Evaluation of the Information Content in Proposed QSAR Descriptors via Machine Learning Meta-Analysis of *In Vivo* Nanotoxicity Experiments

Jeremy M. Gernand | Penn State University, University Park, PA
Elizabeth A. Casman | Carnegie Mellon University, Pittsburgh, PA
Vignesh Ramchandran | Penn State University, University Park, PA

PENNSTATE

Carnegie Mellon

CENT | Center for the Environmental Implications of NanoTechnology

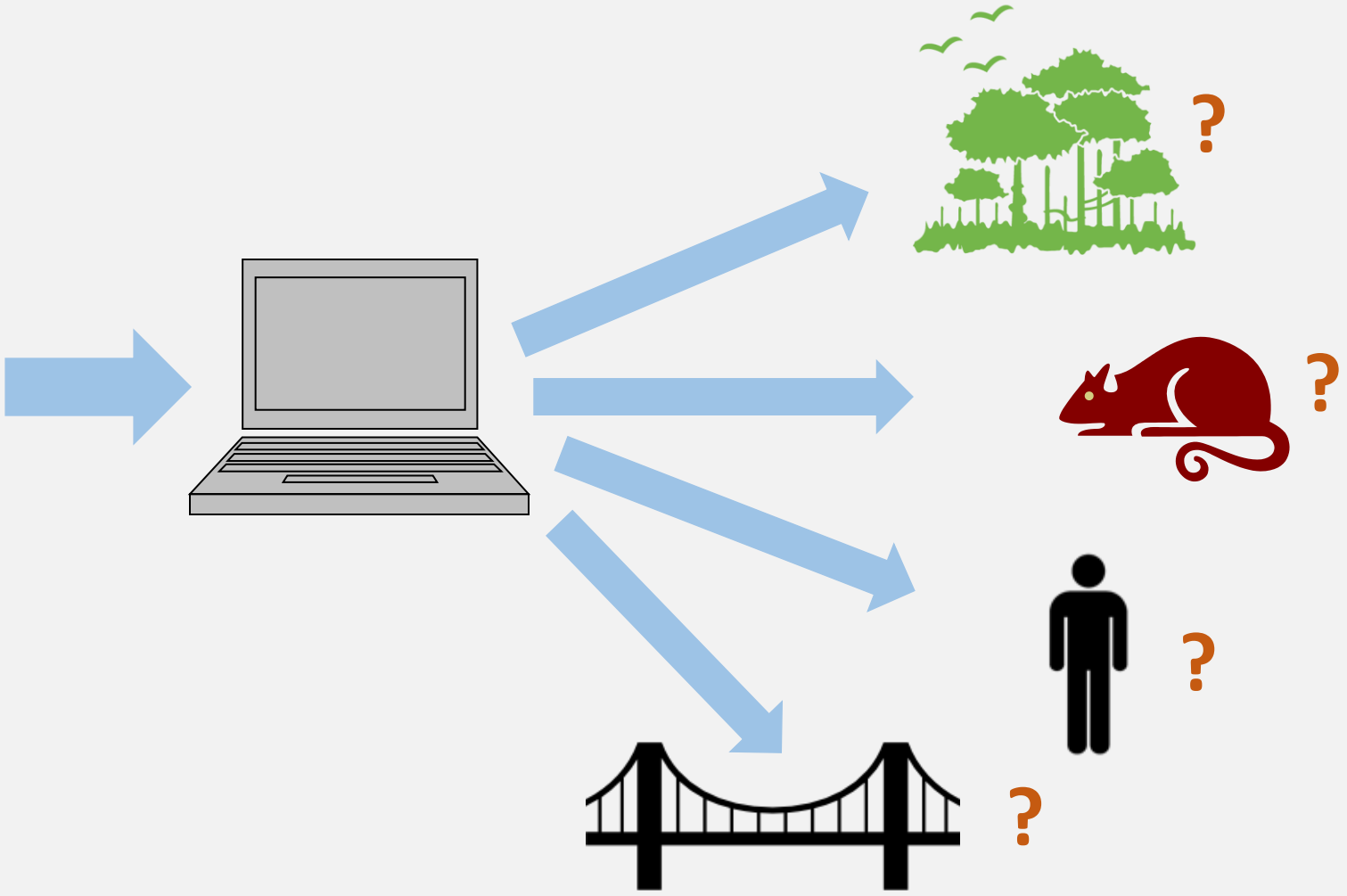
What could we do with models that predict the kinds of interactions nanomaterials and biological organisms have?

