD).[19] This method allows us to identify how similar the structures from validation or prediction are similar to those used for the model training. Also, it helps to identify the outliers from training or validation sets that are structurally similar but are predicted inaccurately. To define the AD for our nano-QSAR model, we constructed a Williams plot that displays both the structural differences of the MWCNTs *via* leverages and the accuracy of predictions *via* standardized residuals (Fig. 2b).

In Fig. 2b, we can see that all predictions for training and validation sets have absolute standardized residuals less than ± 3. This indicates that there are no significant outliers – points for which the model's predictions are incorrect, even if the structure is similar to others. One data point, NP_3, exhibits a leverage of $h = 0.85$ (Table S5), indicating that it is structurally distinct from the other MWCNTs in the training set. NP_3 corresponds to the largest MWCNT (NM-401) tested at 162 μg per mouse concentration. While this is not a significant issue, provided that the prediction is accurate, further investigation of other MWCNTs at similar concentrations could be valuable for advancing both experimental and computational nanotoxicity studies. Such an extension of the existing data will provide essential data for filling the gap in the applicability of models predicting the transcriptomic response to MWCNTs inhalation.

Thus, the developed software ChemBioML Platform successfully integrated essential capabilities required for training the modern QSAR model. It includes comprehensive data preprocessing, a GA-based feature selection algorithm, a hyperparameter optimizer, and a ridge regressor for training robust, well-predictive QSAR models. Applying L2 regularization addresses common limitations like predictor multivariance and high predictor space.[39,40] Furthermore, it evaluates the robustness and predictivity of the trained models by applying CV, external validation, and AD. Additionally, the ChemBioML Platform offers a graphical representation of the obtained models, as shown in Fig. 2, and demonstrates a significant optimization in combination with overall TRI methodology. A single model training, including feature selection and validation, took just 2 : 40 minutes with the use of an AMD Ryzen 5 3500U CPU (which is a usual processor equipped in office laptops). For example, if we were training models for each gene separately, it could take ~230 hours with the same computer, which showcases a strong decrease in the amount of computational resources needed for work with complex transcriptomic data.

In the nano-QSAR equation (eqn (2)), it is evident that descriptors weighted by charge exert the most significant impact on the model's predictions. This might be expected, as the charge is a well-known MWCNT parameter, affecting their toxicity.[55,56] Zeinabad *et al.* investigated the impact of the charge on the interaction between CNTs and biological systems such as cells and proteins, using the example of Tau protein. They hypothesized that the interaction mechanism depends on the charge distribution and protein surfaces.[57] Our study extends this understanding by demonstrating that the charge of MWCNTs may significantly influence biological outcomes following their inhalation. Furthermore, while "$CHARGE_{A\text{–}A\text{–}B\text{–}B\text{–}A\text{–}A}$" has a negative coefficient, the coefficients for other charge-weighted descriptors differ, suggesting that distinct charged regions contribute oppositely to the transcriptomic response. It is well known that the charge distribution on MWCNTs is uneven, higher in the tube ends.[58] Our model indicates that regions with differing charges can exert opposing effects on the transcriptomic response. Notably, the descriptors "$CHARGE_{A\text{–}A\text{–}B\text{–}B\text{–}A\text{–}A}$" and "$CHARGE_{A.A}$" predominantly characterize pristine MWCNTs, whereas "$CHARGE_{B\text{–}B\text{–}B\text{–}C}$" and "$REFRACTIVITY_{B.C}$" are more common in functionalized variants; no descriptor was found to be exclusive to any specific functional group (Table S5). This suggests that the possible impact of the functional groups, such as –OH, –COOH, and $-NH_2$ can be indirect, thereby affecting the charge and refractivity of the MWCNTs, which are crucial for the transcriptomic response. Importantly, in this study, we have used functionalized but uncoated MWCNTs. In the case of preparing nano-QSAR models for coated ENMs, an additional set of descriptors can be calculated by the Dragon software[43] or using the SiRMS

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

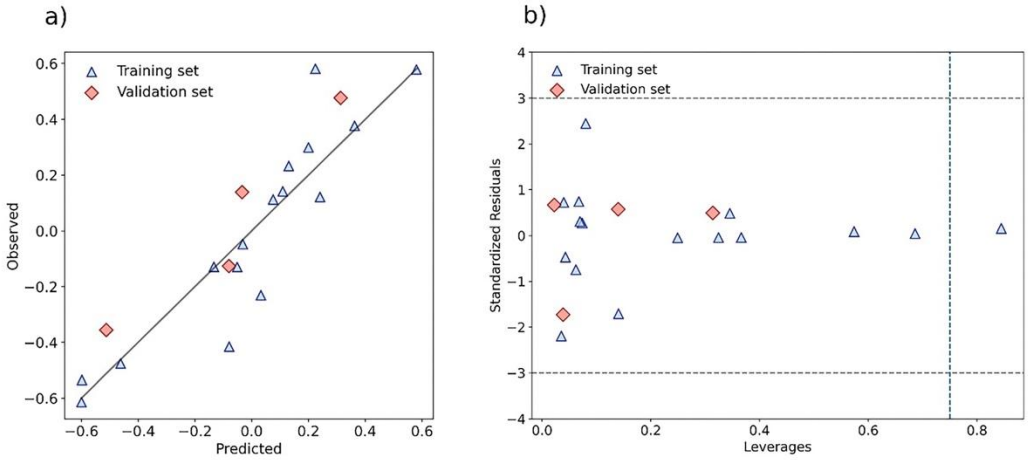


Fig. 2 (a) Observed-predicted plot presenting predictions' accuracy in two-dimensional space; (b) Williams plot presenting the model AD.

method. This approach would ensure that possible parameters of the additional molecules are counted during the modelling.

Thus, our results highlight that charge-weighted descriptors (*i.e.*, descriptors reflecting the distribution of partial charges in a structure) play a significant role in predicting transcriptomic responses. This finding suggests a mechanistic link between MWCNTs' electronic structure and the genes they perturb. One plausible explanation is that charge distribution influences how a molecule interacts with cellular components. For instance, highly electrophilic MWCNTs are known to activate cellular stress pathways. Literature examples support this: reactive electrophiles can trigger robust gene expression responses such as heat-shock protein induction[59] and the activation of detoxification genes *via* the Nrf2-mediated electrophile response element.[60] Our observation that charge-based descriptors are predictive aligns with these examples – molecules or materials bearing charged or polar functional groups may more readily form covalent bonds or induce oxidative stress, leading to upregulation of defensive transcripts. Conversely, compounds with certain charge features might have reduced cell permeability or distinct subcellular localization, affecting which genes are up- or down-regulated. For example, strong polar or ionic compounds may accumulate in specific cellular compartments or interact with membrane proteins, eliciting unique transcriptomic signatures. This notion is consistent with general toxicological principles: baseline cytotoxicity often arises from hydrophobic compounds integrating into membranes,[61] whereas specific gene responses often correlate with reactive functionalities. In line with this, our model's emphasis on charge-weighted descriptors echoes findings from other QSAR studies that identified surface charge properties as key determinants of bioactivity.[62]

In addition to the TRI combined from PC1–19, we have tested the predictability of the index calculated with PC1–5, PC1–10, and PC1–15. The graph shown in Fig. 3 represents that R^2 and $Q_{CV}{}^2$ significantly increased for the TRI combined from PC1–19. This showcases that in our study, the full TRI not only explains 99.9% of the transcriptomic space consisting of 5167 genes, but also gives the best opportunities for training nano-QSAR models predicting the transcriptomic responses to MWCNTs exposure in lungs.

Although the developed model provides a powerful framework for exploring the relationship between MWCNTs' characteristics and biological responses, it has limitations. One primary limitation of this study is the small sample size of compounds, which constrains the diversity of chemical space explored. A limited number of MWCNTs increases the risk of overfitting and reduces confidence that the observed structure–response relationships will hold broadly. Indeed, prior analyses have shown that using too few MWCNTs can yield unstable gene expression signatures and poor predictive performance.[63] Consequently, the AD of our model is narrow; the predictions may not generalize well beyond the MWCNTs represented in the training set.[62] In our study, we provide and analyze the AD using the Williams plot from Fig. 2b. For all new predictions, AD also has to be calculated based on the studied MWCNTs structures, and only the predictions inside AD can be considered as accurate and used for further analysis. This restricted domain makes the model highly specialized and sensitive to outliers and underscores the need for a cautious interpretation of the results. Future experiments should expand the number and chemical diversity dataset to address this limitation. For example, including a broader range of MWCNTs would help ensure that the learned relationships are not idiosyncratic to a small subset. Likewise, collecting more samples per nanotube (*e.g.*, across different concentrations) could improve the robustness of transcriptomic signals and reduce experimental variability. By explicitly acknowledging the sample size issue, we emphasize that further validation on larger, independent datasets is essential before the model's predictions can be

 This journal is © The Royal Society of Chemistry 2025

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

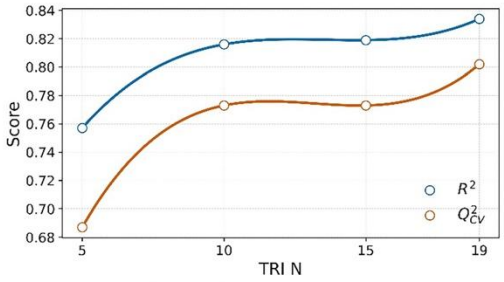


Fig. 3 Estimation of R^2 and $Q_{CV}{}^2$ for TRI combined from PC1–5 to PC1–19.

generalizable. This limitation is not unique to our work – it mirrors challenges in many toxicogenomic modeling studies where a paucity of data can undermine external predictivity.[62]

Conclusions

We have successfully developed the methodology that allows the compression of extensive transcriptomic data into a single endpoint, the so-called TRI, and linked it with the structural parameters of the studied MWCNTs. The trained nano-QSAR model can use TRI to predict overall transcriptomic patterns in response to MWCNTs inhalation with a very high level of predictability. Furthermore, our initial application of SiRMS descriptors in training nano-QSAR models for MWCNTs demonstrates significant potential for enhancing our understanding of the mode of action underlying interactions between inhaled MWCNTs and biological systems. Thus, the integration of PCA and ridge regression within the ChemBioML Platform showcases substantial potential for computational nano-toxicology applications. The proposed TRI showcases a strong potential for further research in the field of computational nanotoxicity, addressing the need to reduce animal testing[64] and computational resource usage in designing safe-by-design nanomaterials. In the meantime, the developed ChemBioML Platform (**https://chembioml.com, https://github.com/chembiodev/ChemBioML-Platform**) provides a powerful free-to-use tool for all researchers, thereby making ML research more accessible for laboratories without a strong programming background.

Author contributions

VM and KJ conceptualization. VM performed formal analysis, including data curation, the development of the model, and validation. VM designed and wrote the manuscript with the support of KJ. TP review and editing. TP and KJ funding acquisition and project administration. All authors discussed the results and commented on the manuscript. VM conceptualized, developed, and supports the ChemBioML Platform.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the SI. Supplementary information: Table S1. Fold change values of 5167 DEGs in response to 20 MWCNTs in different combinations. Table S2. Values of 19 PCs with calculated TRI and explained variance. Table S3 Loading values of 5167 DEGs in PC1–19. Table S4. Set of SiRMS descriptors calculated for 20 MWCNTs. Table S5. Model statistics and the selected SiRMS descriptors. See DOI: **https://doi.org/10.1039/d5nh00330j**

Acknowledgements

This work was funded *via* the National Science Centre in the frame of the TransNANO project (UMO-2020/37/B/ST5/01894). Figures for this work were prepared using the software **https://BioRender.com**. VM acknowledges Epic Games, Inc. for providing access to Unreal Engine 5 under their standard licensing terms.

Notes and references

1 S. S. Gupta, K. P. Singh, S. Gupta, M. Dusinska and Q. Rahman, *Nanomaterials*, 2022, **12**(10), 1708.
2 G. Vietti, D. Lison and S. van den Brule, *Part. Fibre Toxicol.*, 2016, **13**, 11.
3 N. Kobayashi, H. Izumi and Y. Morimoto, *J. Occup. Health*, 2017, **59**, 394–407.
4 J. Nikota, A. Williams, C. L. Yauk, H. Wallin, U. Vogel and S. Halappanavar, *Part. Fibre Toxicol.*, 2016, **13**, 25.
5 C. Nishida, H. Izumi, T. Tomonaga, J. Takeshita, K.-Y. Wang, K. Yamasaki, K. Yatera and Y. Morimoto, *Nanomaterials*, 2020, **10**, 2032.
6 K. Jagiello, S. Halappanavar, A. Rybińska-Fryca, A. Willliams, U. Vogel and T. Puzyn, *Small*, 2021, **17**(15), 2003465.
7 S. Merugu, K. Jagiello, A. Gajewicz-Skretna, S. Halappanavar, A. Willliams, U. Vogel and T. Puzyn, *Small*, 2025, 2501185.
8 V. Fortino, P. A. S. Kinaret and M. Fratello, *et al.*, *Nat Commun.*, 2022, **13**, 3798.
9 R. Concu, V. V. Kleandrova, A. Speck-Planche and M. N. D. S. Cordeiro, *Nanotoxicology*, 2017, **11**, 891–906.
10 F. Luan, V. V. Kleandrova, H. González-Díaz, J. M. Ruso, A. Melo, A. Speck-Planche and M. N. D. S. Cordeiro, *Nanoscale*, 2014, **6**, 10623.
11 V. V. Kleandrova, F. Luan, H. González-Díaz, J. M. Ruso, A. Melo, A. Speck-Planche and M. N. D. S. Cordeiro, *Environ. Int.*, 2014, **73**, 288–294.
12 V. V. Kleandrova, F. Luan, H. González-Díaz, J. M. Ruso, A. Speck-Planche and M. N. D. S. Cordeiro, *Environ. Sci. Technol.*, 2014, **48**, 14686–14694.
13 P. Polishchuk, E. Mokshyna, A. Kosinskaya, A. Muats, M. Kulinsky, O. Tinkov, L. Ognichenko, T. Khristova, A. Artemenko and V. Kuz'min, *Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure*, 2017 pp. 107–147.

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

14 V. M. Alves, T. Bobrowski, C. C. Melo-Filho, D. Korn, S. Auerbach, C. Schmitt, E. N. Muratov and A. Tropsha, *Mol. Inf.*, 2020, **40**(1), 2000113.
15 K. Khan, V. Kumar, E. Colombo, A. Lombardo, E. Benfenati and K. Roy, *Environ. Int.*, 2022, **170**, 107625.
16 V. Kuz'min, A. Artemenko, L. Ognichenko, A. Hromov, A. Kosinskaya, S. Stelmakh, Z. L. Sessions and E. N. Muratov, *Struct. Chem.*, 2021, **32**, 1365–1392.
17 M. Nesterkina, V. Muratov, L. Ognichenko, I. Kravchenko and V. Kuz'min, *Open Chem.*, 2021, **19**, 1184–1192.
18 V. E. Kuz'min, L. N. Ognichenko, N. Sizochenko, V. A. Chapkin, S. I. Stelmakh, A. O. Shyrykalova and J. Leszczynski, *Research Anthology on Synthesis, Characterization, and Applications of Nanomaterials*, IGI Global, 2021, pp. 317–329.
19 (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure – Activity Relationship models, predictions, and results based on multiple predictions, 2023.
20 S. S. Poulsen, A. T. Saber, A. Williams, O. Andersen, C. Købler, R. Atluri, M. E. Pozzebon, S. P. Mucelli, M. Simion, D. Rickerby, A. Mortensen, P. Jackson, Z. O. Kyjovska, K. Mølhave, N. R. Jacobsen, K. A. Jensen, C. L. Yauk, H. Wallin, S. Halappanavar and U. Vogel, *Toxicol. Appl. Pharmacol.*, 2015, **284**, 16–32.
21 S. Halappanavar, L. Rahman, J. Nikota, S. S. Poulsen, Y. Ding, P. Jackson, H. Wallin, O. Schmid, U. Vogel and A. Williams, *NanoImpact*, 2019, **14**, 100158.
22 S. Halappanavar, A. T. Saber, N. Decan, K. A. Jensen, D. Wu, N. R. Jacobsen, C. Guo, J. Rogowski, I. K. Koponen, M. Levin, A. M. Madsen, R. Atluri, V. Snitka, R. K. Birkedal, D. Rickerby, A. Williams, H. Wallin, C. L. Yauk and U. Vogel, *Environ. Mol. Mutagen.*, 2015, **56**, 245–264.
23 A. Saber, N. Jacobsen, A. Mortensen, J. Szarek, P. Jackson, A. Madsen, K. Jensen, I. K. Koponen, G. Brunborg, K. Gützkow, U. Vogel and H. Wallin, *Part. Fibre Toxicol.*, 2012, **9**, 4.
24 A. Saber, A. Mortensen, J. Szarek, N. Jacobsen, M. Levin, I. Koponen, K. Jensen, U. Vogel and H. Wallin, *Hum. Exp. Toxicol.*, 2019, **38**, 11–24.
25 H. Wallin, Z. O. Kyjovska, S. S. Poulsen, N. R. Jacobsen, A. T. Saber, S. Bengtson, P. Jackson and U. Vogel, *Mutagenesis*, 2017, **32**, 47–57.
26 J. A. Bourdon, A. T. Saber, N. R. Jacobsen, K. A. Jensen, A. M. Madsen, J. S. Lamson, H. Wallin, P. Møller, S. Loft, C. L. Yauk and U. B. Vogel, *Part. Fibre Toxicol.*, 2012, **9**, 5.
27 D. J. McCarthy and G. K. Smyth, *Bioinformatics*, 2009, **25**, 765–771.
28 M. J. Peart, G. K. Smyth, R. K. van Laar, D. D. Bowtell, V. M. Richon, P. A. Marks, A. J. Holloway and R. W. Johnstone, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 3697–3702.
29 A. Raouf, Y. Zhao, K. To, J. Stingl, A. Delaney, M. Barbara, N. Iscove, S. Jones, S. McKinney, J. Emerman, S. Aparicio, M. Marra and C. Eaves, *Cell Stem Cell*, 2008, **3**, 109–118.
30 S. Halappanavar, P. Jackson, A. Williams, K. A. Jensen, K. S. Hougaard, U. Vogel, C. L. Yauk and H. Wallin, *Environ. Mol. Mutagen.*, 2011, **52**, 425–439.
31 M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown and R. W. Davis, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 10614–10619.
32 Y. Zhao, D.-F. Guo and K. Rahmouni, *Physiology*, 2024, **39**(S1), DOI: **10.1152/physiol.2024.39.S1.881**.
33 J. Lever, M. Krzywinski and N. Altman, *Nat. Methods*, 2017, **14**, 641–642.
34 I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 2002.
35 D. Saha and A. Manickavasagan, *Curr. Res. Food Sci.*, 2021, **4**, 28–44.
36 P. Polishchuk, O. Tinkov, T. Khristova, L. Ognichenko, A. Kosinskaya, A. Varnek and V. Kuz'min, *J. Chem. Inf. Model.*, 2016, **56**, 1455–1469.
37 V. E. Kuz'min, A. G. Artemenko, P. G. Polischuk, E. N. Muratov, A. I. Hromov, A. V. Liahovskiy, S. A. Andronati and S. Y. Makan, *J. Mol. Model.*, 2005, **11**, 457–467.
38 V. E. Kuz'min, A. G. Artemenko and E. N. Muratov, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 403–421.
39 M. S. Mohammed, N. Paraman, A. A.-H. A. Rahman, F. A. Ghaleb, A. Al-Dhamari and M. N. Marsono, *IEEE Access*, 2021, **9**, 124087–124099.
40 D. Dai, F. Javed, P. Karlsson and K. Månsson, *Ann. Oper. Res.*, 2025, DOI: **10.1007/s10479-025-06486-y**.
41 F.-A. Fortin, U. Marc-André Gardner, M. Parizeau and C. Gagné, *DEAP: Evolutionary Algorithms Made Easy François-Michel De Rainville*, 2012, vol. 13.
42 A. O. Aptula, N. G. Jeliazkova, T. W. Schultz and M. T. D. Cronin, *QSAR Comb. Sci.*, 2005, **24**, 385–396.
43 S. Sengottiyan, A. Mikolajczyk, K. Jagiełło, M. Swirog and T. Puzyn, *ACS Nano*, 2023, **17**, 1989–1997.
44 V. Muratov, K. Jagiello, A. Mikolajczyk, P. H. Danielsen, S. Halappanavar, U. Vogel and T. Puzyn, *J. Hazard. Mater.*, 2025, **493**, 138240.
45 Q. Chen, L. Wu, W. Liu, L. Xing and X. Fan, *Molecules*, 2013, **18**, 10789–10801.
46 T.-H. Pham, Y. Qiu, J. Liu, S. Zimmer, E. O'Neill, L. Xie and P. Zhang, *Patterns*, 2022, **3**, 100441.
47 G. Woo, M. Fernandez, M. Hsing, N. A. Lack, A. D. Cavga and A. Cherkasov, *Bioinformatics*, 2020, **36**, 813–818.
48 I. Iavicoli, V. Leso, L. Fontana and E. Calabrese, *Int. J. Mol. Sci.*, 2018, **19**, 805.
49 L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.*, 2003, **111**, 1361–1375.
50 A. Tropsha, *Mol. Inf.*, 2010, **29**, 476–488.
51 V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, P. G. Polischuk, L. N. Ognichenko, A. V. Liahovsky, A. I. Hromov and E. V. Varlamova, *Virtual Screening and Molecular Design Based on Hierarchical Qsar Technology*, 2010, pp. 127–176.
52 T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T. P. Dasari, A. Michalkova, H.-M. Hwang, A. Toropov, D. Leszczynska and J. Leszczynski, *Nat. Nanotechnol.*, 2011, **6**, 175–178.
53 G. K. Jillella, P. K. Ojha and K. Roy, *Aquat. Toxicol.*, 2021, **238**, 105925.