- Develop safer technological utilization of nanotechnology (reduce risks)
  - Protect workers and consumers
  - Protect patients
  - Protect the environment from new pollutants
- Identify more useful and effective nanomaterials (improve function)
  - Better materials
  - Better drugs
- Enable design tradeoffs between risk and function

We want to connect potential ***risks*** of and ***usefulness*** of nanomaterials to specific particle characteristics

Chemical makeup
Purity
Size
Shape
Surface properties
Surface area
Aggregation state
…

Concentration
# of particles
Duration
Recovery
…



Based primarily on *in vivo* data sets a few nanomaterial QSARs for toxicity have been proposed

Author(s)	Year	Proposed Predictors
Puzyn T. et al.	2011	Enthalpy of formation of a gaseous cation: $Me\ (s) \rightarrow Me^{n+}(g) + n \cdot \overline{e}$ ΔH_{Me+}
Fourches D. et al.	2011	Surface area, atom and bond counts, Kier & Hall connectivity indices, kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and molecular charges
Liu R. et al.	2011	NM and NO: number of metal and Oxygen atoms, mMe (g·mol−1): atomic mass of the nanoparticle metal, mMeO (g·mol−1): molecular weight of the metal oxide, GMe and PMe: group and period of the nanoparticle metal, EMeO (kcal·eqv−1): atomization energy of the metal oxide, d (nm): nanoparticle primary size, Zw (mV): zeta potential (in water at pH=7.4), IEP: isoelectric point.

Data sources for this investigation made up of 162 pulmonary nanomaterial exposure studies in rodents

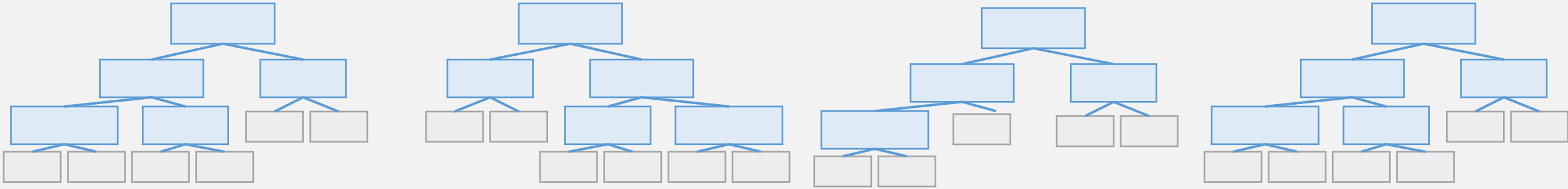
Table 1: Number of unique nanomaterial variants by particle type included in each dataset

Nanomaterial Type	Pulmonary Dataset	*In Vitro* Dataset
Ag	10	8
Al_2O_3	4	0
Au	5	1
C_{60}	4	0
CdO	3	0
CeO_2	20	5
CNT	24	8
Co_3O_4	5	0
Cu	5	9
Fe	3	0
Fe_2O_3	4	0
Fe_3O_4	5	0
MgO	2	0
Ni	6	0
NiO	11	0
SiO_2	35	6
TiO_2	44	13
W	1	0
ZnO	12	3

- Although dominated by titania, silica, CNT, and ceria studies, there is a substantial amount of data existing in published sources on pulmonary exposures to nanomaterials
- 162 separate studies
- 2136 unique exposure groups
- Focused primarily on inflammation and other short term impacts