 This journal is © The Royal Society of Chemistry 2025

54 L. N. Ognichenko, V. E. Kuz'min and A. G. Artemenko, *QSAR Comb. Sci.*, 2009, **28**, 939–945.
55 R. Li, X. Wang, Z. Ji, B. Sun, H. Zhang, C. H. Chang, S. Lin, H. Meng, Y.-P. Liao, M. Wang, Z. Li, A. A. Hwang, T.-B. Song, R. Xu, Y. Yang, J. I. Zink, A. E. Nel and T. Xia, *ACS Nano*, 2013, **7**, 2352–2368.
56 M.-H. Jang and Y. S. Hwang, *PLoS One*, 2018, **13**, e0194935.
57 H. A. Zeinabad, A. Zarrabian, A. A. Saboury, A. M. Alizadeh and M. Falahati, *Sci. Rep.*, 2016, **6**, 26508.
58 K. Lönnecke, O. Eberhardt and T. Wallmersperger, *Acta Mech.*, 2023, **234**, 1–16.
59 M. Muench, C.-H. Hsin, E. Ferber, S. Berger and M. J. Mueller, *J. Exp. Bot.*, 2016, **67**, 6139–6148.
60 H. Sasaki, H. Sato, K. Kuriyama-Matsumura, K. Sato, K. Maebara, H. Wang, M. Tamba, K. Itoh, M. Yamamoto and S. Bannai, *J. Biol. Chem.*, 2002, **277**, 44765–44771.
61 J. Huchthausen, J. Braasch, B. I. Escher, M. König and L. Henneberger, *Chem. Res. Toxicol.*, 2024, **37**, 744–756.
62 Z. Cai, M. Zafferani, O. M. Akande and A. E. Hargrove, *J. Med. Chem.*, 2022, **65**, 7262–7277.
63 C. Stretch, S. Khan, N. Asgarian, R. Eisner, S. Vaisipour, S. Damaraju, K. Graham, O. F. Bathe, H. Steed, R. Greiner and V. E. Baracos, *PLoS One*, 2013, **8**, e65380.
64 European Chemicals Agency., New approach methodologies in regulatory science – Proceedings of a scientific workshop – Helsinki,19–20 April 2016, ECHA, 2016.

Published on 28 August 2025. Downloaded on 9/3/2025 3:02:42 PM.

Nanoscale

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Nanoscale*, 2026, **18**, 7021

An integrative omics-machine learning framework for predicting pulmonary responses to titanium dioxide nanoparticles

Viacheslav Muratov, [a] Karolina Jagiello [*a,b] and Tomasz Puzyn[*a,b]

We address the challenge of leveraging complex lung transcriptomic data for nanosafety by re-optimizing the Transcriptomic Response Index (TRI) as a single, New Approach Methodology (NAM)-ready omics endpoint for instilled titanium dioxide nanoparticles (TiO_2-NPs). Using mouse lung gene-expression profiles after intratracheal instillation of five TiO_2-NPs across three doses and two post-exposure times (30 conditions), we compressed 621 differentially expressed genes (DEGs) into a one-variable index *via* Principal Component Analysis (PCA) and systematically evaluated TRI variants composed of principal components PC1–PC2 up to PC1–PC29 (TRI_2–TRI_{29}). The full composition (TRI_{29}) reconstructs ~99.9% of transcriptomic variance, while a compact TRI_2 maximized predictability when linked to exposure predictors *via* ridge regression with genetic algorithm-based feature selection (R^2 = 0.79; Q^2_{CV} = 0.70; Q^2 = 0.79). Thus, while the full-space TRI_{29} covers the whole transcriptomic variance, a compact TRI_2 captures ≈44% but yields the best predictive performance. Because TRI is built from PC loadings, it remains reversible to genes. Collectively, our results demonstrate that TRI is portable across different nanomaterials, but must be re-optimized per dataset to balance explained variance and predictive performance. TRI_2 provides a compact, interpretable, and externally validated bridge from TiO_2-NP attributes to system-level transcriptomic response. By replacing hundreds of gene-specific models with a single, validated endpoint, this workflow streamlines omics-driven screening, potency ranking, and study-internal extrapolation in regulatory-relevant NAM pipelines.

Received 4th November 2025,
Accepted 20th February 2026

DOI: 10.1039/d5nr04657b

rsc.li/nanoscale

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

Introduction

Engineered titanium dioxide nanoparticles (TiO_2-NPs) are widely used in consumer and industrial products, ranging from cosmetics and food additives to paints, coatings, and catalysts, raising plausible scenarios for inhalation exposure in occupational and environmental settings. Evidence from rodent inhalation studies demonstrates that high airborne concentrations and prolonged TiO_2-NP exposures can elicit acute neutrophilic inflammation with durable particle retention in the distal lung.[1,2] At sufficiently high lung burdens, persistence of TiO_2-NPs can promote maladaptive repair processes, contributing to chronic pathologies, such as fibrosis and, in some cases, tumorigenesis.[2,3] These findings underscore growing concern over the respiratory hazards posed by TiO_2-NPs and, more broadly, by other widely used engineered nanomaterials (ENMs).

Toxicogenomic approaches are increasingly leveraged to characterize nanomaterial hazards beyond conventional toxicology. High-throughput "omics" technologies, such as transcriptomics, capture gene-wide changes in gene expression, offering a comprehensive molecular fingerprint of biological responses to a substance.[4] In alignment with ongoing New Approach Methodologies (NAMs) initiatives,[5,6] transcriptomics is being integrated into multiple facets of hazard and risk assessment, including category formation, mode-of-action (MoA) elucidation, potency ranking, and derivation of transcriptomic-based points of departure (PoD).[6,7]

Importantly, transcriptomic responses to TiO_2-NPs inhalation are strongly influenced by particle attributes, including crystal phase, primary size, surface area, coating, agglomeration state, as well as by how dose is expressed (mass/surface area *vs.* particle number).[8] Additionally, complexity arises from time-dependent sequestration of particles in the lung and macrophage-overload effects, which alter inflammatory and downstream gene expression patterns over time.[9,10] Beyond biological variables, technical issues such as cross-platform batch effects complicate direct comparison of gene-level signals across doses, time points, and independent studies.[11,12]

[a]*University of Gdansk, Faculty of Chemistry, Laboratory of Environmental Chemoinformatics, Wita Stwosza 63, 80-308 Gdansk, Poland. E-mail: karolina.jagiello@ug.edu.pl, tomasz.puzyn@ug.edu.pl*
[b]*QSAR Lab Ltd, Trzy lipy 3, 80-172 Gdansk, Poland*

This journal is © The Royal Society of Chemistry 2026

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

These challenges underscore the need for advanced computational and integrative frameworks capable of harmonizing transcriptomic data across experimental systems. Emerging strategies include pathway-level and network-based analyses, which offer greater robustness than single-gene comparisons by contextualizing molecular changes within conserved biological processes.[13,14] Such approaches hold promise for improving reproducibility, enabling cross-study comparisons, and ultimately refining the application of transcriptomics for regulatory decision-making in nanotoxicology.

However, even with such integrative frameworks, the inherent statistical and computational complexity of transcriptomic data remains a major obstacle. One of the central challenges in applying transcriptomics to (nano)toxicology is the high dimensionality of the data. A single exposure experiment typically quantifies tens of thousands of genes, vastly exceeding the number of samples or materials tested. This imbalance, often described as the "curse of dimensionality", complicates analyses and increases the risk of overfitting.[15] Accordingly, there is a pressing need for scalable data modeling strategies that distill complex gene expression datasets into tractable, interpretable metrics.

Dimensional reduction techniques provide a foundation for such approaches. Unsupervised methods such as principal component analysis (PCA) are widely used to extract the dominant axes of variation in the transcriptome.[13,16] Alternative strategies – non-negative matrix factorization or independent component analysis for component discovery, and single-sample pathway scoring (*e.g.*, ssGSEA/GSVA/singscore) or network modules (*e.g.*, WGCNA) for biology-aligned summaries – offer complementary views but typically yield multi-dimensional outputs that require additional aggregation or interpretation.[17–21]

The Transcriptomic Response Index (TRI) was recently introduced as a novel approach in this field.[22] TRI is constructed as a weighted linear combination of selected principal components (PCs). Conceptually, the implementation of TRI involves using dimensionality-reduction techniques (such as PCA) to capture the major variance in the transcriptome, and then combining PCs to yield a single, continuous TRI score. The result is a quantitative index that integrates the contributions of many genes, rather than focusing on any single "biomarker" gene. Thus, TRI is the weighted linear combination of selected PCs. In practice, TRI can be paired with a supervised modeling, wherein exposure conditions (*e.g.*, material properties, dose, time) are used to estimate TRI values, thereby enabling quantitative exposure-response analysis.

A practical challenge, however, lies in defining how many PCs should be included in the TRI calculation. Prior work demonstrated that the predictive performance of the trained nano-Quantitative Structure–Activity Relationship (QSAR) models can vary with the PC subset used to construct TRI.[22] For instance, in a Multiwalled Carbon Nanotube (MWCNTs) inhalation dataset, Muratov *et al.* reported the best validation for a TRI variant using 19 PCs.[22] However, the most optimal combination of PCs may vary across nanomaterials, exposure context, and study designs, reflecting the diverse nature of ENM-induced perturbations.

Thus, although PCA-derived TRI offer a promising framework for summarizing transcriptomic perturbation, the stability and predictability of TRI for instilled TiO_2-NP, as a function of PC composition, have not been systematically evaluated within the previous study.[22] So far, there is no standard, validated procedure to choose the PC count for TRI calculated in response to TiO_2-NP instillation. In this work, we address this gap by defining and optimizing the PCs composition of TRI using robust model validation and providing standardized, direction-aware gene loadings to contextualize TRI. Through this evaluation, we assess TRI as a NAM-relevant, study-internal omics metric for early screening and prioritization. Importantly, in our recent TRI work on inhaled MWCNTs, we introduced TRI and showed that it works for carbon-based engineered nanomaterials – reducing whole transcriptomic expression data to one predictive variable.[22] In a different study, we trained machine learning (ML) models centred on TiO_2-NPs – those models linked particle attributes and exposure conditions to gene markers of lung damage.[8] The current work pushes both efforts forward. It sets out a general, validation-based way to pick the TRI principal component subset – we show that the component explaining the most variance is not always the one that predicts best. It shows that the optimized TRI idea can be rebuilt and predicted in a separate material class – TiO_2-NPs – as one continuous measure that suits non-animal prioritization and potency ranking.

Experimental section

Data preprocessing

In this study, we analyzed transcriptomic responses from mice exposed to TiO_2-NP, as reported in previously published studies. The underlying animal experiments[2,23–26] and associated global transcriptional changes in lung tissue[2] have been described in detail elsewhere. These animal experiments have resulted in a dataset comprised of 5 TiO_2-NPs in different concentrations (18, 54, and 162 μg per animal), charges, and post-exposure durations (1 and 28 days), which is presented in Table 1.

Gene expression profiles were derived from lung tissues of 5–7-week-old female C57BL/6 mice exposed by single intratracheal instillation to five distinct TiO_2-NPs, with sampling at 1 day and 28 days post-exposure, as described in prior studies.[2,8] Details of exposure conditions, tissue collection, and normalization procedures followed established protocols.[24,25,27] Intratracheal instillation was selected to standardize deposited lung doses across groups, thereby enabling direct comparability of TiO_2-NP-induced biological responses. Groups of 6–8 female mice were exposed to TiO_2-NP, alongside vehicle-treated negative controls and a benchmark group exposed to carbon black NPs (Printex90). The TiO_2-NP doses (18, 54, and 162 μg per mouse) were based on prior toxicological studies and extrapolated occupational exposure scenarios, corres-

 This journal is © The Royal Society of Chemistry 2026

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

Table 1 TiO_2-NP types, their physicochemical characteristics, and exposure conditions

TiO_2-NPs_IDs	NPs	Post-exposure duration [days]	Concentration [μg per animal]	Charge[a]	Size [nm]	BET [m^2 g^{-1}]	Deposited surface area[b] [cm^2]
NP_1	NRCWE-001	1	18	0	10	99	17.8
NP_2	NRCWE-001	1	54	0	10	99	53.5
NP_3	NRCWE-001	1	162	0	10	99	160.4
NP_4	NRCWE-002	1	18	1	10	84	15.1
NP_5	NRCWE-002	1	54	1	10	84	45.4
NP_6	NRCWE-002	1	162	1	10	84	136.1
NP_7	NRCWE-001	28	18	0	10	99	17.8
NP_8	NRCWE-001	28	54	0	10	99	53.5
NP_9	NRCWE-001	28	162	0	10	99	160.4
NP_10	NRCWE-002	28	18	1	10	84	15.1
NP_11	NRCWE-002	28	54	1	10	84	45.4
NP_12	NRCWE-002	28	162	1	10	84	136.1
NP_13	NRCWE-030	1	18	0	10.5	139.1	25.0
NP_14	NRCWE-025	1	18	0	38	28.2	5.1
NP_15	NRCWE-030	1	54	0	10.5	139.1	75.1
NP_16	NRCWE-025	1	54	0	38	28.2	15.2
NP_17	NRCWE-030	1	162	0	10.5	139.1	225.3
NP_18	NRCWE-025	1	162	0	38	28.2	45.7
NP_19	NRCWE-030	28	18	0	10.5	139.1	25.0
NP_20	NRCWE-025	28	18	0	38	28.2	5.1
NP_21	NRCWE-030	28	54	0	10.5	139.1	75.1
NP_22	NRCWE-025	28	54	0	38	28.2	15.2
NP_23	NRCWE-030	28	162	0	10.5	139.1	225.3
NP_24	NRCWE-025	28	162	0	38	28.2	45.7
NP_25	UV-Titan L181	1	18	0	20.6	107.7	19.4
NP_26	UV-Titan L181	1	54	0	20.6	107.7	58.2
NP_27	UV-Titan L181	1	162	0	20.6	107.7	174.5
NP_28	UV-Titan L181	28	18	0	20.6	107.7	19.4
NP_29	UV-Titan L181	28	54	0	20.6	107.7	58.2
NP_30	UV-Titan L181	28	162	0	20.6	107.7	174.5

[a] [0 – no charge; 1 – positive charge (contains positively charged amino groups on the surface[2])]. [b] Calculated based on specific surface area, size and concentration.

ponding to approximately 1.5, 5, and 15 eight-hour working days at the Danish occupational exposure limit for TiO_2-NP (6.0 mg Ti per m^3 or 10 mg TiO_2-NP per m^3).[2,24,25]

Differentially Expressed Genes (DEGs) were identified using a significance threshold of false discovery rate (FDR) $p \leq 0.05$ and an absolute fold change (FC) ≥ 1.5. This FC cut-off aligns strongly with the literature,[2,28–32] ensuring both methodological consistency and comparability with earlier studies. Although more stringent criteria (FDR $p \leq 0.05$ and FC ≥ 2) have also been recommended,[33,34] prior work with the TRI indicated reduced predictive performance at this threshold.[22] Thus, the genes qualifying the criteria of FDR p-value ≤ 0.05 and $|FC| \geq 1.5$ were classified as DEGs in this study and are presented in Table S1.

Thus, to ensure transparency and reproducibility, we took the transcriptomic profiles from publicly available mouse lung microarray data already published in other studies.[2,23–26] The design covers 30 exposure conditions – five TiO_2-NPs, each at three doses, and two post-exposure times. Each condition has at least five biological replicates plus time-matched vehicle controls in the original datasets. We used the already normalized expression values, compared each exposure group to its control at the same time point, but also called DEGs with the same FDR p-value and FC cut-offs used in the source study. The 30 conditions were combined to give the 621 genes listed in Table S1.

Unsupervised machine learning methods for TRI calculation

To reduce the dimensionality of the dataset comprising all DEGs, PCA was employed.[16,35,36] PCA transforms correlated gene expression variables into a set of orthogonal components, termed PCs, which are ordered according to the proportion of variance they explain. The first principal component (PC1) accounts for the largest share of variance, while successive PCs capture progressively smaller portions. This transformation facilitates the simplification and visualization of high-dimensional transcriptomic data.[37] To construct the TRI, each PC was weighted by its explained variance, and the weighted contributions were summed across all components, as described in formula (1), where n denotes the total number of included PCs.

This weighted integration incorporates both dominant and minor transcriptomic signatures, yielding a single composite metric that reflects the overall transcriptomic response. All options of TRI that are comprised of a different number of PCs (starting from PC1 + PC2 to the sum of all calculated PCs) are presented in Table S2.

The calculation of TRI using PCs:

$$\mathrm{TRI} = \mathrm{PC1} \times \mathrm{variance}_1 + \mathrm{PC2} \times \mathrm{variance}_2 + \ldots + \mathrm{PC}n \times \mathrm{variance}_n \quad (1)$$

where n is the total number of all PCs.

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

Supervised machine learning methods for linking TRI with TiO_2-NPs' structural properties

To ensure robust model development, the dataset of 30 observations (Table 1) was partitioned into training and external hold-out test subsets. Initially, all the 30 data points were ranked in ascending order according to their TRI values. To construct the validation set, every fifth condition starting from the third in the ordered sequence was excluded from the training set. This procedure ensured that the validation points were systematically and evenly distributed across the TRI range of the training nanomaterials.

Model training was performed by ChemBioML platform (https://chembioml.com) using the calculation of polynomial features based on the initial set of experimental conditions from Table 1, and ridge regression combined with a genetic algorithm (GA)-based feature selection procedure. Ridge regression, a linear modeling approach with L2 regularization, enables effective handling of cross-correlated predictors[38] and facilitates modeling in high-dimensional predictor spaces.[39] The GA was employed to optimize the predictor selection, ensuring that only the most relevant features were retained for the ML model construction.[40]

Overall, to ensure the best reproducibility without the ChemBioML platform, the supervised workflow always runs the same steps in the same order. The data are split once – the external validation subset is locked before the model is chosen and never influences the selection of the TRI constitution, the predictors, or the hyperparameters. Every predictor is transformed and then standardised with a *z*-score – the mean and standard deviation used for the standardisation are taken only from the training set. New predictors are created as second-degree polynomials from the remaining predictors. The ridge regression model is trained – the strength of the regularisation is chosen by cross-validation inside the training set from a declared grid of values. A GA selects the most optimal predictors.

To mitigate overfitting risks inherent to small datasets, a comprehensive validation strategy was implemented. Model fit was assessed using the coefficient of determination (R^2) and the root mean square error of calibration ($RMSE_C$). Robustness was evaluated *via* leave-one-out cross-validation (CV), yielding the coefficient of determination for CV (Q^2_{CV}) and the root mean square error of CV ($RMSE_{CV}$). Predictive performance was further examined with external validation metrics, including the external validation coefficient (Q^2) and the root mean square error of external validation ($RMSE_{ext}$).

The applicability domain (AD) of the models, defined by the range of response values and structural predictor space, was assessed using the leverage method and visualized through a Williams plot, a well-established approach in ML modeling.[41–43]

Results and discussion

The primary objective of this study was to optimize the application of TRI and supervised ML methods to efficiently compress and predict transcriptomic alterations across hundreds of DEGs within a single ML modeling framework. To achieve this, we utilized a transcriptomic dataset derived from a previously published TiO_2-NP exposure study. Specifically, gene expression profiles were obtained from lung tissues of 5–7-week-old female C57BL/6 mice exposed to TiO_2-NPs *via* intratracheal instillation. To ensure statistical rigor and biological relevance, only genes meeting the dual thresholds of a false discovery rate (FDR)-adjusted $p \leq 0.05$ and an absolute fold change (FC) ≥ 1.5 in at least one exposure condition were retained.[2,28–32] This filtering yielded a set of 621 DEGs significantly affected by TiO_2-NPs exposure in the lungs (Table S1).

By focusing on this curated DEG subset, we aimed to capture the core transcriptional perturbations associated with TiO_2-NP instillation and use them as a foundation for downstream dimensionality reduction, TRI optimization, and predictive ML modeling. This strategy not only enables efficient representation of high-dimensional transcriptomic data but also facilitates interpretable associations between nanoparticle exposure characteristics and biological response magnitude.

Compression of DEGs into a single TRI

To address the challenge of modeling high-dimensional transcriptomic data comprising 621 DEGs in this study (Table S1), we used the TRI and linked this composite metric to TiO_2-NPs properties using a single ML model. Conventional computational approaches typically train independent models for each gene, thereby treating every DEG as a separate endpoint.[8] While practical for small gene sets, such strategies become computationally prohibitive for datasets involving thousands of DEGs. Moreover, measuring or predicting changes in a single gene is insufficient, as such isolated alterations rarely translate into functional changes at the tissue level. In contrast, the TRI provides a unified endpoint that integrates all extractable information from the contributing genes.