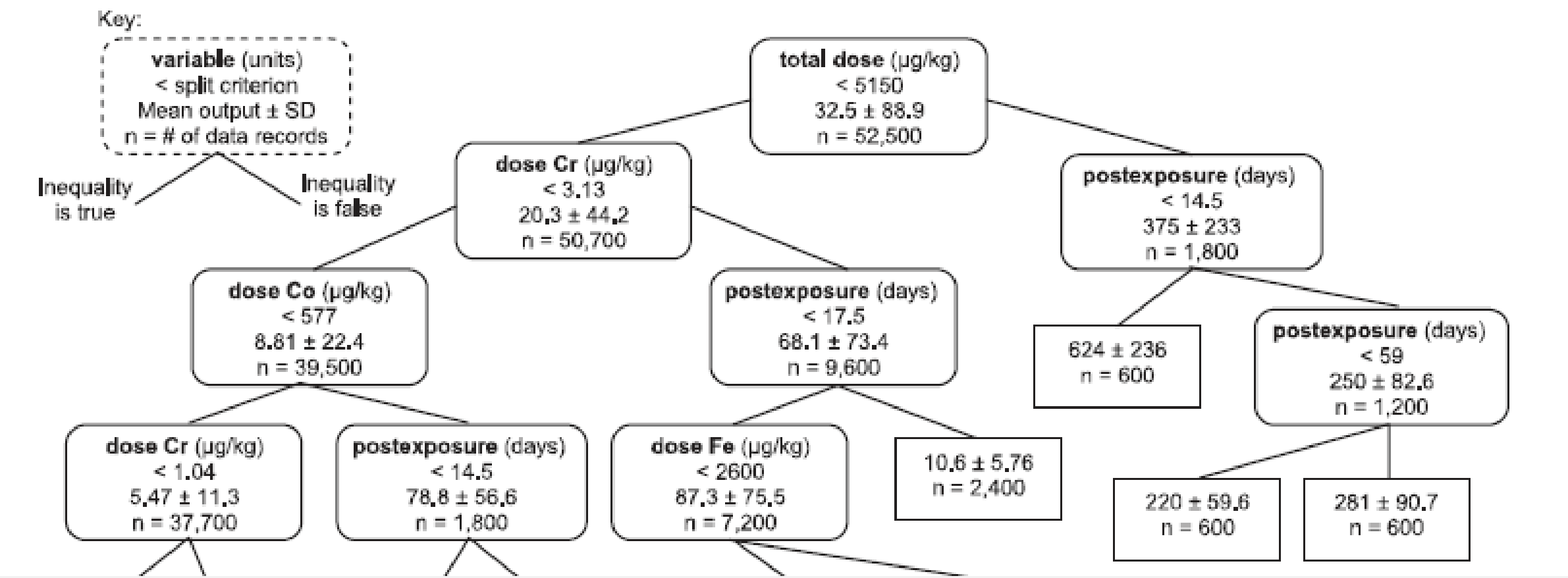
Regression Tree and Random Forest models can help measure information content in input parameters

- These models can be used with missing data without requiring imputation
  - A very important characteristic when incorporating data from many different *in vivo* studies
- The nonlinear nature of the model structure can identify a likely upper limit to the predictive utility of each input variable
  - Careful validation necessary to prevent identification of noise as important
- Regression trees are easily readable unlike other machine learning models



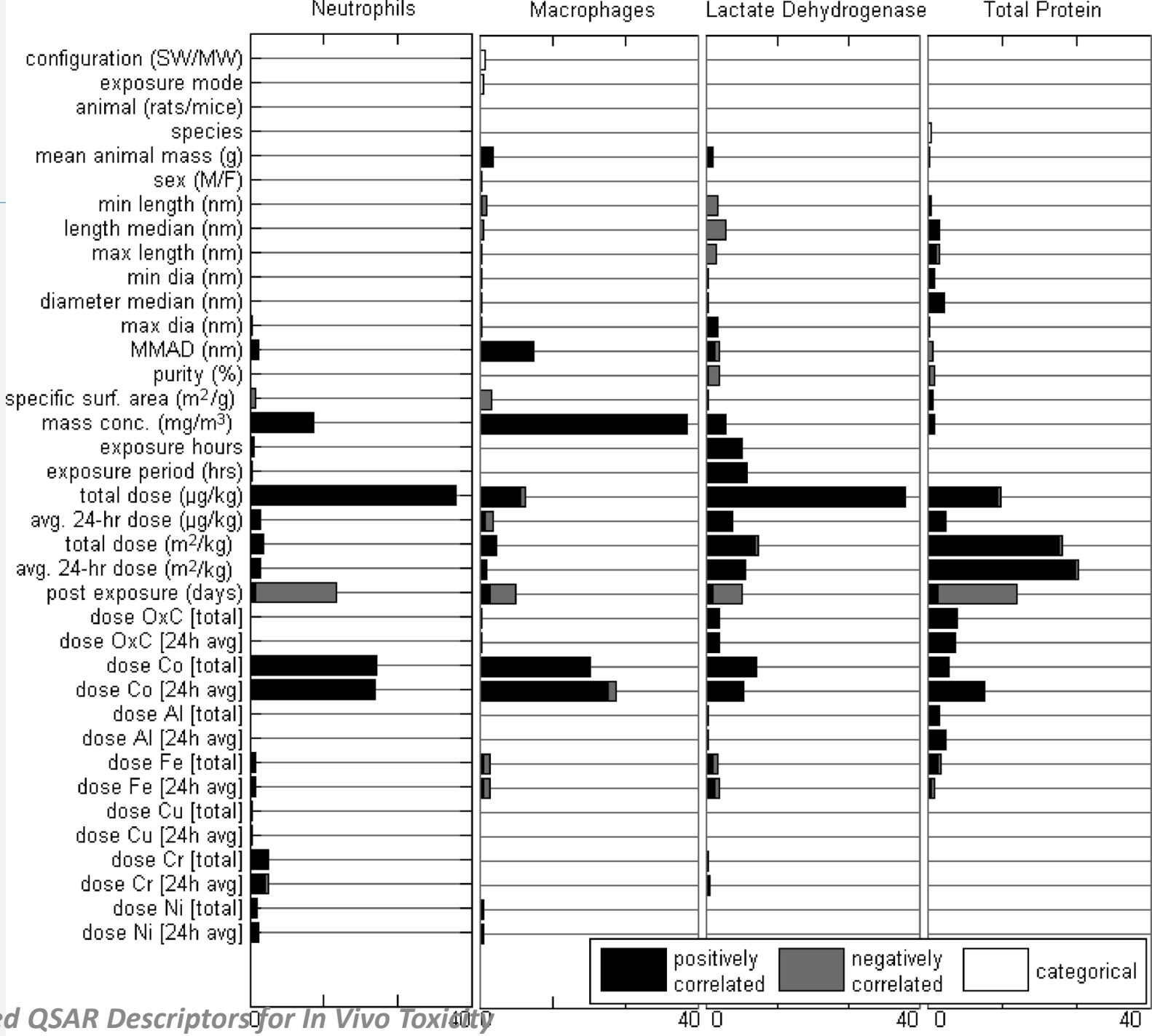
Information gain by the addition of each branch is recorded along with correlation and conditionality

- Measuring the error or variance reduction achieved by each individual branch is a simple expression of variable value to model