Reducing 621 gene-specific endpoints into a single TRI variable markedly streamlines the modeling process. Instead of constructing hundreds of individual ML models, the TRI enables the development of one integrated model, thereby enhancing both computational efficiency and interpretability. Summarizing the global transcriptomic response in this manner not only supports predictive modeling but also facilitates integration with other omics data or phenotypic outcomes. This systems-level representation is particularly advantageous in contexts where coordinated gene activity conveys greater biological insight than isolated gene-level effects.

To calculate the TRI, PCA was applied to the 621 DEGs from Table S1. Combinations of the first 29 PCs (ranging from PC1–PC2 through PC1–PC29) were used to compute TRI metrics of increasing dimensionality (TRI_2 through TRI_{29}), thereby allowing assessment of the optimal balance for predicting global transcriptomic patterns following TiO_2-NP exposure in the lungs. This unsupervised approach captured ~44% of the variance with TRI_2 and nearly 99.9% with TRI_{29}. Importantly, all TRI options are present in Table S2, ensuring the reproducibility of the approach. In this study, we call a PC a "key PC" if,

 This journal is © The Royal Society of Chemistry 2026

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

after supervised validation that checks predictivity and robustness, the component is kept in the final TRI model – we do not keep it only because it adds to the cumulative explained variance. It has a biologically-relevant reason – components that raise cross-validated predictivity might carry a reproducible exposure-linked signal, while extra components often mirror leftover, study-specific noise that raises model complexity without assisting generalization. In order to identify biological information the kept PCs capture, we took the gene loadings for each PC (Table S3), ranked genes by the absolute value of their loadings, and used the sign of each loading to indicate whether expression correlates positively or negatively with the PC score. Shifting along a PC axis indicates the coordinated change in the transcripts that most strongly drive that axis – the optimized TRI can be read as a picture of the main biological programs that are represented by PCs.

In addition, PC1 and PC2 were employed to construct a two-dimensional "transcriptomic space" for visualization of similarities among different TiO_2-NPs based on DEG fold-change profiles (Fig. 1). This low-dimensional representation provides an intuitive overview of transcriptomic relationships, while the full set of 29 PCs ensures high-fidelity reconstruction of the data for TRI development.

Projection of the 30 TiO_2-NP conditions onto the first two PCs (PC1 = 29.62%; PC2 = 14.31%) reveals a largely continuous transcriptomic space rather than discrete clusters (Fig. 1). A weak left-to-right color gradient is visible along PC1, suggesting a partial alignment between deposited dose and this axis of variation; however, several high-dose conditions occupy the left/negative PC1 space (*e.g.*, NP_23, NP_30), indicating that dose alone does not determine the primary transcriptomic response. This pattern implies that other experimental attributes (*e.g.*, TiO_2-NPs charge, size, post-exposure periods) can modulate the global response sufficiently to override simple mass-dose effects.

Several TiO_2-NPs instillation omics biomarkers[8] contributing to the PC1 through loadings (Table S3), including *Saa3*, *Saa1*, *Serpina3n*, are directly associated with acute phase response and inflammation.[28,44,45] The response mediated by these genes is a well-established systemic reaction that not only reflects stress but is also closely linked to the onset and progression of cardiovascular diseases.[46,47] At the protein level, correlation between *Saa3* plasma levels and neutrophil levels induced by metal oxides inhalation was also reported,[48] consistent with the role of *Ccl2* protein expression in bronchoalveolar lavage fluid, which correlates with neutrophil influx induced by inhalation of diesel exhaust.[49] In contrast, PC2 captures a distinct set of biological processes represented by key biomarkers, such as *Ccl3* (associated with lung fibrosis[50]), *Hspa1a* (related to cardiovascular diseases[51]) and other genes.

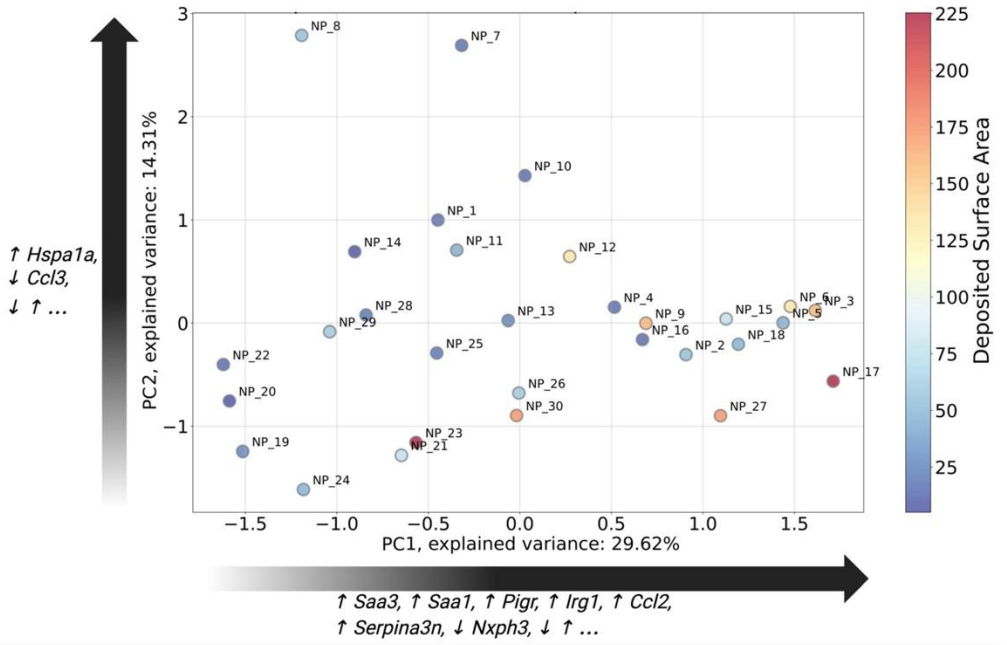


Fig. 1 PC1–PC2 scores plot showing the two-dimensional transcriptomic space underlying TRI_2. Each point represents one TiO_2-NP exposure condition, and distances between points reflect similarity of DEG FC profiles. Dose is expressed as deposited surface area and is encoded for visualization by color, illustrating how increasing deposited surface area shifts samples along the dominant transcriptomic axes.

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

Overall, along PC2, a small but distinct grouping emerges – NP_7, NP_8, and NP_10 – corresponding to NRCWE-001 (NP_7/NP_8) and NRCWE-002 measured 28 days post-exposure at 18 μg per animal. The positive PC2 displacement of these data points, contrasted with the broader scatter of the remaining materials, is consistent with a transcriptomic response that is specific to these legacy particles and/or to persistent responses at the late time point.

Together, these observations indicate a quasi-gradient with dose along PC1 tempered by material-specific biology, and a secondary axis (PC2) that separates a subset of two NRCWE materials at day 28, consistent with a distinct, longer-lasting response phenotype. Also, these observations provide important and meaningful information on TiO_2-NPs induced toxicity MoA through the transcriptomic biomarkers[8] loadings contributing to the PC1 and PC2 (Table S3). Importantly, the obtained results not only indicate the important genes but also quantitatively measure their dependence on the PCs scores (for example, *Saa3*, *Saa1*, *Pigr*, *Irg1*, *Ccl2*, and *Serpina3n* expression increases with the larger values of PC1, and *Nxph3* decreases, while PC2 is positively impacted by *Hspa1a* and negatively by *Ccl3*, Fig. 1).

Applying supervised ML algorithms to optimize TRI constitution and link the index to TiO_2-NPs instillation conditions

Training of the supervised ML models predicting differentially constituted TRI (from TRI_2 to TRI_{29}) and selection of the best one is the most important step of this work. As the primary aim of this work is the optimization of TRI based on supervised metrics, Q^2_{CV} is the main metric used for model optimization and selection. A set of ML models predicting different TRIs was trained. With this approach, every TRI, constituted from combinations of PC1–PC2 to PC1–PC29 (Table S2), was used as a separate endpoint for training an ML model. Such a comparison with the previous study[22] would show whether TRI needs any constitution optimization during the study, or the maximum number of PCs can be included, as in the prior study.

For training and validating the models, a set of 2nd degree polynomial predictors was generated based on physicochemical properties and exposure conditions from Table 1, and the ChemBioML platform was utilized (https://chembioml.com). Such a combined approach helped us to catch the complex interactions between the instilled TiO_2-NPs (as polynomial features are known to cover more information than original ones[52,53]) and to prepare accurate and regulatory acceptable ML models.[22] The predictor dataset was first preprocessed by removing predictors with constant values and standardizing them. Traditional ML approaches, such as multiple linear regression (MLR), often require the exclusion of predictors when cross-correlation exceeds a given threshold (*e.g.*, $R > 0.8$), which may lead to the loss of potentially relevant predictive information. To circumvent this limitation, we applied ridge regression, which incorporates L2 regularization to manage multicollinearity.[38] Unlike MLR, ridge regression penalizes large coefficients rather than discarding correlated variables outright, thereby retaining a broader predictor space and enhancing both model robustness and predictive accuracy. Consequently, predictors with high intercorrelation were retained in our modeling.

To rigorously evaluate predictive performance, six evenly distributed data points were excluded from training and reserved as an external hold-out test set. These points, representative of the full dataset, were not used in model calibration or optimization. The remaining 24 data points were employed for model training and internal validation.

Model performance for the training set was assessed using the R^2 and $RMSE_C$, which indicated a strong goodness-of-fit. Robustness was further confirmed through CV, where the corresponding Q^2_{CV} and $RMSE_{CV}$ values were consistent with reliable model generalization. Finally, predictive ability was quantified using the Q^2 and the $RMSE_{ext}$. Fig. 2 was used to identify the best balance between the model's Q^2_{CV} and the cumulative explained variance by TRI.

We built a set of TRI scores, labelled TRI_2 through TRI_{29}. Each score is the simple sum of the first n PCs – TRI_2 equals PC1 plus PC2, TRI_3 equals PC1 plus PC2 plus PC3, and so on until TRI_{29}, which adds all components from PC1 to PC29. Every score served once as the response variable in its own ridge regression run. The same pool of physicochemical and exposure variables was offered to a genetic algorithm selector each time. For every run, we saved R^2 and $RMSE_{CV}$ to evaluate the fit, Q^2_{CV}, and $RMSE_{CV}$ to assess internal stability. We inspected those numbers, along with the share of transcriptomic variance already accounted for by the components that entered the score. TRI_2 gave the best balance – its Q^2_{CV} reached the highest value while the two components already explained roughly 44% of the variance. When we kept adding components, the captured variance rose – yet Q^2_{CV} fell step by step.

Thus, as it is visible in Fig. 2, the model's Q^2_{CV} and the explained variance by TRI have the closest match at TRI_2, which

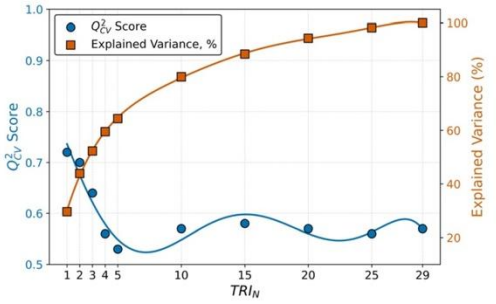


Fig. 2 Plot illustrating how TRI performance and information content change with the number of included principal components (PCs). The curve for cross-validated predictivity (Q^2_{CV}; higher is better) is shown alongside the cumulative explained variance (higher is better) captured by TRI as PCs are added sequentially. This visualization supports selecting an optimal TRI constitution by balancing robust predictive performance against retaining sufficient transcriptomic variance.

This journal is © The Royal Society of Chemistry 2026

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

represents a combination of PC1 and PC2. After that, the model's quality significantly decreases (Q^2_{CV} for TRI_3 is closer to 0.6, which is not the best choice in case of the availability of a more robust model for TRI_2), and the growth of TRI explainability slows down. This observation showcases that TRI_2 is the best option for predicting the global transcriptomic response to the TiO_2-NPs instillation. In contrast, the prior study introducing TRI showcases TRI_{19} (a combination of PC1–PC19) as the most predictable index.[22] These findings suggest that the different nature of transcriptomic data affects the optimal constitution of TRI based on its predictability. Comparing this and the prior study, where a set of MWCNTs was studied, we hypothesize that the difference in the MoA between responses to MWCNTs and TiO_2-NPs instillation may affect the optimal constitution of TRI. Transcriptomic MoA of TiO_2-NPs exposure is characterized by immune activation (cytokine, complement, acute-phase genes) and antioxidant responses, with comparatively modest fibrotic or proliferative gene signatures.[54,55] In contrast, MWCNTs, with their fiber-like shape, act as a more aggressive pulmonary hazard at lower doses: they incite a robust and persistent inflammatory transcriptomic program, marked by both Th1 and Th2 cytokines, and strongly tip the balance toward fibrosis-related and cell-cycle pathways.[56] The transcriptomic hallmarks of MWCNT exposure include surges in osteopontin and IL-13-linked genes (signifying pro-fibrotic Th2 activity), alongside upregulation of cell division and survival genes that may underpin fibrosis and carcinogenesis.[56,57] Meanwhile, TiO_2-NP's transcriptomic hallmarks are more aligned with a "particle overload" inflammatory response (neutrophil chemotaxis, acute phase proteins, *etc.*) and oxidative stress mitigation.[55,58] These differences underscore that MWCNTs, by virtue of their shape and biopersistence, can invoke qualitatively similar pathways as TiO_2-NPs (inflammation, ROS) but to a greater degree and with additional fiber-specific pathways (*e.g.*, frustrated phagocytosis leading to *NLRP3* inflammasome activation and fibrogenic signaling). These findings align with the experimental data we have worked with – TiO_2-NPs provoked a significant response of 621 DEGs, while the prior study introducing TRI analyzed a dataset with over 5000 DEGs significantly affected by MWCNTs inhalation.[22] Thus, our study underscores that the constitution of TRI must not always contain all calculated PCs, but it has to be optimized every time for a new study, balancing between high predictability and the explained variance level of the index. Selecting TRI_2 balances fidelity and generalization: although TRI_{29} reconstructs essentially the entire expression space (≈99.9%), TRI_2 (≈44%) captures the principal axes that drive predictability while reducing noise and overfitting risk. This framing treats TRI_2 as a compact, predictive proxy of the global response rather than a complete reconstruction. Because TRI is composed of PC loadings (Table S3), it remains reversible – gene-level attributions can be recovered from TRI_2, preserving mechanistic transparency.

The statistics of the selected ML model predicting TRI_2 are presented in Table 2, where the determination coefficients and RMSE for training, CV, and external validation is available. The overall evaluation of this model demonstrated strong goodness-of-fit, robustness, and external predictability. Importantly, model quality is determined not only by the absolute values of performance metrics but also by the difference between R^2 and Q^2, which should not exceed 0.3 to avoid indications of overfitting.[59,60] For our ML model, this difference is 0, confirming that the model is not overfitted. From a nanosafety modeling perspective, the predictivity we reached is competitive with published studies that link nanomaterial properties to biological effects.[42,59–61] For example, Puzyn *et al.* demonstrated that descriptors based on physical and chemical mechanisms can give reliable predictions of metal oxide nanoparticle toxicity. This shows that structure–activity modeling is feasible in nanotoxicology. Our endpoint is not a single assay readout but a single transcriptomic index that integrates *in vivo* lung responses. This setting contains more biological complexity than most *in vitro* endpoints. The external predictivity we reached shows that the transcriptomic response index works as a quantitative endpoint that NAMs can use. This view agrees with nanosafety modeling guidance, which states that regulatory relevant ML models require strict validation, a clear AD, and predictors that have a mechanistic meaning, as set out in the NanoSafety Modelling Cluster perspective.[61]

Table 2 Statistics of the ML model predicting TRI_2

Statistic	Value
R^2	0.79
$RMSE_C$	0.15
Q^2_{CV}	0.70
$RMSE_{CV}$	0.18
Q^2	0.79
$RMSE_{ext}$	0.15

The correspondence between predicted and observed values is illustrated in Fig. 3a. While both external validation and CV were employed, it should be noted that the relatively small sample size (24 training data points) may constrain the generalizability of the model. To address this limitation and minimize the probability of inaccurate predictions, an AD was estimated for the trained ML model. It is an essential component of ML model evaluation.[62]

The AD provides a framework for determining how structurally similar validation or prediction compounds are to those included in the training set, while also identifying potential outliers – cases where structurally comparable compounds are nevertheless predicted inaccurately. To delineate the AD of our ML model, we employed a Williams plot, which simultaneously visualizes structural diversity (through leverage values) and predictive accuracy (*via* standardized residuals) (Fig. 3b).

As shown in Fig. 3b, all compounds from both the training and validation sets exhibited absolute standardized residuals within ±3. There are also no data points with out-of-domain leverages, showcasing no structural outliers (where the TiO_2-NPs physicochemical properties or experimental conditions are significantly different from the rest of the dataset). This outcome indicates the absence of outliers, suggesting that the model's predictions are consistent across structurally diverse

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

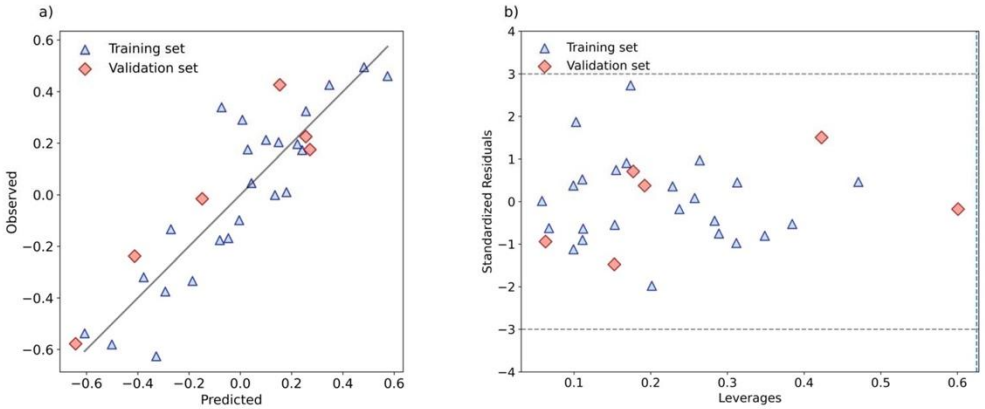


Fig. 3 (a) Observed–predicted scatter plot demonstrating agreement between experimental and model-predicted TRI values, illustrating predictive accuracy across the modeled range; (b) Williams plot (standardized residuals *vs.* leverage) used to assess the model's applicability domain (AD), identify potential outliers and influential observations, and verify that predictions are generated within the predictor space supported by the training data.

TiO_2-NPs and that no individual materials with high similarity were poorly predicted. However, additional testing on larger, independent datasets will therefore be necessary to fully substantiate predictive performance and further minimize potential overfitting risks.

The modeling software ChemBioML platform (https://chembioml.com) has also provided a linear regression equation for the obtained ML model (formula (2)) and the model's hyperparameters: 'alpha': 0.1, 'fit_intercept': false, 'solver': 'sparse_cg', 'tol': 0.1, where alpha is the value controlling regularisation, "sparse_cg" is the conjugate gradient solver, and "tol" is the precision of the solution, specifying norms of residuals smaller then "tol". Analyzing the coefficients of the regression equations is a very common method for performing a mechanistic interpretation of linear models like ours.[13,22,41,59] They are interpreted as the expected change in an outcome for a one-unit change in a predictor, holding other predictors constant, which allows scientists to turn statistical models into quantitative statements in the units of their fields. These coefficients are used to quantify effect sizes, compare relative importance, test hypotheses with uncertainty, and translate models into domain meaning with units, transformations, and interactions.