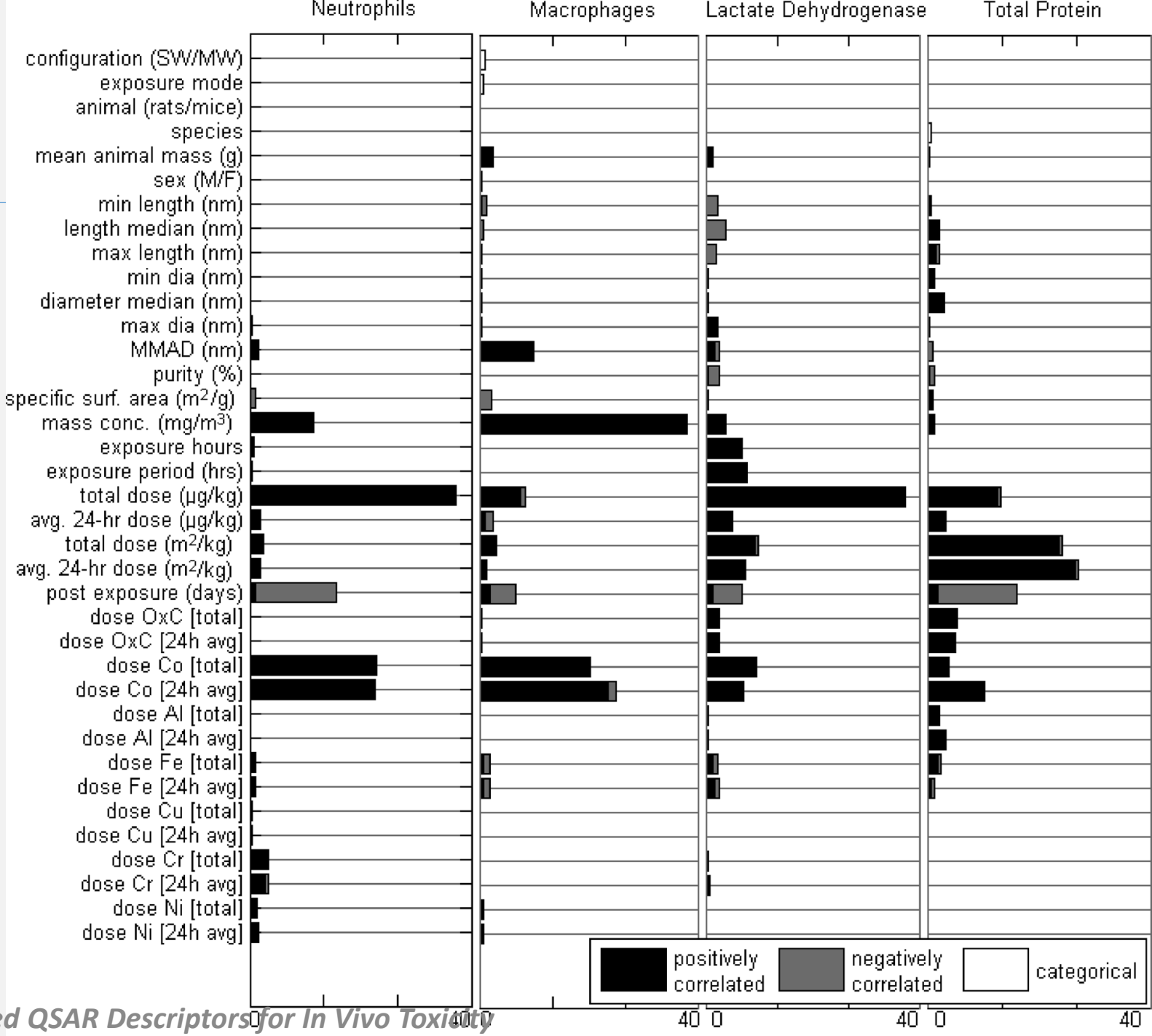
Information content of CNT tox predictors

- Assembling the variance reduction values per variable for many different toxic endpoints provides a picture of information value consistence across different endpoint measures