Final ridge-regression equation used to predict TRI_2 from the selected TiO_2-NP physicochemical descriptors and exposure-condition variables, reporting the fitted coefficients for the optimized model.

$$\begin{aligned} TRI_2 = {} & 0.13 \times \text{concentration} - 0.195 \times \text{BET}^2 \\ & - 0.071 \times \text{time} \times \text{concentration} - 0.167 \times \text{time} \\ & \times \text{size} - 0.201 \times \text{size} \end{aligned} \quad (2)$$

Formula (2) is showcasing that TRI_2 depends the most on the instilled TiO_2-NPs BET, size, exposure concentration, and the post-exposure time. Previous computational studies and empirical toxicogenomics align on this point, showing that specific surface area scales with cellular association/uptake and with activation of oxidative-stress and inflammatory cascades – canonical NRF2, MAPK, and NF-κB responses that propagate to downstream cell-cycle, DNA-damage, and metabolic pathways.[63,64] Primary size co-varies with BET and therefore with surface-mediated effects. Smaller TiO_2-NPs generally produce more extensive DEG sets and stronger pathway enrichment, which is mechanistically plausible given higher curvature, more defect sites, and facilitated intracellular access of smaller primaries and compact aggregates.[65–68] Size also modulates the kinetics and ceiling of uptake: smaller particles reach higher intracellular numbers at equal mass, thereby amplifying oxidative and inflammatory programs and shifting dose–response curves leftward.[67,69] At the same time, cells and tissues do not encounter primary particles in isolation. Aggregation produces transport-relevant hydrodynamic sizes that determine delivered dose and, for inhalation, regional lung deposition. Studies comparing ~166 nm agglomerates to larger clusters demonstrate distinct mRNA signatures, underscoring that aggregate size can be an important driver of transcriptomic outcome independent of primary size.[66] These observations argue for explicit reporting and control of agglomeration and corona formation in transcriptomic studies, and for interpreting "size effects" through the lens of delivered reactive surface. Meantime, dose remains the most robust predictor of transcriptomic amplitude across systems. Multiple datasets show clear concentration-response behavior, with low microgram-per-milliliter ranges already engaging adaptive stress programs in sensitive cell types and conditions.[70–72] As concentration increases, signatures progress from antioxidant and proteostatic defenses toward apop-

 This journal is © The Royal Society of Chemistry 2026

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

tosis, necroinflammation, and cell-cycle arrest, consistent with a transition from compensatory to injurious states.[70–73] Threshold-like behavior is common: above ~10 μg mL^{-1} in submerged exposures, cytotoxicity accelerates, reflecting saturation of defenses and secondary inflammatory signaling.[67] Time after exposure introduces an additional effect to the response. Acute sampling (≈24 h) frequently yields thousands of DEGs spanning oxidative stress, cytokine signaling, and metabolic rewiring; these signatures often contract substantially with time as homeostatic mechanisms restore baseline or as exposed cells are replaced, with only dozens of DEGs persisting in progeny populations.[58,72,74] Across days to weeks, pathway emphasis shifts: early time points (1–3 d) are dominated by innate inflammation and redox responses, whereas later windows (≈28 d) increasingly reflect tissue remodeling, extracellular matrix organization, and immune resolution programs.[23,58] Thus, the selected predictors align strongly with the existing knowledge of MoA induced by the TiO_2-NPs instillation.

Together, our results demonstrate that TRI is portable beyond MWCNTs and must be re-optimized for each dataset to maximize predictability, rather than being assumed to require a fixed number of PCs. In the TiO_2-NP dataset, the balance between model robustness (high Q^2_{CV}) and captured variance is achieved at TRI_2 (PC1 + PC2), not at the higher-dimensional solution reported for MWCNTs (TRI_{19}) in prior work, underscoring study-specific MoA as a determinant of the optimal TRI composition. Consistent with this, PC1 (29.62%) and PC2 (14.31%) capture a meaningful two-axis "transcriptomic space" in which dose aligns only partially with PC1 and is modulated by particle attributes and post-exposure time – supporting the need for a compact, responsive TRI for TiO_2-NPs. Coupling the optimized TRI_2 to a regularized, feature-selected ML model built on particle and exposure predictors gives consistent performance on both internal and external tests, indicating a stable signal rather than overfitting. AD checks support reliable use within the studied predictor space, with no influential outliers. Mechanistically, the model is driven by BET, primary size, exposure concentration, and post-exposure time, in line with established TiO_2-NPs biology linking reactive surface-driven oxidative/inflammatory programs and their resolution over time. Together, these results deliver the main aim: TRI can be transferred beyond MWCNTs and, when re-optimized to the dataset, serves as a compact, interpretable, and predictive bridge from material attributes and conditions to system-level transcriptomic response. While this work continues our earlier TRI and TiO_2-NPs biomarker identification studies,[8,22] this new study supplies a step that regulators need before they adopt TRI as an omics endpoint. Regulatory acceptance requires more than proof-of-concept for one material alone – it also requires that the endpoint definition and the model-building choices be chosen through transparent validation rules and that the same rules deliver the same endpoint when they are applied to different ENM classes. The methodological advance is therefore a validated selection rule and a standardized workflow. Those two factors let TRI serve as a single scalable endpoint so that users no longer have to keep a separate model for every gene when they screen, rank, or extrapolate within a study.

In practical terms, this study extends the TRI framework from MWCNTs to TiO_2-NPs and formalizes a data-driven rule for TRI selection: selection of the PC subset that jointly improves Q^2_{CV} and preserves sufficient explained variance, then validation with external hold-out validation and AD checks. Summarizing all the validation steps, our ML model aligns with the criteria defined by the Organization for Economic Cooperation and Development (OECD) and is supportive of NAM workflows.[62]

To show how regulators can use it, TRI can be inserted into an NAM pipeline as a clear-cut number that measures bioactivity at the omics level inside a step-by-step integrated approaches to testing and assessment (IATA).[75] Every transcriptomic study that builds or refreshes the TRI must be reported with full metadata and quality details (study design, controls, normalization, definition of DEGs) so that any reviewer or regulator can repeat the work. The TRI is defined by the chosen gene set, centering/scaling, and principal-component loadings – if this definition is supplied, any lab can recalculate the TRI under matching conditions. The externally validated ML model (list of predictors, coefficients, and AD terms) predicts the TRI from material attributes and exposure settings, which supports early screening and potency ranking and supplies a numerical basis for grouping/read-across hypotheses within a stated domain. When a quantitative decision is needed instead of mere prioritization, the TRI can serve as a continuous dose–response endpoint to find transcriptomic points of departure – those points can be combined with exposure data and other NAM evidence inside an IATA. Broader regulatory use will require further validation across studies and laboratories, and decision rules that fit each context (for example, the exact threshold that separates a "low" from a "high" TRI score).

Thus, by compressing 621 DEGs into a single, validated endpoint and predicting it accurately from exposure conditions, the workflow reduces omics complexity to an NAM-ready metric suitable for screening, potency ranking, and study-internal extrapolation. However, the regulatory interpretation of TRI should be considered supportive rather than definitive at the current stage. In this work, TRI is derived from lung transcriptomics and evaluated as an internally consistent, predictive omics endpoint; we do not directly quantify its relationship to concurrent apical pathological outcomes (*e.g.*, histopathology, inflammatory cell influx, fibrosis/collagen endpoints) or to formally measured Adverse Outcome Pathway (AOP) key-event metrics within the same experimental set.[76] Therefore, TRI should be viewed as a compact measure of molecular response intensity and direction that is mechanistically interpretable through gene loadings and consistent with established TiO_2-NPs pulmonary exposure biology, but not yet validated for adversity. Strengthening this linkage for broader regulatory use will require independent datasets that include

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

matched apical outcomes, enabling explicit TRI-outcome associations, comparison of transcriptomic and apical points of departure, and development of context-specific decision thresholds for IATA/NAM applications. While the present sample size cautions that broader validation will further strengthen generalizability, the convergence of predictive performance, biological plausibility, and AD coverage indicates that TRI can be systematically optimized and transferred to new ENMs, enabling transcriptomics to function as a scalable, quantitative bridge between material attributes and integrated biological response. PCA turns the activity of hundreds of different genes into a small set of new directions called PCs. Every component stands for a distinct, major way in which the genes move together across the exposures – each exposure gets a value on every component that shows how much of that pattern it displays. TRI next compresses those multiple values into one value, adding the first n components, giving greater weight to those that account for a larger explanation of the total gene activity variation. The result is a concise score that keeps the main shape of the transcriptomic response and leaves out the minor fluctuation carried by other components.

It is important to distinguish the external hold-out validation reported here from fully independent external omics validation. Every step – definition of DEGs, derivation of TRI, and supervised modelling – relied on one TiO_2-NPs pulmonary exposure transcriptomic study that was produced under a single experimental and analytical framework. Because of that, the Q_2 values obtained here cannot be assumed to apply to other studies. Whether the index transfers to independent TiO_2 pulmonary transcriptomic datasets (*e.g.*, datasets from different laboratories, batches, platforms like RNA-seq, or studies with alternative exposure designs) still needs to be indicated. Future work should re-derive TRI and test model performance through inter-study validation, measure how sensitive results are to preprocessing choices, and specify the context-of-use boundaries (material class, exposure route, and time points and platform) within which TRI can function as a regulatory-grade NAM endpoint.

This transferability means only that the workflow itself can be applied to another project, but it does not mean that one fixed TRI definition will suit every kind of nanomaterial. As soon as TRI is derived from PCA, its numeric scale and meaning depend on the transcriptomic matrix that produced it – the chosen genes, exposure design, time points, and the preprocessing steps used by the platform. TRI values remain comparable only inside the single harmonized dataset from which they were produced – they are not directly comparable across fundamentally different ENM classes or study designs unless the entire procedure is repeated. When a new nanomaterial type is studied (for example, another metal oxide, a carbon-based ENM, or quantum dots), the investigator must recalculate TRI from the corresponding DEG space and repeat the full optimization pipeline (selection of principal component subset, external validation, and AD assessment). This confirms that the retained components capture reproducible, material-specific biological signal instead of variance that belongs only to the dataset. Within those limits, TRI is useful in three practical ways: it supplies a quantitative endpoint that is ready for non animal methods and that can be used to build models linking material properties to biological response inside a single material class; it supplies a mechanistically anchored omics summary that supports potency ranking but also grouping or read across inside the same class; and it connects high dimensional transcriptomics to regulatory workflows – enabling exposure response modeling and prioritization without the need to examine each gene separately.

Conclusions

We re-optimized the TRI for TiO_2-NP, compressing lung transcriptomics into a single NAM-ready endpoint and identifying TRI_2 as the best balance of predictability and generalization. Coupled with ridge regression and GA-based feature selection, the model trained by the ChemBioML platform (https://chembioml.com) is compact, interpretable, and externally validated ($R^2 = 0.79$; $Q^2_{CV} = 0.70$; $Q^2 = 0.79$), with a clean AD. Coefficients highlight BET surface area, primary size, concentration, and post-exposure time as principal drivers of the transcriptomic response to the TiO_2-NP exposure in lungs, aligning with known MoA. Practically, this replaces hundreds of gene-level models with a single reversible index, streamlining screening and potency ranking in regulatory workflows.

Author contributions

V. M. and K. J. conceptualization. V. M. performed formal analysis, including data curation, the development of the model, and validation. V. M. designed and wrote the manuscript with the support of K. J. T. P. review and editing. T. P. and K. J. funding acquisition and project administration. All authors discussed the results and commented on the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information: Table S1: fold change values of 621 DEGs in response to 5 TiO_2-NPs in different combinations; Table S2: values of TRI1–TRI29 calculated based on PCs; Table S3: loading values of 621 DEGs in PC1–29. See DOI: https://doi.org/10.1039/d5nr04657b.

This journal is © The Royal Society of Chemistry 2026

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

Acknowledgements

This work was funded *via* the National Science Centre in the frame of the TransNANO project (UMO-2020/37/B/ST5/01894). Figures for this work were prepared using the software BioRender.com. ML models were trained using the software ChemBioML platform (https://chembioml.com).

References

1 E. M. Rossi, L. Pylkkänen, A. J. Koivisto, M. Vippola, K. A. Jensen, M. Miettinen, K. Sirola, H. Nykäsenoja, P. Karisola, T. Stjernvall, E. Vanhala, M. Kiilunen, P. Pasanen, M. Mäkinen, K. Hämeri, J. Joutsensaari, T. Tuomi, J. Jokiniemi, H. Wolff, K. Savolainen, S. Matikainen and H. Alenius, *Toxicol. Sci.*, 2010, **113**, 422–433.

2 S. Halappanavar, A. T. Saber, N. Decan, K. A. Jensen, D. Wu, N. R. Jacobsen, C. Guo, J. Rogowski, I. K. Koponen, M. Levin, A. M. Madsen, R. Atluri, V. Snitka, R. K. Birkedal, D. Rickerby, A. Williams, H. Wallin, C. L. Yauk and U. Vogel, *Environ. Mol. Mutagen.*, 2015, **56**, 245–264.

3 L. Rahman, D. Wu, M. Johnston, A. William and S. Halappanavar, *Mutagenesis*, 2017, **32**, 59–76.

4 S. Lee, T. Weiss, M. Bühler, J. Mena, Z. Lottenbach, R. Wegmann, M. Sun, M. Bihl, B. Augustynek, S. P. Baumann, S. Goetze, A. van Drogen, P. G. A. Pedrioli, D. Penton, Y. Festl, A. Buck, D. Kirschenbaum, A. M. Zeitlberger, M. C. Neidert, F. Vasella, E. J. Rushing, B. Wollscheid, M. A. Hediger, M. Weller and B. Snijder, *Nat. Med.*, 2024, **30**, 3196–3208.

5 K. Paul Friedman, R. S. Thomas, J. F. Wambaugh, J. A. Harrill, R. S. Judson, T. J. Shafer, A. J. Williams, J.-Y. J. Lee, L.-H. Loo, M. Gagné, A. S. Long, T. S. Barton-Maclaren, M. Whelan, M. Bouhifd, M. Rasenberg, U. Simanainen and T. Sobanski, *Toxicol. Sci.*, 2025, **205**, 74–105.

6 European Chemicals Agency, *New approach methodologies in regulatory science – Proceedings of a scientific workshop – Helsinki, 19–20 April* 2016, ECHA, 2016.

7 *OECD Omics Reporting Framework (OORF): Guidance on reporting elements for the regulatory use of omics data from laboratory-based toxicology studies*, OECD, 2023.

8 V. Muratov, K. Jagiello, A. Mikolajczyk, P. H. Danielsen, S. Halappanavar, U. Vogel and T. Puzyn, *J. Hazard. Mater.*, 2025, **493**, 138240.

9 S. Kumar, C. Mickael, R. Kumar, R. R. Prasad, N. V. Campbell, H. Zhang, M. Li, B. A. McKeon, T. E. Allen, B. B. Graham, Y.-R. A. Yu and K. R. Stenmark, *Front. Immunol.*, 2024, **15**, 1372959.

10 L. Perez, J. Ambroise, B. Bearzatto, Y. Yakoub, M. Palmai-Pallag, C. Bouzin, L. Ryelandt, C. Pavan, F. Huaux and D. Lison, *Arch. Toxicol.*, 2025, **99**, 4941–4953.

11 J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly and R. A. Irizarry, *Nat. Rev. Genet.*, 2010, **11**, 733–739.

12 Y. Yu, Y. Mai, Y. Zheng and L. Shi, *Genome Biol.*, 2024, **25**, 254.

13 K. Jagiello, S. Halappanavar, A. Rybińska-Fryca, A. Willliams, U. Vogel and T. Puzyn, *Small*, 2021, **17**, 2003465.

14 S. Merugu, K. Jagiello, A. Gajewicz-Skretna, S. Halappanavar, A. Willliams, U. Vogel and T. Puzyn, *Small*, 2025, **21**, 2501185.

15 A. Serra, M. Fratello, L. Cattelani, I. Liampa, G. Melagraki, P. Kohonen, P. Nymark, A. Federico, P. A. S. Kinaret, K. Jagiello, M. K. Ha, J.-S. Choi, N. Sanabria, M. Gulumian, T. Puzyn, T.-H. Yoon, H. Sarimveis, R. Grafström, A. Afantitis and D. Greco, *Nanomaterials*, 2020, **10**, 708.

16 I. T. Jolliffe and J. Cadima, *Philos. Trans. R. Soc., A*, 2016, **374**, 20150202.

17 P. Langfelder and S. Horvath, *BMC Bioinf.*, 2008, **9**, 559.

18 D. D. Bhuva, J. Cursons and M. J. Davis, *Nucleic Acids Res.*, 2020, **48**, e113.

19 M. Foroutan, D. D. Bhuva, R. Lyu, K. Horan, J. Cursons and M. J. Davis, *BMC Bioinf.*, 2018, **19**, 404.

20 S. Hänzelmann, R. Castelo and J. Guinney, *BMC Bioinf.*, 2013, **14**, 7.

21 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.

22 V. Muratov, K. Jagiełło and T. Puzyn, *Nanoscale Horiz.*, 2025, **10**, 3116–3126.

23 H. Wallin, Z. O. Kyjovska, S. S. Poulsen, N. R. Jacobsen, A. T. Saber, S. Bengtson, P. Jackson and U. Vogel, *Mutagenesis*, 2017, **32**, 47–57.

24 A. Saber, N. Jacobsen, A. Mortensen, J. Szarek, P. Jackson, A. Madsen, K. Jensen, I. K. Koponen, G. Brunborg, K. Gützkow, U. Vogel and H. Wallin, *Part. Fibre Toxicol.*, 2012, **9**, 4.

25 A. Saber, A. Mortensen, J. Szarek, N. Jacobsen, M. Levin, I. Koponen, K. Jensen, U. Vogel and H. Wallin, *Hum. Exp. Toxicol.*, 2019, **38**, 11–24.

26 M. Husain, A. T. Saber, C. Guo, N. R. Jacobsen, K. A. Jensen, C. L. Yauk, A. Williams, U. Vogel, H. Wallin and S. Halappanavar, *Toxicol. Appl. Pharmacol.*, 2013, **269**, 250–262.

27 M. Wang, J. Li, S. Dong, X. Cai, A. Simaiti, X. Yang, X. Zhu, J. Luo, L.-H. Jiang, B. Du, P. Yu and W. Yang, *Part. Fibre Toxicol.*, 2020, **17**, 23.

28 S. Halappanavar, P. Jackson, A. Williams, K. A. Jensen, K. S. Hougaard, U. Vogel, C. L. Yauk and H. Wallin, *Environ. Mol. Mutagen.*, 2011, **52**, 425–439.

29 S. Halappanavar, L. Rahman, J. Nikota, S. S. Poulsen, Y. Ding, P. Jackson, H. Wallin, O. Schmid, U. Vogel and A. Williams, *NanoImpact*, 2019, **14**, 100158.

30 D. J. McCarthy and G. K. Smyth, *Bioinformatics*, 2009, **25**, 765–771.

31 M. J. Peart, G. K. Smyth, R. K. van Laar, D. D. Bowtell, V. M. Richon, P. A. Marks, A. J. Holloway and R. W. Johnstone, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 3697–3702.