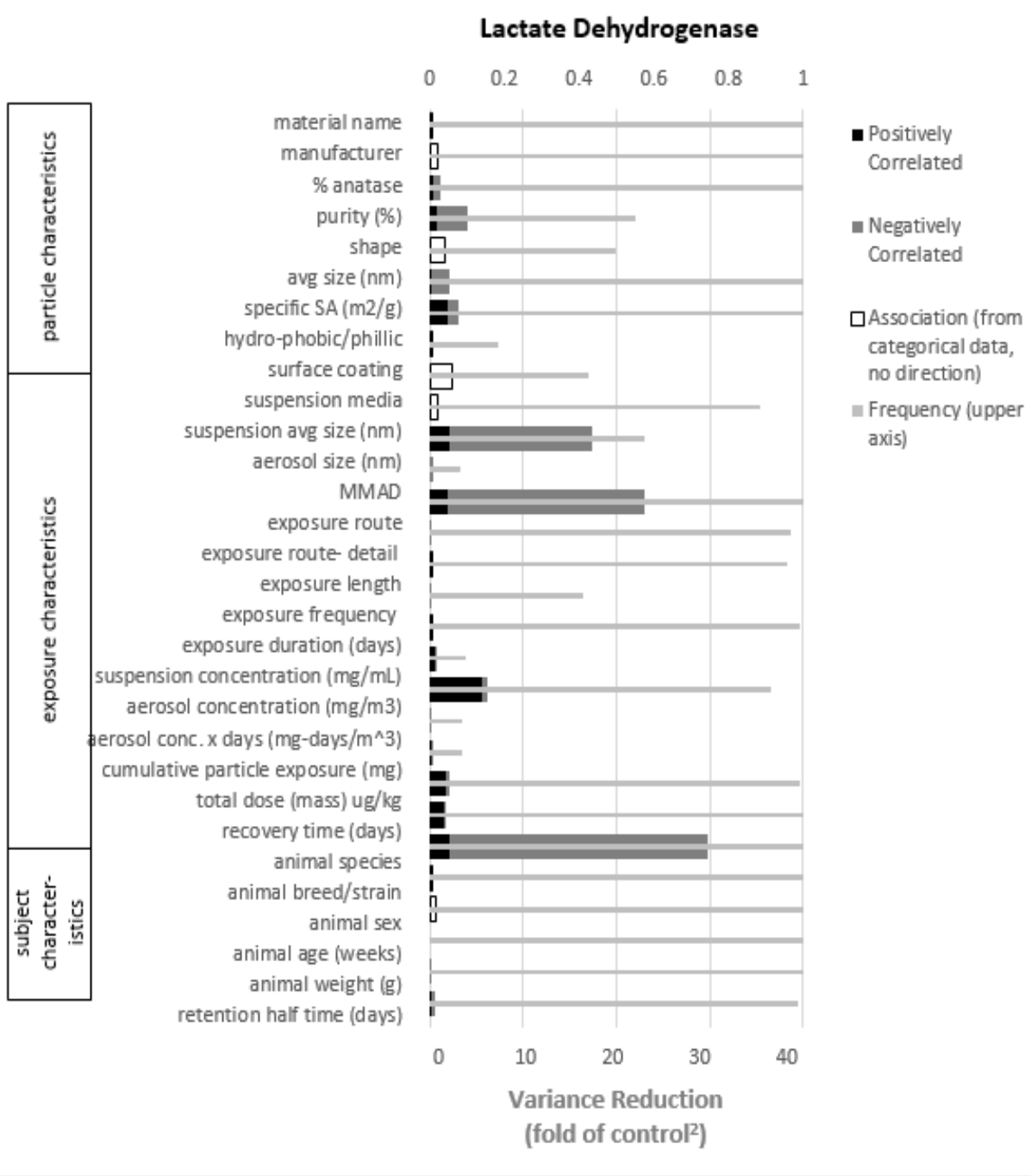
Information content of CNT tox predictors

- In CNT studies some QSAR-like descriptors were identified as important predictors of toxicity
  - Length and Diameter
  - Aggregation
  - Metal impurity content (Co, Fe, Cr, Ni)