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

32 A. Raouf, Y. Zhao, K. To, J. Stingl, A. Delaney, M. Barbara, N. Iscove, S. Jones, S. McKinney, J. Emerman, S. Aparicio, M. Marra and C. Eaves, *Cell Stem Cell*, 2008, **3**, 109–118.
33 J. Zschaler, D. Schlorke and J. Arnhold, *Crit. Rev. Immunol.*, 2014, **34**, 433–454.
34 M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown and R. W. Davis, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 10614–10619.
35 J. Lever, M. Krzywinski and N. Altman, *Nat. Methods*, 2017, **14**, 641–642.
36 I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 2002.
37 D. Saha and A. Manickavasagan, *Curr. Res. Food Sci.*, 2021, **4**, 28–44.
38 M. S. Mohammed, N. Paraman, A. A.-H. A. Rahman, F. A. Ghaleb, A. Al-Dhamari and M. N. Marsono, *IEEE Access*, 2021, **9**, 124087–124099.
39 D. Dai, F. Javed, P. Karlsson and K. Månsson, *Ann. Oper. Res.*, 2025, DOI: 10.1007/s10479-025-06486-y.
40 F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau and C. Gagné, *DEAP: Evolutionary Algorithms Made Easy*, 2012, vol. 13.
41 S. Sengottiyan, A. Mikolajczyk, K. Jagiełło, M. Swirog and T. Puzyn, *ACS Nano*, 2023, **17**, 1989–1997.
42 A. O. Aptula, N. G. Jeliazkova, T. W. Schultz and M. T. D. Cronin, *QSAR Comb. Sci.*, 2005, **24**, 385–396.
43 M. Nesterkina, V. Muratov, L. Ognichenko, I. Kravchenko and V. Kuz'min, *Open Chem.*, 2021, **19**, 1184–1192.
44 P. H. Danielsen, S. S. Poulsen, K. B. Knudsen, P. A. Clausen, K. A. Jensen, H. Wallin and U. Vogel, *Environ. Toxicol. Pharmacol.*, 2024, **107**, 104413.
45 J. Nikota, A. Williams, C. L. Yauk, H. Wallin, U. Vogel and S. Halappanavar, *Part. Fibre Toxicol.*, 2016, **13**, 25.
46 N. Hadrup, V. Zhernovkov, N. R. Jacobsen, C. Voss, M. Strunz, M. Ansari, H. B. Schiller, S. Halappanavar, S. S. Poulsen, B. Kholodenko, T. Stoeger, A. T. Saber and U. Vogel, *Small*, 2020, **16**, 1907476.
47 A. T. Saber, N. R. Jacobsen, P. Jackson, S. S. Poulsen, Z. O. Kyjovska, S. Halappanavar, C. L. Yauk, H. Wallin and U. Vogel, *WIREs Nanomed. Nanobiotechnol.*, 2014, **6**, 517–531.
48 C. T. Gutierrez, C. Loizides, I. Hafez, A. Brostrøm, H. Wolff, J. Szarek, T. Berthing, A. Mortensen, K. A. Jensen, M. Roursgaard, A. T. Saber, P. Møller, G. Biskos and U. Vogel, *Part. Fibre Toxicol.*, 2023, **20**, 4.
49 S. McCarrick, V. Malmborg, L. Gren, P. H. Danielsen, M. Tunér, L. Palmberg, K. Broberg, J. Pagels, U. Vogel and A. R. Gliga, *Arch. Toxicol.*, 2025, **99**, 797–814.
50 Y. Ishida, A. Kimura, T. Kondo, T. Hayashi, M. Ueno, N. Takakura, K. Matsushima and N. Mukaida, *Am. J. Pathol.*, 2007, **170**, 843–854.
51 S. Li and P. Yang, *J. Thorac. Dis.*, 2021, **13**, 5517–5533.
52 Z. Zhang and K. K. Parhi, in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2015, pp. 5748–5751.
53 J. Li, S. Xiang and X. Song, *J. Comput. Biol.*, 2024, **31**, 670–690.
54 L. Perez, J. Ambroise, B. Bearzatto, A. Froidure, C. Pilette, Y. Yakoub, M. Palmai-Pallag, C. Bouzin, L. Ryelandt, C. Pavan, F. Huaux and D. Lison, *Part. Fibre Toxicol.*, 2025, **22**, 8.
55 L. Chézeau, S. Sébillaud, R. Safar, C. Seidel, D. Dembélé, M. Lorcin, C. Langlais, S. Grossmann, H. Nunge, S. Michaux, H. Dubois-Pot-Schneider, B. Rihn, O. Joubert, S. Binet, F. Cosnier and L. Gaté, *Toxicol. Appl. Pharmacol.*, 2018, **356**, 54–64.
56 T. M. Sager, C. M. Umbright, G. M. Mustafa, J. R. Roberts, M. S. Orandle, J. L. Cumpston, W. G. McKinney, T. Boots, M. L. Kashon and P. Joseph, *Inhalation Toxicol.*, 2022, **34**, 200–218.
57 C. Guo, M. D. Wright, A. Buckley, A. Laycock, T. Berthing, U. Vogel, F. Cosnier, L. Gaté, M. O. Leonard and R. Smith, *Toxics*, 2025, **13**, 401.
58 S. A. Valentino, L. Chézeau, C. Seidel, S. Sébillaud, M. Lorcin, M. Chalansonnet, F. Cosnier and L. Gaté, *Nanomaterials*, 2021, **11**, 1466.
59 T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T. P. Dasari, A. Michalkova, H.-M. Hwang, A. Toropov, D. Leszczynska and J. Leszczynski, *Nat. Nanotechnol.*, 2011, **6**, 175–178.
60 L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell and P. Gramatica, *Environ. Health Perspect.*, 2003, **111**, 1361–1375.
61 T. Puzyn, N. Jeliazkova, H. Sarimveis, R. L. Marchese Robinson, V. Lobaskin, R. Rallo, A.-N. Richarz, A. Gajewicz, M. G. Papadopulos, J. Hastings, M. T. D. Cronin, E. Benfenati and A. Fernández, *Food Chem. Toxicol.*, 2018, **112**, 478–494.
62 *(Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions*, OECD, 2023.
63 A. Mikolajczyk, N. Sizochenko, E. Mulkiewicz, A. Malankowska, M. Nischk, P. Jurczak, S. Hirano, G. Nowaczyk, A. Zaleska-Medynska, J. Leszczynski, A. Gajewicz and T. Puzyn, *Beilstein J. Nanotechnol.*, 2017, **8**, 2171–2180.
64 A. M. Fresegna, C. L. Ursini, A. Ciervo, R. Maiello, S. Casciardi, S. Iavicoli and D. Cavallo, *Nanomaterials*, 2021, **11**, 253.
65 R. F. Hamilton, N. Wu, D. Porter, M. Buford, M. Wolfarth and A. Holian, *Part. Fibre Toxicol.*, 2009, **6**, 35.
66 J. Okuda-Shimazaki, S. Takaku, K. Kanehira, S. Sonezaki and A. Taniguchi, *Int. J. Mol. Sci.*, 2010, **11**, 2383–2392.
67 J. K. Suthar, B. Rakesh, A. Vaidya and S. Ravindran, *J. Xenobiot.*, 2023, **13**, 662–684.
68 S. S. Mano, K. Kanehira, S. Sonezaki and A. Taniguchi, *Int. J. Mol. Sci.*, 2012, **13**, 3703–3717.
69 K. S. El-Said, E. M. Ali, K. Kanehira and A. Taniguchi, *J. Nanobiotechnol.*, 2014, **12**, 48.
70 W. Li, M. X. Jia, J. Deng, J. H. Wang, Z. Zuberi, S. Yang, J. Ba and Z. Chen, *Cancers*, 2020, **12**, 1236.
71 E. Fuster, H. Candela, J. Estévez, E. Vilanova and M. A. Sogorb, *Int. J. Mol. Sci.*, 2021, **22**, 2084.

This journal is © The Royal Society of Chemistry 2026

72 D. T. Jayaram, A. Kumar, L. E. Kippner, P.-Y. Ho, M. L. Kemp, Y. Fan and C. K. Payne, *RSC Adv.*, 2019, **9**, 25039–25047.
73 M. Alswady-Hoff, J. S. Erdem, S. Phuyal, O. Knittelfelder, A. Sharma, D. de M. Fonseca, Ø. Skare, G. Slupphaug and S. Zienolddiny, *Int. J. Mol. Sci.*, 2021, **22**, 5349.
74 M. G. Domingo, M. Kurtz, G. Maglione, M. Martin, F. Brites, D. R. Tasat and D. G. Olmedo, *J. Biomed. Mater. Res., Part B*, 2024, **112**, e35443.
75 *Overview of Concepts and Available Guidance related to Integrated Approaches to Testing and Assessment (IATA)*, OECD, 2020.
76 S. Halappanavar, M. Sharma, S. Solorio-Rodriguez, H. Wallin, U. Vogel, K. Sullivan and A. Clippinger, *Substance interaction with the pulmonary resident cell membrane components leading to pulmonary fibrosis*, 2023.

Published on 23 February 2026. Downloaded by University of Gdansk on 5/14/2026 5:05:51 PM.

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

The role of machine learning in predicting titanium dioxide nanoparticles induced pulmonary pathology using transcriptomic biomarkers

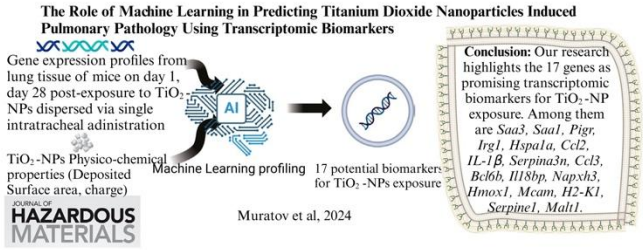
Viacheslav Muratov [a], Karolina Jagiello [a,b,*], Alicja Mikolajczyk [a,b], Pernille Høgh Danielsen [c], Sabina Halappanavar [d,e], Ulla Vogel [c], Tomasz Puzyn [a,b,*]

[a] University of Gdansk, Faculty of Chemistry, Laboratory of Environmental Chemoinformatics, Wita Stwosza 63, Gdansk 80-308, Poland
[b] QSAR Lab Ltd., Trzy lipy 3, Gdansk 80-172, Poland
[c] The National Research Centre for the Working Environment, Copenhagen DK-2100, Denmark
[d] Environmental Health Science and Research Bureau, Health Canada, Ottawa, Ontario K1A 0K9, Canada
[e] Department of Biology, University of Ottawa, Ontario, Canada

HIGHLIGHTS

- 17 genes linked to lung inflammation/fibrosis are actively altered by NP exposure.
- NP size and deposited surface area drive gene changes triggering lung pathologies.
- ML integrates omics data and offers robust tools for detecting adverse effects.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Nanosafety
Transcriptomics
Pulmonary toxicity
Titanium dioxide nanoparticles
Machine learning

ABSTRACT

This study explores the application of machine learning (ML) in identifying transcriptomic changes associated with pulmonary pathologies induced by titanium dioxide nanoparticles (TiO_2-NPs). Such an approach significantly contributes to understanding the underlying mode-of-action of TiO_2-NP inhalation and follows the European Chemicals Agency's recommendations on applying Novel Approach Methodologies designed for reducing animal studies. The lung gene expression profiles from mice exposed via single intratracheal instillations to TiO_2-NPs with varying physicochemical properties on day 1, and day 28 post-exposure were analyzed to develop computational models for predicting the lung pathologies of rutile TiO_2-NPs. More than 600 random forest models were generated and rigorously validated, leading to the identification of 17 high-quality models with an average accuracy of 0.95. These models link nanoparticle-deposited surface area, charge, and post-exposure sampling time with dysregulation in key genes, including serum amyloid *Saa1* (59.7-fold increase), *Saa3* (253.7-fold increase), and the cytokine *Ccl2* (3.4-fold increase). These genes are strongly associated with lung inflammation and fibrosis, key pathological responses to nanomaterial exposure. The study highlights critical nanoparticle features that drive transcriptomic changes. Hierarchical clustering confirmed the mechanistic links between nanoparticle properties and transcriptomic changes. This study demonstrates ML's potential to integrate omics data for nanosafety, offering a robust framework for early detection of adverse effects. The models enable

* Corresponding authors at: University of Gdansk, Faculty of Chemistry, Laboratory of Environmental Chemoinformatics, Wita Stwosza 63, Gdansk 80-308, Poland.
E-mail addresses: karolina.jagiello@ug.edu.pl (K. Jagiello), tomasz.puzyn@ug.edu.pl (T. Puzyn).

https://doi.org/10.1016/j.jhazmat.2025.138240
Received 18 December 2024; Received in revised form 7 April 2025; Accepted 9 April 2025

the prediction of gene expression changes based on nanoparticle features, aiding in potential Safe and Sustainable-by-design of nanomaterials.

1. Introduction

Biomarkers are measurable indicators of abnormal biological processes, organisms' responses to interventions, or xenobiotic exposure [6]. They have found applications in the healthcare system and medical research [12,43,55,69]. For instance, F2-isoprostanes tests are extensively used in measuring lipid peroxidation, commonly associated with diseases such as atherosclerosis [48]. In cardiovascular medicine, biomarkers are essential as they reflect the pathophysiological processes of heart disease, provide meaningful information about prognosis, guide clinical decision-making, and offer potential targets for drug development [13]. Similarly, different types of inflammatory biomarkers, including Interleukin (IL)-6, Tumor Necrosis Factor (TNF)-alpha, C-reactive protein, and other immune-related effectors are actively used as indicators of the magnitude of infection or inflammation in clinical settings, which guide the treatment regimen and in monitoring the effectiveness of interventions [5].

The importance of biomarkers in the hazard assessment of chemicals and human safety biomonitoring is well-recognized. This recognition has been further enhanced by recent advancements in genomics, proteomics, and metabolomics, which, when integrated with bioinformatics, offer a critical understanding of the underlying mechanisms of the perturbed biological processes or toxicity. The omics (e.g., gene, protein, or metabolite expression) biomarkers, measured early after exposure and offering insights into the mode of action (MoA) of stressor-induced pathologies [74], have gained interest in predictive modeling. The European Chemicals Agency (ECHA) highlighted the significance of omics techniques and omics-based biomarkers analysis in chemical risk assessment during a workshop dedicated to New Approach Methodologies (NAMs) [17]. The workshop supported the use of "high--throughput screening" and "high-content techniques" as effective NAMs, indicating that omics will play a crucial role in the regulatory safety assessment of chemicals and nanomaterials/nanoparticles (NMs/NPs). Linking the omics biomarkers that are induced early after exposure and associated with or anchored to a specific toxicity or a disease, with the properties of a stressor might provide a better understanding of how different materials' characteristics influence their toxicity [15,19].

The increasing use of NMs/NPs across various industries (i.e., electrical, optical, and pharmaceutical applications) underscores the need to understand, measure, and manage their risks [40]. Among NMs, titanium oxide NPs (TiO_2-NPs) are widely manufactured and utilized in various consumer products and processes, including photocatalytic processes, paint matrices, sunscreens, cancer treatment, and antibacterial solutions [1]. With the production of TiO_2-NPs and other NMs projected to reach millions of metric tons per year by the end of 2025, their widespread presence in consumer and industrial products raised concerns regarding potential environmental and human impacts. Growing evidence from experimental studies indicates that TiO_2-NPs can induce lung pathologies following inhalation, including inflammation, emphysema, and lung cancer [10,33,67]. These findings suggest that chronic exposure to TiO_2-NPs may contribute to respiratory disease, emphasizing the need for a comprehensive assessment of their toxicity mechanism. In recent years, transcriptomic analyses have provided deeper insights into the molecular responses triggered by TiO_2-NP exposure [42,56,70]. Prior studies have demonstrated that TiO_2-NPs elicit distinct gene expression changes associated with pulmonary inflammation, fibrosis, cytotoxicity, and lipoproteinosis, thereby uncovering potential biomarkers linked to toxicity pathways [36,37,49,71].

Given that inhalation is a primary exposure route in occupational settings during TiO_2-NPs production and product manufacturing, assessing the pulmonary hazards of TiO_2-NPs is critical. Thus, the goal of this study was to examine gene expression changes after lung exposures to several TiO_2-NP types to identify material features that may influence the molecular response and deliver models capable of predicting the changes in gene expression associated with lung responses to TiO_2-NPs. We have applied supervised machine learning (ML) methodologies with the synthetic minority oversampling techniques (SMOTE) that offer robust and efficient alternatives to experimental testing methods and allow to address data imbalance, which is frequently observed while analyzing active vs. nonactive responses. By generating synthetic data points for underrepresented classes, SMOTE enhances model performance and prevents biased predictions, thus improving the reliability and accuracy of the results. Supervised ML models are particularly well-suited for predicting complex dose-response relationships, as they can learn from labeled datasets and generalize patterns to doses with missing data. In this study, we leveraged a range of supervised learning algorithms, selecting the most accurate models based on evaluation metrics such as accuracy, precision, and recall. The application of these models not only maximized predictive power but also facilitated the identification of key features influencing biological responses to nanoparticles. While experimental safety testing is valuable, it is often time-consuming, resource-intensive, and may fail to fully capture the complex interactions between nanoparticles properties and biological responses. ML models, in contrast, offer the ability to handle high-dimensional datasets typical of toxicogenomics, optimizing feature selection, and improving both the accuracy and stability of toxicity prediction and biomarker identification [66]. ML has demonstrated its potential to significantly enhance predictive capabilities across various scientific and industrial domains [41]. For instance, it has been instrumental in optimizing wastewater treatment processes [62]; enabling the early detection of skin cancer through advanced diagnostic algorithms [38]; and enhancing of efficacy of cancer therapies by personalizing treatment protocols [50].

In this study, we explored the influence of primary particle size, deposited surface area, post-exposure sampling duration, and surface charge of rutile-type TiO_2-NPs on the expression of individual or group of genes that might serve as potential biomarkers for lung pathologies induced by TiO_2-NPs [31,36]. A battery of models was developed for over 600 differentially expressed genes, encompassing the entire set of statistically significant gene responses. As a result, a set of genes most actively altered by nanoparticle exposure affected by properties of TiO_2-NPs was identified.

While omics-based biomarkers are recognized for their potential in chemical risk assessment, their practical integration into nanotoxicology is still in its early stages. Previous studies have successfully identified transcriptomic responses following TiO_2-NP exposure using conventional statistical methods, providing valuable insights into biological pathways related to pulmonary inflammation and fibrosis [29,31,30,49,71]. However, such conventional methods often lack the predictive power and ability to simultaneously integrate multiple nanoparticle features, limiting their real-world applicability in early hazard prediction and nanomaterial design.

In this study, we address these limitations by integrating ML approaches with transcriptomic biomarker analysis, enabling the establishment of robust, predictive relationships between TiO_2-NP physicochemical properties (e.g., particle size, deposited surface area, and charge) and specific gene expression changes. Our ML-based framework not only identifies mechanistically relevant biomarkers but also provides quantitative predictions of transcriptomic responses, substantially enhancing the practical utility of omics data. This

predictive capability aligns well with the ECHA emphasis on NAMs for early hazard identification and chemical safety assessments. By facilitating rapid and reliable prediction of adverse biological effects based on nanoparticle characteristics, our approach supports practical applications in regulatory toxicology, Safe-and-Sustainable-by-Design (SSbD) strategies, and occupational health protection, ultimately contributing to the safer development and use of nanomaterials.

2. Materials and methods

In this study, the transcriptomic response from mice exposed to various TiO_2-NPs in previously published studies was analyzed to identify potential biomarkers for lung pathologies induced by TiO_2-NPs and to explore their link with the NP's characteristics. The animal studies [31,36,60,61,72] and the global transcriptional changes in lung tissue [31] were previously reported in detail. We classified genes as responsive or non-responsive, validated our models with independent test sets with supervised classification methods, and applied hierarchical cluster analysis (HCA) for pattern recognition. Fig. 1 illustrates the overall workflow, from transcriptomic data processing and model development to identifying the potential biomarkers for lung pathologies induced by TiO_2-NPs, well-linked with the properties of the inhaled NPs.

2.1. Selection of genes for modeling and classes assignment (Step 1 and Step 2)

Gene expression profiles from lung tissues of 5–7-weeks-old female C57BL/6 mice exposed via single intratracheal instillation to five different TiO_2-NPs measured at 1 day and 28 days post-exposure were collected from previously published studies [31]. The details of animal exposure conditions, sample collection procedures, and data normalization were conducted following established protocols, as described in the original studies [60,61,72]. In brief, we used intratracheal instillation as an exposure method to ensure that the lung-deposited doses were similar. This ensures that biological responses to TiO_2 exposure were directly comparable. Groups of 6–8 female mice were exposed to TiO_2-NPs alongside with negative controls exposed to vehicle and a benchmark group exposed to carbon black NPs (Printex90). The TiO_2-NP doses used in the original experiments were 18, 54, and 162 μg/mouse, selected based on a combination of previous toxicological studies and estimated real-world exposure levels. These exposure levels correspond to 1.5, 5 and 15 eight hours working days at the maximum permitted Danish occupational level for TiO_2 which is 6.0 mg Ti/m^3 or 10 mg TiO_2/m^3 ([31]; A. [60]; A. [61]). The gene expression changes observed by microarray analyses were validated by real-time quantitative PCR (62 of 126 genes [36] and commercially available ELISA arrays were used to assess changes in protein levels of 31 chemokines, cytokines, and their receptors in the lung tissue.

The collected gene profiles were filtered to identify differentially expressed genes (DEGs) that exhibited significant changes in response to at least one TiO_2-NP, using selection criteria of p-value ≤ 0.05 and an absolute fold change (FC) ≥ 1.5. The FC cut-off of 1.5 is grounded in well-established scientific literature, ensuring consistency with previous research in the field [29,31,30,47,53,57]. Thus, employing the same threshold enhances the comparability of our findings with existing studies and ensures methodological consistency. However, since there are also recommendations to adopt more stringent criteria, specifically a p-value ≤ 0.05 and an absolute fold change ≥ 2 [63,76], we have explored how applying these stricter parameters influences the results. Using a threshold of $|FC| = >1.5$ demonstrated a strong relationship between the properties of inhaled TiO_2-NP and the response of the genes, which exhibited the highest FC values. In contrast, models based on the $|FC| = >2$ thresholds showed poor predictability. Based on these findings and the biological relevance demonstrated, we have retained the initial threshold of $|FC| = >1.5$.

Genes meeting the p-value ≤ 0.05 and an absolute fold change thresholds (FC) ≥ 1.5) were classified as DEGs (up/down-regulated, Table S2,), while all others were categorized as non-DEGs (no response). Each gene was then assigned a binary code – '1' for DEGs and '0' for non-DEGs (see Table S2).

2.2. Classification model development and its quality assessment (Step 3 and Step 4)

To develop predictive models for gene expression based on the properties of TiO_2-NPs, we employed a combination of methods: the Random Forest (RF) Classifier and GridSearchCV, both from the Scikit-learn library. The TiO_2-NP properties considered in the models were retrieved from the literature [31] and are summarized in Table 1. XGBoost and LightGBM were used as alternative Decision Tree methods for comparing results from different estimators.

To confirm the predictive power of the obtained models, we performed validation using independent test sets. Specifically, we randomly selected and removed eight data points (approximately 25 % of the total dataset) from the training set, ensuring that these points were not used when calibrating the models. This procedure helped avoid bias and allowed us to assess the generalizability and predictability of our models [45] The retained test sets were then used to validate model predictions, thus demonstrating that trained models can reliably inform future analyses even when tested on data not used in model development [51]. To further support model generalizability, we conducted an external validation using previously published data on *Saa3* mRNA levels determined by quantitative real-time PCR from independent studies [10,25] on TiO_2-NP exposure in lungs where the same experimental protocol for the animal study was used (Table S8). In these two datasets, post-exposure durations ranged from 1 to 180 days, and deposited exposure totals (DET) ranged from 13.2 to 481.3 cm^2. All TiO_2-NPs used in this dataset were uncharged. Fold changes (FC) in gene expression were calculated in two ways: (i) using mean values across the studied

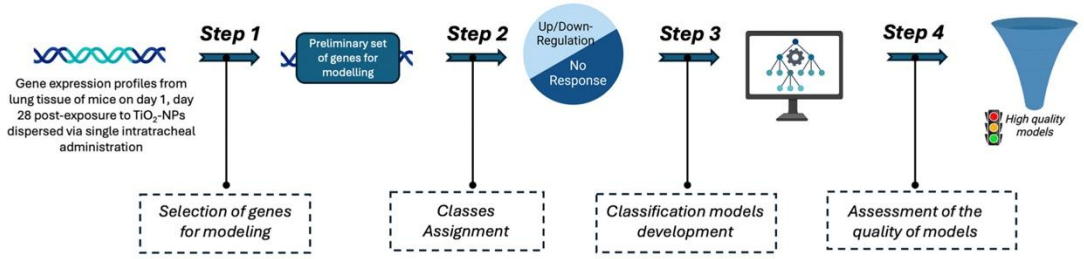


Fig. 1. Overall workflow, representing each step of the study, starting from transcriptomic data processing and finishing with the set of high-quality models that link properties of TiO_2-NPs with biomarkers of lung pathologies.

Table 1
TiO_2-NP types, their physicochemical characteristics, and exposure conditions [31].

TiO_2-NPs_IDs	NPs	Post-exposure duration [days]	Concentration [μg/animal]	Charge*	Size [nm]	BET [m^2/g]	Deposited surface area** [cm^2]
NP_1	NRCWE−001	1	18	0	10	99	17.8
NP_2	NRCWE−001	1	54	0	10	99	53.5
NP_3	NRCWE−001	1	162	0	10	99	160.4
NP_4	NRCWE−002	1	18	1	10	84	15.1
NP_5	NRCWE−002	1	54	1	10	84	45.4
NP_6	NRCWE−002	1	162	1	10	84	136.1
NP_7	NRCWE−001	28	18	0	10	99	17.8
NP_8	NRCWE−001	28	54	0	10	99	53.5
NP_9	NRCWE−001	28	162	0	10	99	160.4
NP_10	NRCWE−002	28	18	1	10	84	15.1
NP_11	NRCWE−002	28	54	1	10	84	45.4
NP_12	NRCWE−002	28	162	1	10	84	136.1
NP_13	NRCWE−030	1	18	0	10.5	139.1	25.0
NP_14	NRCWE−025	1	18	0	38	28.2	5.1
NP_15	NRCWE−030	1	54	0	10.5	139.1	75.1
NP_16	NRCWE−025	1	54	0	38	28.2	15.2
NP_17	NRCWE−030	1	162	0	10.5	139.1	225.3
NP_18	NRCWE−025	1	162	0	38	28.2	45.7
NP_19	NRCWE−030	28	18	0	10.5	139.1	25.0
NP_20	NRCWE−025	28	18	0	38	28.2	5.1
NP_21	NRCWE−030	28	54	0	10.5	139.1	75.1
NP_22	NRCWE−025	28	54	0	38	28.2	15.2
NP_23	NRCWE−030	28	162	0	10.5	139.1	225.3
NP_24	NRCWE−025	28	162	0	38	28.2	45.7
NP_25	UV-Titan L181	1	18	0	20.6	107.7	19.4
NP_26	UV-Titan L181	1	54	0	20.6	107.7	58.2
NP_27	UV-Titan L181	1	162	0	20.6	107.7	174.5
NP_28	UV-Titan L181	28	18	0	20.6	107.7	19.4
NP_29	UV-Titan L181	28	54	0	20.6	107.7	58.2
NP_30	UV-Titan L181	28	162	0	20.6	107.7	174.5

* [0 - no charge; 1 - positive charge]; ** calculated based on specific surface area, size and concentration

Table 2
Statistics of the selected 17 models that link properties of TiO2-NPs with the biomarkers for lung pathologies.

Biomarker	Model calibration					Model validation				
	Accuracy	Precision	Recall	F1	ROC_AUC	Accuracy	Precision	Recall	F1	ROC_AUC
Saa3	0.89	0.94	0.83	0.88	0.89	0.88	1.00	0.83	0.91	0.92
Saa1	0.81	0.81	0.81	0.81	0.81	0.75	0.75	0.75	0.75	0.75
Pigr	0.85	0.85	0.85	0.85	0.85	1.00	1.00	1.00	1.00	1.00
Irg1	0.77	0.80	0.73	0.76	0.77	1.00	1.00	1.00	1.00	1.00
Hspa1a	0.85	0.85	0.85	0.85	0.85	0.88	0.80	1.00	0.89	0.88
Ccl2	0.77	0.73	0.85	0.79	0.77	0.75	0.75	0.75	0.75	0.75
Il1b	0.79	0.82	0.75	0.78	0.79	0.88	1.00	0.75	0.86	0.88
Serpina3n	0.90	0.83	1.00	0.91	0.90	1.00	1.00	1.00	1.00	1.00
Ccl3	0.73	0.73	0.73	0.73	0.73	1.00	1.00	1.00	1.00	1.00
Bcl6b	0.83	0.80	0.89	0.84	0.83	1.00	1.00	1.00	1.00	1.00
Il18bp	0.88	0.88	0.88	0.88	0.88	1.00	1.00	1.00	1.00	1.00
Nxph3	0.82	0.80	0.84	0.82	0.82	1.00	1.00	1.00	1.00	1.00
Hmox1	0.94	0.89	1.00	0.94	0.94	1.00	1.00	1.00	1.00	1.00
Mcam	0.84	0.81	0.89	0.85	0.84	1.00	1.00	1.00	1.00	1.00
H2-K1	0.90	0.86	0.95	0.90	0.90	1.00	1.00	1.00	1.00	1.00
Serpine1	0.84	0.84	0.84	0.84	0.84	1.00	1.00	1.00	1.00	1.00
Malt1	0.88	0.88	0.88	0.88	0.88	1.00	1.00	1.00	1.00	1.00

animal groups, and (ii) using corresponding standard deviation values (Table S8).

Balancing of the training datasets for genes with different distributions of classes was achieved with the SMOTE [7]. This method balances class distributions in datasets by creating synthetic samples for the minority class, addressing the overfitting problem often associated with random oversampling. Thus, the technique effectively generates new data points in the feature space, enhancing the diversity and quantity of the minority class in the dataset.

Additionally, the k-fold cross-validation (cv) was performed with the number of folds k = 5 to evaluate the model's performance. This method divides the dataset by k number of folds and uses k - 1 folds to fit models and 1-fold to test one by one, making it a common practice to determine the quality of models [54]. Accuracy, Precision, Recall, F1-score, Area Under the Receiver Operating characteristic curve (AUC ROC), and the Confusion matrix were used as performance metrics to assess model quality. Accuracy quantifies the proportion of correct predictions among all predictions, while Precision captures the proportion of predicted positive instances that are truly positive, and Recall measures the fraction of actual positive instances that are correctly identified. The F1-score, as the harmonic mean of Precision and Recall, balances these two metrics in a single value. The AUC ROC measures the model's ability to rank positive instances higher than negative ones across a range of decision thresholds, and the Confusion Matrix provides a detailed breakdown of prediction outcomes in terms of true and false positives and negatives. This stands in contrast to regression metrics such as R^2,

RMSE, and MAE, which assess how closely continuous predictions match the actual numeric values, indicating variance explained and average prediction error.

This research used a straightforward approach to calculate classification models' applicability domain (AD) based on standardizing the model's features [58]. The method involved identifying X-outliers in the training set, which were compounds with features highly dissimilar to the rest of the compounds in the training set. The AD was defined as the response and chemical structure space characterized by the properties of the molecules in the training set. This approach simplified AD estimation and enhanced user-friendliness while maintaining transparency and reproducibility in AD assessment. Test set compounds resembling X-outliers likely had inaccurate predictions as the model might not have captured the features of the underrepresented and distinct training set compounds.

Moreover, for exploring and visualizing multivariable datasets the Hierarchical Cluster Analysis (HCA) was applied [22]. Here, two-way HCA and heatmap visualization have been used to investigate the similarities (or dissimilarities) between target TiO_2-NPs and their induced transcriptomic response. The Euclidean distances between each pair of TiO_2-NPs in the space of transcriptomic response (two-fold change) were calculated. Subsequently, HCA with Ward's clustering algorithm was conducted. HCA was performed in R language using the ggplot2 package11.

3. Results and discussion

3.1. Development and validation of machine learning models

The purpose of these studies was to explore the possibilities of applying machine learning approaches to predict gene expression changes that might serve as potential biomarkers of lung pathologies induced by TiO_2-NP exposure. For this purpose, in Step 1, the dataset comprised 621 differentially expressed genes, measured at 1 day and 28 days post-exposure across 30 distinct dose-level combinations for five individual TiO_2-NPs, which varied in size, concentration, and deposited surface area, was collected. The results are summarized in Table S1. The 621 differentially expressed genes that showed significant changes in response to at least one TiO_2-NP (Table S2) were assigned to classes: DEG and non-DEG (Step 2) and linked with the properties of TiO_2-NPs summarized in Table 1 by using an RF Classifier (Step 3). The quality and performance of each predictive model were rigorously assessed (Step 4). Alternatively, we have performed a similar modeling procedure using XGBoost (Table S5) and LightGBM (Table S6) classifiers. The purpose of this additional step was to identify if there are other techniques able to enhance the modeling results and provide more insights into the relationship between the properties of the inhaled TiO_2-NPs and the transcriptomic response. Due to XGBoost and LightGBM did not provide better results, thus the RF models were kept for further analysis.

In effect, a battery of RF models (621 models in total) was developed to accurately associate the properties of TiO_2-NPs with the corresponding gene expression changes. The ML models were trained using parameters like the post-exposure time, the deposited surface area, size, dose of the exposure TiO_2-NPs, and their charge, each of which could potentially influence the resulting toxicological responses.

The nanoparticle deposited surface area (expressed in cm^2) was used to represent TiO_2-NPs doses. This approach facilitates a nuanced assessment of the TiO_2-NPs' dose by incorporating the physical dimensionality alongside the concentration, offering a refined metric for dose quantification that inherently accounts for the variable biological impacts attributable to nanoparticle size [64,9]. The rationale behind this methodology is to ensure differentiation between the actual quantities of TiO_2-NPs administered, which is pivotal for establishing a precise dose-response relationship within the study context. For instance, an examination of two NPs within Table 1, characterized by an identical nominal concentration of 18 μg, reveals a significant discrepancy in the calculated exposure amounts when including the nanoparticle size. This discrepancy underscores the critical influence of nanoparticle size on the effective dose, suggesting the hypothesis that even with equivalent concentrations, the transcriptomic response can significantly diverge based on the NPs' surface area.

Another feature considered within our analysis is the post-exposure time, which helps us not only identify the relation between dose and transcriptomic response but also to see how the transcriptomic response can change over time [14]. It is important to identify how the post-exposure sampling time point affects the transcriptomic response, as it is known that even a year after NM exposure, pulmonary toxicity can be observed [68]. In addition to that, we added a qualitative feature, "Charge", which represents if the TiO_2-NPs are positively charged (value of the feature – 1) or not (the value is 0) (Table 1). The influence of the crystalline structure of the TiO_2-NPs, which, according to previous studies [30], might be crucial since it affects surface area and reactivity, was not considered here as we have only used rutile TiO_2-NPs for training the model.

The physicochemical properties of TiO_2-NPs that were used in this study were previously discussed as affecting lung toxicity [10,31]. The size and the corresponding specific surface area and surface modifications of titanium oxide-based NPs were identified as key factors significantly influencing gene expression profiles. For instance, the observed gene expression response to 38 nm TiO_2-NPs was substantially lower compared to the response induced by 10 nm TiO_2-NPs [31], highlighting the critical role of nanoparticle size in modulating biological effects.

In the next step, developed RF models (621 models in total) were validated. For this purpose, we applied criteria based on established metrics such as Accuracy, Precision, Recall, F1-score, and the AUC score. Accuracy measures the proportion of correct predictions made by the model out of all predictions. Precision, also known as the false positive rate, measures the accuracy of positive predictions. Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives that are correctly identified. The F1-score is the harmonic mean of precision and recall, giving more weight to low values and favoring classifiers with similar precision and recall. The ROC-AUC curve is a graphical representation of the performance of a binary classifier, plotting the true positive rate against the false positive rate for different threshold values. The AUC summarizes the model performance, with a value of 1 indicating perfect accuracy and 0.5 indicating a model that performs no better than random chance. Additionally, the Confusion Matrix was provided to each model, which is a table layout visualizing the performance of an algorithm. It contains information about actual and predicted classifications done by the classifier [16]. The models were tested through k-Fold cross-validation and with external datasets, which helps verify that the models perform well not just on one dataset but consistently across different samples. Table S3 summarizes the detailed information about each developed RF model (~600 models in total), while Fig. 2 presents as example of model performance for predicting the response of the *Saa3* gene. All the calculated models, along with their graphical representations and the Python script used for training, are publicly available in the GitHub repository corresponding to this paper (https://github.com/chembiodev/TransNano-TiO2-Classifier). By making the models and training scripts publicly accessible, this project fosters transparency and encourages further use and refinement of these tools in the study of nanoparticle-induced gene responses.

In addition to the validation through random splitting of the initial dataset into training and test subsets, we also performed an external validation of the model predicting *Saa3* gene response using an independent dataset [10,25]. This dataset included TiO_2-NPs with physicochemical and exposure characteristics not previously seen by the model. The model achieved a high level of balanced accuracy (0.83; Table S8), demonstrating its ability to accurately predict biological responses to novel TiO_2-NP exposures. Table S8 also highlights the sensitivity of the applicability domain (AD): predictions for NPs 19–24 fell outside the AD

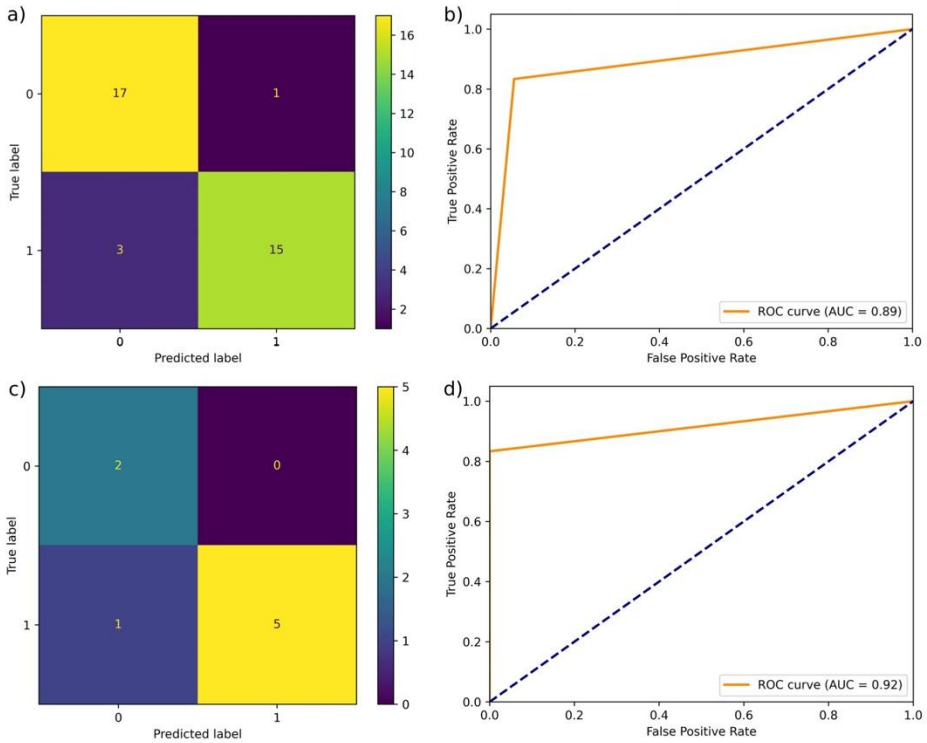


Fig. 2. An example of model performance for predicting the response of the Saa3 gene, with: (a) Confusion Matrix for k-fold cross-validation, (b) ROC-AUC curve for cross-validation, (c) Confusion Matrix for external validation (d) ROC-AUC curve for external validation.

due to extended post-exposure durations and elevated DET values. Despite these outliers, the model retained a high level of predictive performance, further underscoring its robustness and potential utility in supporting toxicological assessments based on transcriptomic responses. Notably, the model also demonstrated strong performance for anatase TiO_2-NPs - a specific crystalline form - highlighting its versatility and suitability for evaluating a wide range of nanomaterial types. This capability enhances its relevance for broader nanosafety applications, especially where comprehensive biological response predictions are required.

3.2. *Property-driven biomarkers identification*

Since our goal was to determine property-driven biomarkers, we next identified the most effective models for predicting gene regulation associated with lung pathologies. To achieve this, models were filtered using an AUC score of 0.7 or higher, meaning they had at least a 70 % chance of correctly distinguishing between positive and negative cases [20,24,35,44]. This approach enabled us to focus on gene responses strongly linked to the properties of inhaled TiO_2-NPs. We also looked at other metrics - Accuracy, Precision, Recall, and F1-score - using the same 0.7 threshold to ensure a comprehensive evaluation.

After this rigorous filtering process, we identified 22 models that met all our criteria, as detailed in Table S4. This table also provides information if any data point from the training or test set (last two columns) is outside of the AD. In the case of our 22 curated models, there was no outlier in any set, suggesting that all the data points are similar to each other enough to provide trustful predictions. These models are considered high-quality and reliable for predicting gene regulation linked to lung diseases. The models are designed to function with TiO_2-NPs that have similar properties to those studied, specifically within the concentration range of 18–162 μg and using the same biological species. This limitation arises from the nature of ML methodologies, which can only generate reliable predictions for data similar to those used during training. Predictions outside the AD should be approached with caution.

Through the analyses of the 22 models summarized in Table S4, we observed that five genes lack informative symbols, which may present challenges for further interpretation and downstream analysis. These genes, while exhibiting significant expression changes in response to TiO_2-NPs, would remain unnamed or unannotated in the current dataset. Thus, they were excluded from further analysis. This refinement resulted in a final list of 17 models meeting rigorous quality criteria corresponding to 17 property-driven genes. These models (Table S7) consider the following genes response: *Saa3, Saa1, Pigr, Irg1, Hspa1a, Ccl2, IL-1β, Serpina3n, Ccl3, Bcl6b, Il18bp, Napxh3, Hmox1, Mcam, H2-K1, Serpine1, Malt1*. The prediction is based on a range of input variables, including material properties and exposure conditions. The summary of model selection, with information about excluded models, is presented in Fig. 3.

The majority of the well-predicted genes (17 genes for which high-quality models were developed) are associated with inflammation and acute phase response (e.g., *Saa3, Saa1, IL-1β, Serpina3n*), which aligns with the observation that inflammation and acute phase responses, key responses to nanomaterial exposure [11,29,49]. Additionally, we identified genes related to lung fibrosis (e.g., *Malt1, Ccl3*) [21,39]. Of note, the acute phase response is causally linked to cardiovascular diseases (*Saa3, Saa1, Il-1β, Serpina3n*) [28,59]. At the protein level, some of the 17 gene products have also been associated with particle-induced

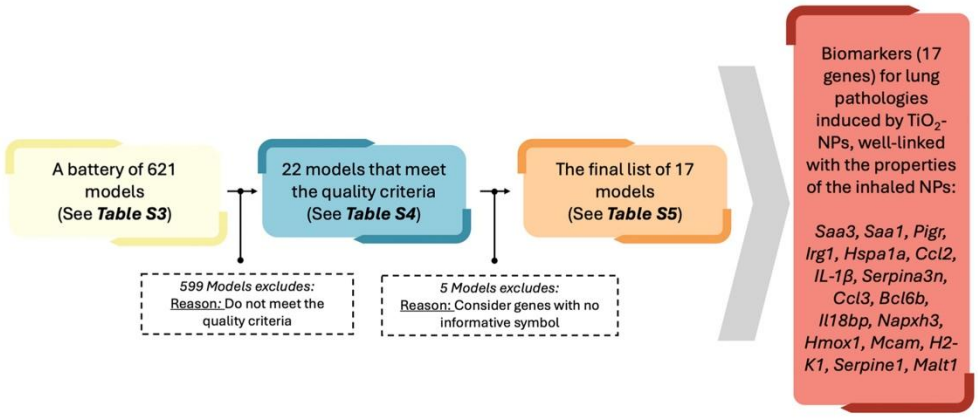


Fig. 3. Workflow of selection of models to predict biomarkers for lung pathologies induced by TiO_2-NPs well-linked with the properties of the inhaled NPs.

inflammation; CCL2, IL1B and CCL3 were all detected in lung tissue following pulmonary exposure to UV-Titan L181 [36]. CCL2 protein levels in broncho-alveolar lavage fluid were also recently shown to correlate with neutrophil influx following pulmonary exposure to diesel exhaust [46], and SAA3 plasma levels were reported to correlate with neutrophil levels following pulmonary exposure to metal oxides [25].

An essential aspect of ML research is the mechanistic interpretation of the results, which provides a deeper understanding of the relationships between input features and model predictions. In our study, we have performed the Feature Importance (FI) technique, which provides valuable insights into the relationship between the properties of exposed TiO_2-NPs and the transcriptomic response. By quantifying the contribution of deposited surface area, charge, and post-exposure time to the model's predictive performance, FI enables to interpret the underlying biological mechanism driving the observed gene expression changes. Examining the relative importance of these features across multiple models allowed us to identify the dominant drivers of gene expression changes and elucidate how variations in nanoparticle properties contribute to the observed biological effects. Table 3 illustrates the FI values for each of the model predicting the selected 17 biomarkers. Correspondingly, Fig. 4 shows the average FIs for these 17 models, providing a visual representation of feature importance across the predictive landscape.