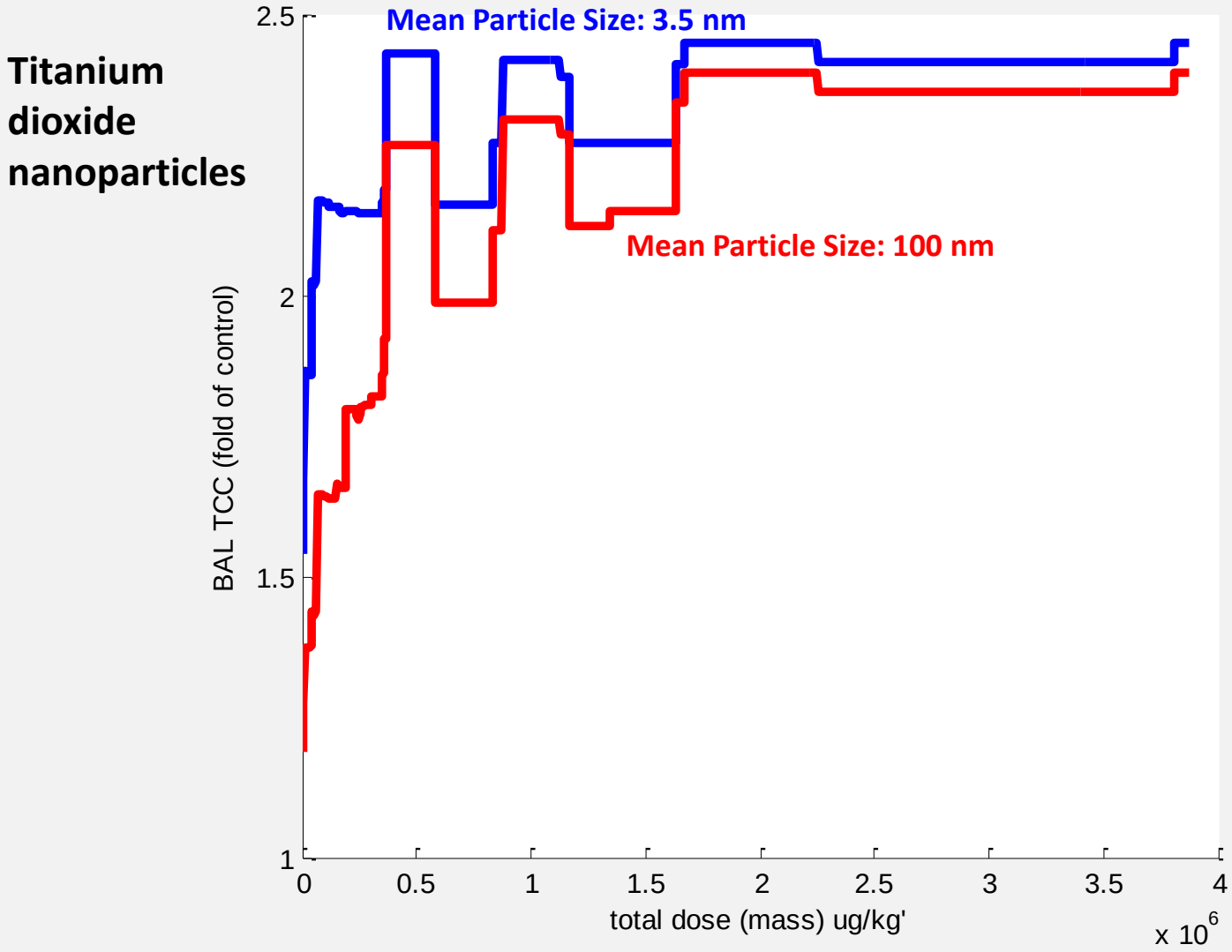
Considering titania studies against one another

- Within TiO_2 studies, crystalline structure seems relatively unimportant compared to dose metrics, aggregation, and recovery time
- Particle size and purity had consistent though relatively small effects



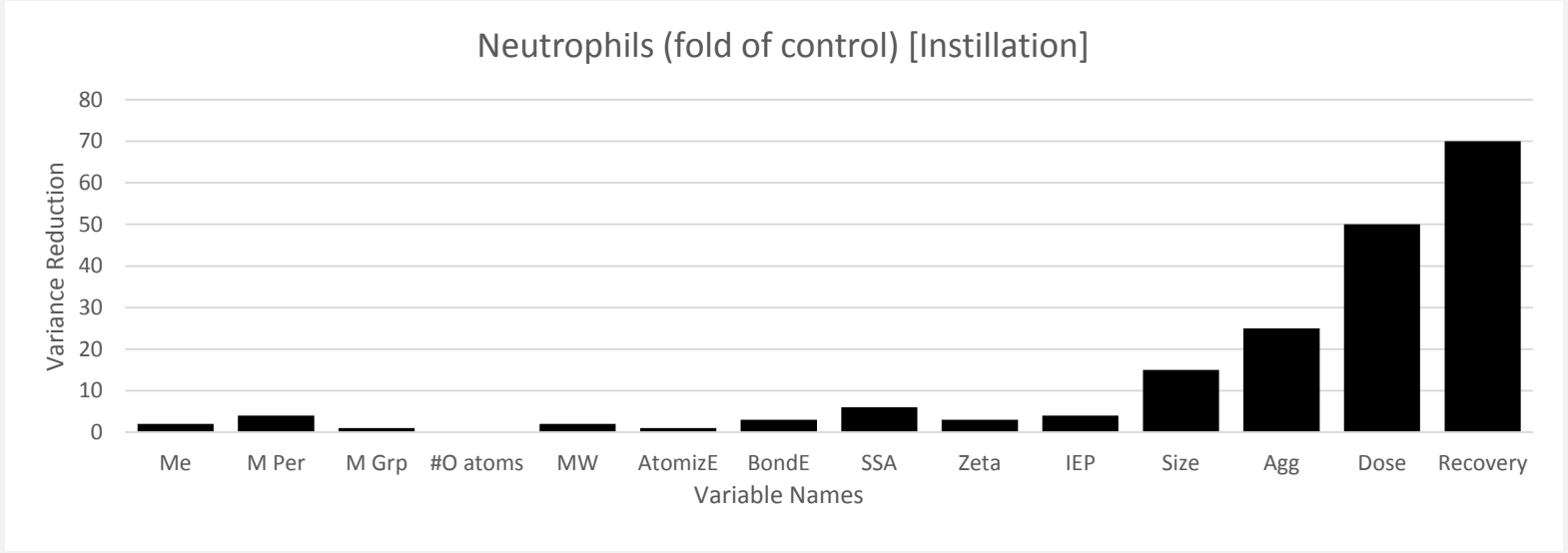
Random Forest models do appear to find known relationships and identify the relative importance of different properties