We can see that DET consistently exhibits the highest average impact

Table 3
Feature Importance values for the TiO_2-NPs' DET, Charge and the Post-Exposure Time for each of the model predicting the selected 17 biomarkers.

Biomarkers	Time	DET	Charge
Saa3	0.26	0.60	0.14
Saa1	0.18	0.68	0.14
Pigr	0.24	0.42	0.34
Irg1	0.32	0.44	0.24
Hspa1a	0.28	0.35	0.38
Ccl2	0.20	0.54	0.26
Il1b	0.45	0.20	0.35
Serpina3n	0.40	0.28	0.33
Ccl3	0.39	0.44	0.17
Bcl6b	0.35	0.49	0.16
Il18bp	0.34	0.37	0.29
Nxph3	0.34	0.33	0.33
Hmox1	0.43	0.36	0.21
Mcam	0.34	0.49	0.18
H2-K1	0.38	0.33	0.30
Serpine1	0.30	0.56	0.14
Malt1	0.22	0.39	0.39

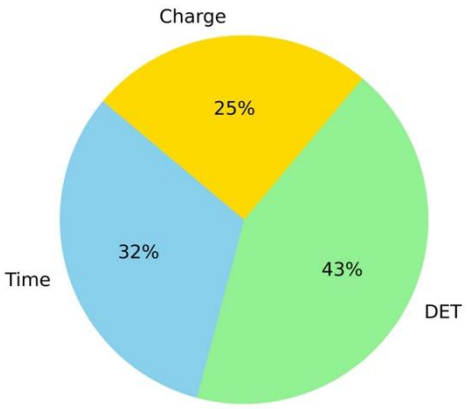


Fig. 4. Average Feature Importances for the models predicting the selected 17 biomarkers.

on the models' predictions, suggesting that the deposited surface area of TiO_2-NPs plays a pivotal role in driving transcriptomic changes. This finding aligns with existing literature, which emphasizes the significance of surface area in nanoparticle-cell interactions, cellular uptake, and subsequent biological responses [31,4,64]. For instance, studies have shown that the size and surface area of nanoparticles can influence cellular uptake efficiency and gene expression profiles. The relationship between particle size and deposited surface area is particularly important, as smaller nanoparticles tend to have a higher surface area per unit mass, which enhances their interaction with biological systems. This can lead to greater cellular uptake, increased bioactivity, and more pronounced effects on gene expression [34]. The deposited surface area also determines the extent of nanoparticle interactions with lung tissue, influencing inflammatory and oxidative stress responses [73].

Moreover, the relatively high FI values for charge indicate its influence on the cellular internalization and distribution of TiO_2-NPs, potentially affecting gene expression through mechanisms related to membrane interaction and cellular signaling. Research has demonstrated that nanoparticle surface charge significantly impacts cellular uptake and subsequent biological responses [18]. The role of surface charge is also critical, as positively charged nanoparticles generally exhibit higher cellular uptake and prolonged retention, which may

amplify gene expression changes over time, particularly in pathways related to oxidative stress and immune responses.

Meanwhile, post-exposure sampling time contributes to the temporal dynamics of transcriptomic responses, reflecting how the duration of exposure modulates the biological impact of nanoparticles. A short-term exposure to high-surface-area nanoparticles is often associated with acute inflammation and immune activation, while longer exposure durations may contribute to persistent inflammation, immune dysregulation, and potential fibrotic remodeling [2,52].

The interplay of these factors suggests that nanoparticle-induced gene expression changes are not driven by a single property but rather by a combination of physical and chemical characteristics that influence toxicity. A comprehensive analysis integrating these parameters is essential for understanding the complexity of nanomaterial interactions with biological systems.

The findings of this study highlight the critical importance of developing machine learning models to establish links between the physicochemical properties of TiO_2-NPs and their transcriptomic responses or transcriptomic-based biomarkers of lung pathologies. Such models not only enhance our understanding of the molecular mechanisms underlying nanoparticle-induced toxicity but also enable the prediction of potential health effects based on a limited set of key nanoparticle features, such as surface area, charge, and post-exposure periods. This predictive capability is essential for early hazard identification, allowing the development of safer nanomaterials with minimal adverse biological effects. Thus, these models provide a valuable tool for predicting gene expression changes in response to TiO_2-NP exposure, supporting the Safe and Sustainable-by-Design (SSbD) approach in nanotechnology.

By leveraging ML, it becomes possible to integrate complex transcriptomic data and NP characteristics, ultimately advancing the field of nanotoxicology and contributing to the development of safer nanotechnology applications. The predictive models developed in this study offer a novel approach to assessing the toxicity of TiO_2-NPs through transcriptomic biomarkers, aligning well with the ECHA guidelines on integrating omics data into regulatory risk assessment. ECHA has recognized the importance of NAMs and high-throughput transcriptomic techniques in chemical safety evaluations, particularly in the context of adverse outcome pathways (AOPs) and mechanistic toxicology [17]. Thus, our models are important for the scientific and regulatory integrity of the studied TiO_2-NPs.

3.3. Revealing molecular signatures of TiO_2-NP exposure: insights from clustering approaches

Although the developed models provide a powerful framework for exploring the relationship between NP characteristics and biological responses, enabling researchers to predict the behavior of specific genes under various experimental conditions, they are not without limitations. The primary constraint of these models lies in the binary classification of genes into two categories: non-responsive ($|FC < 1.5$) and significantly affected ($|FC \geq 1.5$). This classification scheme, while simplifying data

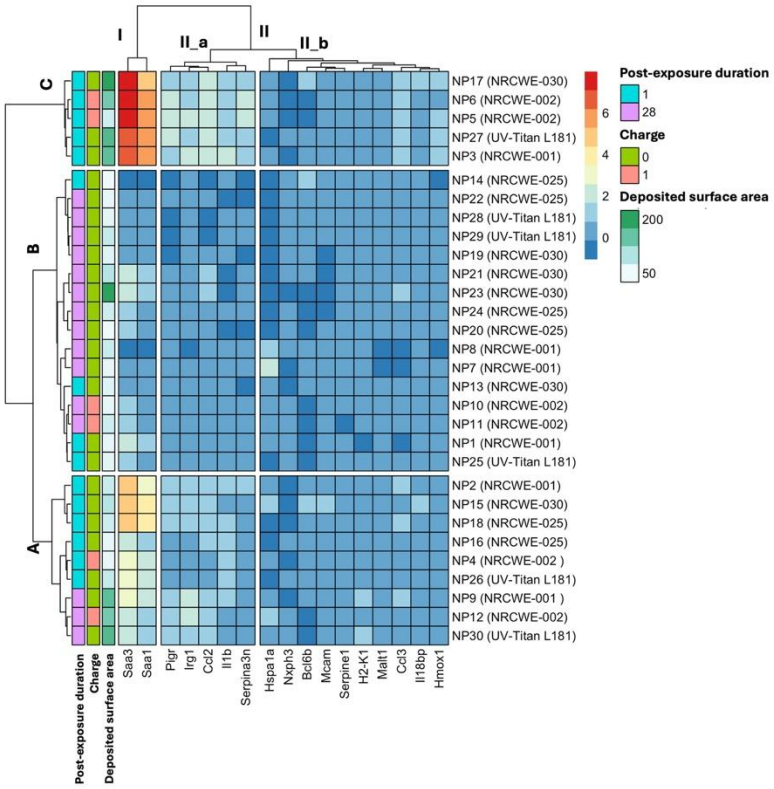


Fig. 5. The heat map represents similarities and dissimilarities between TiO2-NPs and their transcriptomic response expressed as log2FC (Euclidean distance and Ward's method for clustering). Charge: 0 – neutral, 1 – positively charged; post-exposure duration: 1 day or 28 days; Dose expressed as Deposited Surface Area.

interpretation, results in a substantial loss of valuable information regarding the direction and magnitude of gene expression changes (i.e., upregulation or downregulation). The binary cutoff approach ignores the subtle gradations of gene expression changes that could hold critical insights into the molecular mechanisms underlying nanoparticle-induced biological effects. Thus, to gain further insight into the influence of the studied properties on well-predicted genes (17 genes listed in Table S7), we have performed a hierarchical cluster analysis (HCA) using two-fold change (log2FC) of the most structural properties-responsive biomarkers (Fig. 5). By using log2FC instead of classes, we focused on the genes most affected, identified by the highest values of FC, while retaining information about their responses. This analysis offers several advantages. First, it preserves the granularity of gene expression data, enabling a more nuanced understanding of how specific genes are affected. Second, it allows for the identification of materials with similar response patterns, which can provide insights into shared regulatory pathways and mechanisms of action.

Among the well-predicted genes are *Saa3* and *Saa1*, both of which exhibit significant upregulation in response to TiO_2-NPs clustered in clusters A and C. However, the highest perturbations of these genes are observed for TiO_2-NPs grouped in Cluster C (Fig. 5), which represents short-term exposures to materials with the highest deposited surface area. This finding underscores a potential mechanistic link between nanoparticle surface area, size, and dose, and the mRNA levels of *Saa1* and *Saa3*. Further evidence for this connection comes from the observation that TiO_2-NPs with lower deposited surface areas result in less pronounced upregulation of *Saa1* and *Saa3*, after 1-day exposure (Cluster A, Fig. 5)

Predicting *Saa3* expression is particularly noteworthy, as its upregulation is associated with the recruitment of inflammatory cells mediated by acute-phase response proteins and cytokines such as CCL2 and CCL3 [23]. *Saa3* is not only a critical biomarker for particle-induced lung inflammation but also plays a significant role in systemic inflammatory responses linked to cardiovascular diseases. Previous studies have established that the acute-phase protein response, including SAA family proteins, is a hallmark of particle-induced toxicity, correlating with both local and systemic inflammation [25,27,29,3].

Beyond cardiovascular implications, NPs have been shown to translocate from the lungs into the systemic circulation, where they can interact with the liver, kidneys, and immune system, potentially triggering multi-organ effects [37]. Prior studies indicate that exposure to certain nanoparticles leads to hepatic inflammation, dyslipidemia, and metabolic alterations, suggesting a broader impact on systemic homeostasis [25,3]. Additionally, NMs have been implicated in neuroinflammation, with emerging evidence suggesting that inhaled NPs may cross the blood-brain barrier, leading to oxidative stress and neuronal dysfunction [75].

The expression of *Saa3* in lung tissue correlates strongly with other key inflammatory markers, such as chemokines (e.g., *Ccl2* and *Ccl3*) and cytokines (e.g., *IL-6* and *TNF-α*) [65]. This underscores its role as a central mediator in orchestrating the inflammatory cascade triggered by TiO_2-NPs exposure. Similar observations have been reported in nanoparticle-induced inflammation models, where *Saa3* has been implicated in amplifying immune cell recruitment and promoting a sustained pro-inflammatory state.

The 28 days recovery period following a single dose exposure to materials with low deposited surface area (classified within Cluster B) has a relatively limited impact on genes like *Saa1* and *Saa3*. In this scenario, the upregulation of *Saa1* and *Saa3* is only observed for the highest dose levels in terms of surface area, like NP-9, NP-12, NP-23, and NP30, here upregulation in genes like *Ccl2* and *Ccl3* is also shown. This finding aligns with existing evidence suggesting that longer post-exposure durations to particulate matter often tend to suppress the expression of genes associated with acute inflammatory responses, favoring pathways involved in sustained immune activation and tissue remodeling [75].

The conducted studies, encompassing both RF models and HCA, highlight the critical importance of considering the total surface area of the pulmonary deposited particles, which is influenced by the size and dose of TiO_2-NPs, as well as post-exposure duration when evaluating their health impacts. Short exposure to high-surface-area NPs is associated with acute inflammation and acute phase response, whereas late response has been shown to potentially drive persistent inflammation, disrupt immune regulation, and contribute to fibrosis or other long-term pathologies. These findings underscore the necessity for detailed dose-response studies to comprehensively characterize these effects and inform risk assessment strategies.

Despite the promising results, this study has several limitations. First, our analysis relies on publicly available transcriptomic datasets, which, while rigorously controlled, may not fully capture the diversity of exposure conditions and biological responses across different experimental settings. Since our model is trained on a specific dataset derived from a controlled exposure setting, its applicability to real-world exposure conditions - such as occupational or environmental inhalation, chronic low-dose exposure, or co-exposures with other pollutants - requires further investigation. Second, the binary classification of gene expression ($|FC| < 1.5$ as non-responsive and $|FC| \geq 1.5$ as significantly affected) may lead to the loss of nuanced regulatory patterns, potentially overlooking genes with moderate but biologically relevant changes. On the other hand, the studies included were all performed using the same experimental protocol. We used pulmonary dosing, which ensures that the lung burden was similar between studies. In one of the studies, the observed changes in global transcription were validated at gene expression level by PCR and at protein level by quantification of protein levels in lung tissue. The physico-chemical properties of the studied TiO_2-NPs were experimentally determined.

The variability in transcriptomic responses between species, such as human versus murine models might also represent a significant limitation in these studies. While mouse models are widely used in nanotoxicology due to their genetic and physiological similarities to humans, key differences in immune responses, metabolic processing, and nanoparticle clearance mechanisms may affect the direct extrapolation of results [77]. To address this, future research should incorporate human-relevant models, such as primary lung epithelial cells, 3D organotypic cultures, and lung-on-chip systems, to enhance translational applicability [8].

By addressing these limitations and building upon our findings, we can further refine predictive models and improve the assessment of nanoparticle-induced biological effects, ultimately contributing to the safe and responsible development of nanotechnology.

4. Conclusions

Our research highlights the 17 genes as promising transcriptomic biomarkers for TiO_2-NP exposure, which may be predicted based on material properties. Among them are *Saa3, Saa1, Pigr, Irg1, Hspa1a, Ccl2, IL-1β, Serpina3n, Ccl3, Bcl6b, Il18bp, Napxh3, Hmox1, Mcam, H2-K1, Serpine1, Malt1*. These biomarkers might be linked to pathways associated with pulmonary inflammation and fibrosis, which are critical endpoints in nanosafety evaluations. These findings are consistent with previous studies demonstrating the role of inflammatory and acute-phase response genes in nanomaterial-induced toxicity [30,49]. Such molecular markers might be proposed for use in regulatory frameworks to establish dose-response relationships and predict long-term health effects.

The identified 17-gene signature not only provides predictive power for assessing lung response to TiO_2-NP exposure but also reflects biologically meaningful processes relevant to adverse outcomes. Several genes in the signature, such as *Saa3*, *Lcn2*, *Cxcl5*, *Ccl2*, *Il1b*, *Timp1*, and *Serpine1* are associated with inflammatory signaling, extracellular matrix remodeling, and immune cell recruitment. These processes represent key events in adverse outcome pathways (AOPs) linked to

cardiovascular disease, pulmonary fibrosis, and chronic inflammation [26,32].

For instance, *Saa3* and *Lcn2* are classic acute-phase proteins indicative of early inflammatory responses. *Timp1* and *Serpine1* are involved in matrix regulation and fibrosis, while *Ccl2* and *Il1b* mediate monocyte and macrophage activation - mechanistic steps frequently observed in lung injury and remodeling. As such, the gene signature captures both initiating and propagating events in the toxicity cascade, providing mechanistic relevance beyond empirical pattern recognition.

Importantly, this mechanistic underpinning enables translational applications. Many of the genes in the signature are responsive in *in vitro* models using relevant lung cell types, such as alveolar macrophages and epithelial cells. This opens the possibility of developing simplified, high-throughput *in vitro* assays guided by this gene panel to support nanoparticle hazard screening and prioritization. Additionally, the signature could be integrated into *in silico* toxicogenomic frameworks to inform read-across strategies or to simulate key biological transitions in AOP-informed computational models. By linking predictive outcomes to known biological pathways, the model facilitates the interpretation of transcriptomic changes and supports its practical application in both regulatory and research contexts.

Developed models enhance our ability to understand and detect the effects of TiO_2-NP exposure at an early stage. By developing ML models capable of predicting biomarkers for TiO_2-NP exposure, we contribute to the creation of safer nano-related products and practices, ultimately protecting the environment through precise and early detection of nanoparticle exposure. RF models can be implemented in the early design of nanomaterials to preemptively mitigate health risks by enabling predictive toxicology and safer material design. These models can assess the potential lung pathologies associated with different TiO_2-NP formulations before synthesis, allowing researchers to modify physicochemical properties such as size, surface coating, and functionalization to reduce adverse biological effects. This approach allows for the ranking of design candidates based on their predicted impact on transcriptomic biomarkers, prioritizing safer options for further experimental validation.

By linking nanoparticle physicochemical properties with transcriptomic responses, our approach contributes to a more mechanistic understanding of TiO_2-NP toxicity as it relates to their unique properties, which is essential for regulatory decision-making. Expanding the model's applicability domain to include a broader range of NMs and exposure scenarios will be a crucial step toward regulatory validation. Additionally, future efforts should explore the use of multi-omics approaches (e.g., proteomics, metabolomics) to strengthen the mechanistic link between TiO_2-NP exposure and adverse health outcomes, further reinforcing the regulatory applicability of omics-based biomarkers.

Additionally, our study demonstrates the power of machine learning in toxicogenomics, providing a robust framework for future research in nanomaterial risk assessment. Integrating machine learning models with transcriptomic data offers a powerful approach for predicting gene expression changes and understanding the molecular mechanisms of nanomaterial-induced toxicity. This strategy enhances the accuracy and reliability of toxicity predictions, supporting the development of targeted mitigation strategies to minimize the adverse effects of NMs.

Environmental implication

This study demonstrates the critical role of machine learning (ML) in predicting the toxicological effects of titanium dioxide nanoparticles (TiO_2-NPs) based on their physicochemical properties. By identifying transcriptomic changes linked to pulmonary inflammation and fibrosis, the findings offer a novel framework for assessing nanoparticle safety. The research highlights the pivotal role of nanoparticle size and deposited surface area in determining biological responses, underscoring the necessity for the development of safe and sustainable nanomaterials. This approach can minimize environmental and human health risks by enabling the early detection of hazardous effects, informing regulatory standards, and promoting the development of safer nanotechnology applications.

Authors contributions

KJ conceptualization. VM performed formal analysis, including data curation, the set of model's development, and validation. VM and KJ designed and wrote the manuscript. AM, PHD, SH, UV and TP review and editing. TP and KJ funding acquisition and project administration. All authors discussed the results and commented on the manuscript.

CRediT authorship contribution statement

Muratov Viacheslav: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jagiello Karolina:** Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Høgh Danielsen Pernille:** Writing – review & editing. **Vogel Ulla:** Writing – review & editing, Methodology, Data curation. **Mikolajczyk Alicja:** Writing – review & editing, Validation. **Halappanavar Sabina:** Writing – review & editing, Methodology, Investigation. **Puzyn Tomasz:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded via the National Science Centre in the frame of the TransNANO project (UMO-2020/37/B/ST5/01894) with the support of the European Union Horizon 2020 research and innovation programme under grant No.101008099 (CompSafeNano). SH acknowledges research funds received from Health Canada's Genomics Research and Development Initiative. UV received funding from FFIKA2, Focused Research Effort on Chemicals in the Working Environment, from the Danish Government.

Figures for this work were prepared using the software BioRender.com. The computations were enabled by resources provided by the Centre of Informatics TASK (CI TASK), Gdansk University of Technology.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2025.138240.

Data availability

I have included the data in the attached supplementary materials. Additionally, the code for model development is available on GitHub (link in the manuscript).

References

[1] Anandgaonker, P., Kulkarni, G., Gaikwad, S., Rajbhoj, A., 2019. Synthesis of TiO2 nanoparticles by electrochemical method and their antibacterial application. Arab J Chem 12 (8), 1815–1822. https://doi.org/10.1016/j.arabjc.2014.12.015.

[2] Bonner, J.C., 2010. Nanoparticles as a potential cause of pleural and interstitial lung disease. Proc Am Thorac Soc 7 (2), 138–141. https://doi.org/10.1513/pats.200907-061RM.

[3] Bourdon, J.A., Halappanavar, S., Saber, A.T., Jacobsen, N.R., Williams, A., Wallin, H., et al., 2012. Hepatic and pulmonary toxicogenomic profiles in mice intratracheally instilled with carbon black nanoparticles reveal pulmonary inflammation, acute phase response, and alterations in lipid homeostasis. Toxicol Sci 127 (2), 474–484. https://doi.org/10.1093/toxsci/kfs119.

[4] Boyadzhiev, A., Wu, D., Avramescu, M.-L., Williams, A., Rasmussen, P., Halappanavar, S., 2023. Toxicity of metal oxide nanoparticles: looking through the lens of toxicogenomics. Int J Mol Sci 25 (1), 529. https://doi.org/10.3390/ijms25010529.

[5] Brenner, D.R., Scherer, D., Muir, K., Schildkraut, J., Boffetta, P., Spitz, M.R., et al., 2014. A review of the application of inflammatory biomarkers in epidemiologic cancer research. Cancer Epidemiol, Biomark Prev 23 (9), 1729–1751. https://doi.org/10.1158/1055-9965.EPI-14-0064.

[6] Califf, R.M., 2018. Biomarker definitions and their applications. Exp Biol Med 243 (3), 213–221. https://doi.org/10.1177/1535370217750088.

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16, 321–357. https://doi.org/10.1613/jair.953.

[8] Ciura, K., Moschini, E., Stępnik, M., Serchi, T., Gutleb, A., Jarzyńska, K., et al., 2024. Toward nano-specific in silico NAMs: how to adjust nano-QSAR to the recent advancements of nanotoxicology? Small 20 (6). https://doi.org/10.1002/smll.202305581.

[9] Cosnier, F., Seidel, C., Valentino, S., Schmid, O., Bau, S., Vogel, U., et al., 2021. Retained particle surface area dose drives inflammation in rat lungs following acute, subacute, and subchronic inhalation of nanomaterials. Part Fibre Toxicol 18 (1), 29. https://doi.org/10.1186/s12989-021-00419-w.

[10] Danielsen, P.H., Knudsen, K.B., Štrancar, J., Umek, P., Koklič, T., Garvas, M., et al., 2020. Effects of physicochemical properties of TiO2 nanomaterials for pulmonary inflammation, acute phase response and alveolar proteinosis in intratracheally exposed mice. Toxicol Appl Pharmacol 386, 114830. https://doi.org/10.1016/j.taap.2019.114830.

[11] Danielsen, P.H., Poulsen, S.S., Knudsen, K.B., Clausen, P.A., Jensen, K.A., Wallin, H., et al., 2024. Physicochemical properties of 26 carbon nanotubes as predictors for pulmonary inflammation and acute phase response in mice following intratracheal lung exposure. Environ Toxicol Pharmacol 107, 104413. https://doi.org/10.1016/j.etap.2024.104413.

[12] de Araujo, L.S., Ribeiro-Alves, M., Wipperman, M.F., Vorkas, C.K., Pessler, F., Saad, M.H.F., 2021. Transcriptomic biomarkers for tuberculosis: validation of NPC2 as a single mRNA biomarker to diagnose TB, predict disease progression, and monitor treatment response. Cells 10 (10), 2704. https://doi.org/10.3390/cells10102704.

[13] Dhingra, R., Vasan, R.S., 2017. Biomarkers in cardiovascular disease: statistical assessment and section on key novel heart failure biomarkers. Trends Cardiovasc Med 27 (2), 123–133. https://doi.org/10.1016/j.tcm.2016.07.005.

[14] Dhruba, S.R., Rahman, A., Rahman, R., Ghosh, S., Pal, R., 2019. Recursive model for dose-time responses in pharmacological studies. BMC Bioinforma 20 (S12), 317. https://doi.org/10.1186/s12859-019-2831-4.

[15] Ebner, J.N., 2021. Trends in the application of "omics" to ecotoxicology and stress ecology. Genes 12 (10), 1481. https://doi.org/10.3390/genes12101481.

[16] Egwim, C.N., Alaka, H., Toriola-Coker, L.O., Balogun, H., Sunmola, F., 2021. Applied artificial intelligence for predicting construction projects delay. Mach Learn Appl 6, 100166. https://doi.org/10.1016/j.mlwa.2021.100166.

[17] European Chemicals Agency. (2016). *New approach methodologies in regulatory science – Proceedings of a scientific workshop – Helsinki,19–20 April 2016*. ECHA. https://doi.org/10.2823/543644.

[18] Forest, V., Pourchez, J., 2017. Preferential binding of positive nanoparticles on cell membranes is due to electrostatic interactions: a too simplistic explanation that does not take into account the nanoparticle protein corona. Mater Sci Eng: C 70, 889–896. https://doi.org/10.1016/j.msec.2016.09.016.

[19] Fortino, V., Kinaret, P.A.S., Fratello, M., Serra, A., Saarimäki, L.A., Gallud, A., et al., 2022. Biomarkers of nanomaterials hazard from multi-layer data. Nat Commun 13 (1), 3798. https://doi.org/10.1038/s41467-022-31609-5.

[20] Franklin, R.J., Mohana, Dabbagol, V., 2020. Anomaly detection in videos for video surveillance applications using neural networks. 2020 Fourth Int Conf Invent Syst Control (ICISC) 632–637. https://doi.org/10.1109/ICISC47916.2020.9171212.

[21] Fusco, R., Siracusa, R., D'Amico, R., Cordaro, M., Genovese, T., Gugliandolo, E., et al., 2020. Mucosa-associated lymphoid tissue lymphoma translocation 1 inhibitor as a novel therapeutic tool for lung injury. Int J Mol Sci 21 (20), 7761. https://doi.org/10.3390/ijms21207761.

[22] Granato, D., Santos, J.S., Escher, G.B., Ferreira, B.L., Maggio, R.M., 2018. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: a critical perspective. Trends Food Sci Technol 72, 83–90. https://doi.org/10.1016/j.tifs.2017.12.006.

[23] Groves, D.T., Jiang, Y., 1995. Chemokines, a family of chemotactic cytokines. Crit Rev Oral Biol Med 6 (2), 109–118. https://doi.org/10.1177/10454411950060020101.

[24] Guéniche, N., Huguet, A., Bruyere, A., Habauzit, D., Le Hégarat, L., Fardel, O., 2021. Comparative in silico prediction of P-glycoprotein-mediated transport for 2010–2020 US FDA-approved drugs using six Web-tools. Biopharm Drug Dispos 42 (8), 393–398. https://doi.org/10.1002/bdd.2299.

[25] Gutierrez, C.T., Loizides, C., Hafez, I., Brostrøm, A., Wolff, H., Szarek, J., et al., 2023. Acute phase response following pulmonary exposure to soluble and insoluble metal oxide nanomaterials in mice. Part Fibre Toxicol 20 (1), 4. https://doi.org/10.1186/s12989-023-00514-0.

[26] Gutierrez, C., Poulsen, S., Sørli, J., Wallin, H., Halappanavar, S., Yauk, C., (2025). Substance interaction with lung resident cell membrane components leading to atherosclerosis. AOP-Wiki. aopwiki.org/aops/237.

[27] Hadrup, N., Rahmani, F., Jacobsen, N.R., Saber, A.T., Jackson, P., Bengtson, S., et al., 2019. Acute phase response and inflammation following pulmonary exposure to low doses of zinc oxide nanoparticles in mice. Nanotoxicology 13 (9), 1275–1292. https://doi.org/10.1080/17435390.2019.1654004.

[28] Hadrup, N., Zhernovkov, V., Jacobsen, N.R., Voss, C., Strunz, M., Ansari, M., et al., 2020. Acute phase response as a biological mechanism-of-action of (nano)particle-induced cardiovascular disease. Small 16 (21). https://doi.org/10.1002/smll.201907476.

[29] Halappanavar, S., Jackson, P., Williams, A., Jensen, K.A., Hougaard, K.S., Vogel, U., et al., 2011. Pulmonary response to surface-coated nanotitanium dioxide particles includes induction of acute phase response genes, inflammatory cascades, and changes in microRNAs: a toxicogenomic study. Environ Mol Mutagen 52 (6), 425–439. https://doi.org/10.1002/em.20639.

[30] Halappanavar, S., Rahman, L., Nikota, J., Poulsen, S.S., Ding, Y., Jackson, P., et al., 2019. Ranking of nanomaterial potency to induce pathway perturbations associated with lung responses. NanoImpact 14, 100158. https://doi.org/10.1016/j.impact.2019.100158.

[31] Halappanavar, S., Saber, A.T., Decan, N., Jensen, K.A., Wu, D., Jacobsen, N.R., et al., 2015. Transcriptional profiling identifies physicochemical properties of nanomaterials that are determinants of the in vivo pulmonary response. Environ Mol Mutagen 56 (2), 245–264. https://doi.org/10.1002/em.21936.

[32] Halappanavar, S., Sharma, M., Solorio-Rodriguez, S., Wallin, H., Vogel, U., Sullivan, K., (2023). *OECD Series on Adverse Outcome Pathways*, No. 33, OECD Publishing, Paris, https://doi.org/10.1787/10372cb8-en.

[33] Heinrich, U., Fuhst, R., Rittinghausen, S., Creutzenberg, O., Bellmann, B., Koch, W., et al., 1995. Chronic inhalation exposure of wistar rats and two different strains of mice to diesel engine exhaust, carbon black, and titanium dioxide. Inhal Toxicol 7 (4), 533–556. https://doi.org/10.3109/08958379509015211.

[34] Hoshyar, N., Gray, S., Han, H., Bao, G., 2016. The effect of nanoparticle size on in vivo pharmacokinetics and cellular interaction. Nanomedicine 11 (6), 673–692. https://doi.org/10.2217/nnm.16.5.

[35] Hosseini, M., Ramazani, J., 2016. Evaluation of acute physiology and chronic health evaluation II and sequential organ failure assessment scoring systems for prognostication of outcomes among Intensive Care Unit's patients. Saudi J Anaesth 10 (2), 168. https://doi.org/10.4103/1658-354X.168817.

[36] Husain, M., Saber, A.T., Guo, C., Jacobsen, N.R., Jensen, K.A., Yauk, C.L., et al., 2013. Pulmonary instillation of low doses of titanium dioxide nanoparticles in mice leads to particle retention and gene expression changes in the absence of inflammation. Toxicol Appl Pharmacol 269 (3), 250–262. https://doi.org/10.1016/j.taap.2013.03.018.

[37] Husain, M., Wu, D., Saber, A.T., Decan, N., Jacobsen, N.R., Williams, A., et al., 2015. Intratracheally instilled titanium dioxide nanoparticles translocate to heart and liver and activate complement cascade in the heart of C57BL/6 mice. Nanotoxicology 9 (8), 1013–1022. https://doi.org/10.3109/17435390.2014.996192.

[38] Irianto, S.Y., Yunandar, R., Hasibuan, M.S., Dewi, D.A., Pitsachart, N., 2024. Early identification of skin cancer using region growing technique and a deep learning algorithm. HighTech Innov J 5 (3), 640–662. https://doi.org/10.28991/HIJ-2024-05-03-07.

[39] Ishida, Y., Kimura, A., Kondo, T., Hayashi, T., Ueno, M., Takakura, N., et al., 2007. Essential roles of the CC chemokine ligand 3-CC chemokine receptor 5 axis in bleomycin-induced pulmonary fibrosis through regulation of macrophage and fibrocyte infiltration. Am J Pathol 170 (3), 843–854. https://doi.org/10.2353/ajpath.2007.051213.

[40] Jafari, S., Mahyad, B., Hashemzadeh, H., Janfaza, S., Gholikhani, T., Tayebi, L., 2020. Biomedical applications of TiO2 nanostructures: recent advances. Int J Nanomed 15, 3447–3470. https://doi.org/10.2147/IJN.S249441.

[41] Li, Yan, Huang, Haoran, Li, Ye, Ye, Zi, Li, Xiang, Liu, Ke, et al., 2025. Characterizing soil COPs eco-risk in China. J Hazard Mater 489, 137588. https://doi.org/10.1016/j.jhazmat.2025.137588.

[42] Li, F., Liu, J., Gao, Z., Yang, C., Sun, L., Liao, Y., et al., 2024. Transcriptomic analyses of Pinctada fucata martensii responses under stress of titanium dioxide nanoparticles. Front Mar Sci 11. https://doi.org/10.3389/fmars.2024.1462589.

[43] Li, X., Zhou, X., Ding, S., Chen, L., Feng, K., Li, H., et al., 2022. Identification of transcriptome biomarkers for severe COVID-19 with machine learning methods. Biomolecules 12 (12), 1735. https://doi.org/10.3390/biom12121735.

[44] Ling, Z., Liu, X., Luo, Y., Wu, X., Yuan, L., Tong, X., et al., 2013. Associations between vaginal pathogenic community and bacterial vaginosis in Chinese reproductive-age women. PLoS ONE 8 (10), e76589. https://doi.org/10.1371/journal.pone.0076589.

[45] Masand, Vijay H., Mahajan, Devidas T., Nazeruddin, Gulam M., Hadda, Taibi Ben, Rastija, Vesna, Alfeefy, Ahmed M., 2015. Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model. Med Chem Res 24 (3), 1241–1264. https://doi.org/10.1007/s00044-014-1193-8.

[46] McCarrick, S., Malmborg, V., Gren, L., Danielsen, P.H., Tunér, M., Palmberg, L., et al., 2025. Pulmonary exposure to renewable diesel exhaust particles alters protein expression and toxicity profiles in bronchoalveolar lavage fluid and plasma of mice. Arch Toxicol 99 (2), 797–814. https://doi.org/10.1007/s00204-024-03915-y.

[47] McCarthy, D.J., Smyth, G.K., 2009. Testing significance relative to a fold-change threshold is a TREAT. Bioinforma (Oxf, Engl) 25 (6), 765–771. https://doi.org/10.1093/bioinformatics/btp053.

[48] Morrow, J.D., 2005. Quantification of isoprostanes as indices of oxidant stress and the risk of atherosclerosis in humans. Arterioscler, Thromb, Vasc Biol 25 (2), 279–286. https://doi.org/10.1161/01.ATV.0000152605.64964.c0.
[49] Nikota, J., Williams, A., Yauk, C.L., Wallin, H., Vogel, U., Halappanavar, S., 2015. Meta-analysis of transcriptomic responses as a means to identify pulmonary disease outcomes for engineered nanomaterials. Part Fibre Toxicol 13 (1), 25. https://doi.org/10.1186/s12989-016-0137-5.
[50] Nobel, S.M.N., Sifat, O.F., Islam, M.R., Sayeed, M.S., Amiruzzaman, M., 2024. Enhancing GI cancer radiation therapy: advanced organ segmentation with ResECA-U-net model. Emerg Sci J 8 (3), 999–1015. https://doi.org/10.28991/ESJ-2024-08-03-012.
[51] OECD, 2023. Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions. In: OECD Series on Testing and Assessment, 386. OECD Publishing, Paris. https://doi.org/10.1787/d96118f6-en.
[52] Pandey, A., Mishra, A.K., 2022. Immunomodulation, toxicity, and therapeutic potential of nanoparticles. BioTech 11 (3), 42. https://doi.org/10.3390/biotech11030042.
[53] Peart, M.J., Smyth, G.K., van Laar, R.K., Bowtell, D.D., Richon, V.M., Marks, P.A., et al., 2005. Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. Proc Natl Acad Sci USA 102 (10), 3697–3702. https://doi.org/10.1073/pnas.0500369102.
[54] Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., et al., 2011. Scikit-learn: machine learning in python. J Mach Learn Res 12, 2825–2830 (http://scikit-learn.sourceforge.net).
[55] Pedrotty, D.M., Morley, M.P., Cappola, T.P., 2012. Transcriptomic biomarkers of cardiovascular disease. Prog Cardiovasc Dis 55 (1), 64–69. https://doi.org/10.1016/j.pcad.2012.06.003.
[56] Rahman, L., Wu, D., Johnston, M., Williams, A., Halappanavar, S., 2017. Toxicogenomics analysis of mouse lung responses following exposure to titanium dioxide nanomaterials reveal their disease potential at high doses. Mutagenesis 32 (1), 59–76. https://doi.org/10.1093/mutage/gew048.
[57] Raouf, A., Zhao, Y., To, K., Stingl, J., Delaney, A., Barbara, M., et al., 2008. Transcriptome analysis of the normal human mammary cell commitment and differentiation process. Cell stem Cell 3 (1), 109–118. https://doi.org/10.1016/j.stem.2008.05.018.
[58] Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst 145, 22–29. https://doi.org/10.1016/j.chemolab.2015.04.013.
[59] Saber, A.T., Jacobsen, N.R., Jackson, P., Poulsen, S.S., Kyjovska, Z.O., Halappanavar, S., et al., 2014. Particle-induced pulmonary acute phase response may be the causal link between particle inhalation and cardiovascular disease. WIREs Nanomed Nanobiotechnol. 6 (6), 517–531. https://doi.org/10.1002/wnan.1279.
[60] Saber, A., Jacobsen, N., Mortensen, A., Szarek, J., Jackson, P., Madsen, A., et al., 2012. Nanotitanium dioxide toxicity in mouse lung is reduced in sanding dust from paint. Part Fibre Toxicol 9 (1), 4. https://doi.org/10.1186/1743-8977-9-4.
[61] Saber, A., Mortensen, A., Szarek, J., Jacobsen, N., Levin, M., Koponen, I., et al., 2019. Toxicity of pristine and paint-embedded TiO 2 nanomaterials. Hum Exp Toxicol 38 (1), 11–24. https://doi.org/10.1177/0960327118774910.
[62] Santis, A., Arbeláez, O., Cardenas, L.A., Castellanos, J., Velasquez, P., 2024. Optimizing Cr(VI) reduction in plastic chromium plating wastewater: particle size, irradiation, titanium dose. Emerg Sci J 8 (1), 17–27. https://doi.org/10.28991/ESJ-2024-08-01-02.
[63] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davis, R.W., 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci 93 (20), 10614–10619. https://doi.org/10.1073/pnas.93.20.10614.
[64] Schmid, O., Stoeger, T., 2016. Surface area is the biologically most effective dose metric for acute nanoparticle toxicity in the lung. J Aerosol Sci 99, 133–143. https://doi.org/10.1016/j.jaerosci.2015.12.006.
[65] Sen'kova, A.V., Savin, I.A., Brenner, E.V., Zenkova, M.A., Markov, A.V., 2021. Core genes involved in the regulation of acute lung injury and their association with COVID-19 and tumor progression: A bioinformatics and experimental study. PLoS One 16 (11), e0260450. https://doi.org/10.1371/journal.pone.0260450.
[66] Serra, A., Fratello, M., Cattelani, L., Liampa, I., Melagraki, G., Kohonen, P., et al., 2020. Transcriptomics in toxicogenomics, part III: data modelling for risk assessment. Nanomaterials 10 (4), 708. https://doi.org/10.3390/nano10040708.
[67] Shi, H., Magaye, R., Castranova, V., Zhao, J., 2013. Titanium dioxide nanoparticles: a review of current toxicological data. Part Fibre Toxicol 10 (1), 15. https://doi.org/10.1186/1743-8977-10-15.
[68] Shvedova, A.A., Yanamala, N., Kisin, E.R., Tkach, A.V., Murray, A.R., Hubbs, A., et al., 2014. Long-term effects of carbon containing engineered nanomaterials and asbestos in the lung: one year postexposure comparisons. Am J Physiol-Lung Cell Mol Physiol 306 (2), L170–L182. https://doi.org/10.1152/ajplung.00167.2013.
[69] Supplitt, S., Karpinski, P., Sasiadek, M., Laczmanska, I., 2021. Current achievements and applications of transcriptomics in personalized cancer medicine. Int J Mol Sci 22 (3), 1422. https://doi.org/10.3390/ijms22031422.
[70] Tilton, S.C., Karin, N.J., Tolic, A., Xie, Y., Lai, X., Hamilton, R.F., et al., 2014. Three human cell types respond to multi-walled carbon nanotubes and titanium dioxide nanobelts with cell-specific transcriptomic and proteomic expression patterns. Nanotoxicology 8 (5), 533–548. https://doi.org/10.3109/17435390.2013.803624.
[71] Valentino, S.A., Seidel, C., Lorcin, M., Sébillaud, S., Wolff, H., Grossmann, S., et al., 2023. Identification of a gene signature predicting (nano)particle-induced adverse lung outcome in rats. Int J Mol Sci 24 (13), 10890. https://doi.org/10.3390/ijms241310890.
[72] Wallin, H., Kyjovska, Z.O., Poulsen, S.S., Jacobsen, N.R., Saber, A.T., Bengtson, S., et al., 2017. Surface modification does not influence the genotoxic and inflammatory effects of TiO 2 nanoparticles after pulmonary exposure by instillation in mice. Mutagenesis 32 (1), 47–57. https://doi.org/10.1093/mutage/gew046.
[73] Wang, M., Li, J., Dong, S., Cai, X., Simaiti, A., Yang, X., et al., 2020. Silica nanoparticles induce lung inflammation in mice via ROS/PARP/TRPM2 signaling-mediated lysosome impairment and autophagy dysfunction. Part Fibre Toxicol 17 (1), 23. https://doi.org/10.1186/s12989-020-00353-3.
[74] Xu, H., Lam, S.H., Shen, Y., Gong, Z., 2013. Genome-wide identification of molecular pathways and biomarkers in response to arsenic exposure in zebrafish liver. PLoS ONE 8 (7), e68737. https://doi.org/10.1371/journal.pone.0068737.
[75] Zhang, R., Chen, S., Chen, L., Ye, L., Jiang, Y., Peng, H., et al., 2022. Single-cell transcriptomics reveals immune dysregulation mediated by IL-17A in initiation of chronic lung injuries upon real-ambient particulate matter exposure. Part Fibre Toxicol 19 (1), 42. https://doi.org/10.1186/s12989-022-00483-w.
[76] Zhao, Y., Guo, D.-F., Rahmouni, K., 2024. Proteomic profiling points to the significance of adipocyte-derived extracellular vesicles in diet-induced obesity. Physiology 39 (S1). https://doi.org/10.1152/physiol.2024.39.S1.881.
[77] Zschaler, J., Schlorke, D., Arnhold, J., 2014. Differences in innate immune response between man and mouse. Crit Rev Immunol 34 (5), 433–454.