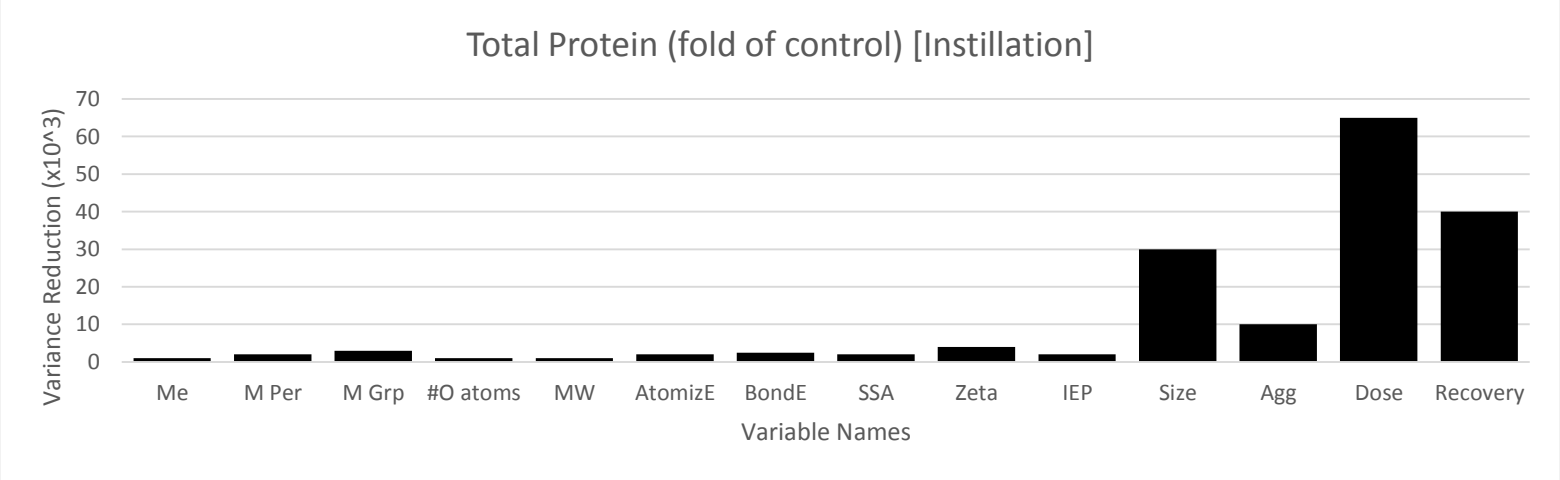
- Although Random Forest models are "dumb"—ignorant of any underlying data structure, they often uncover plausible looking dose-response relationships assembled out of step functions



What is the value to QSAR descriptors for metal oxides when considered as a class

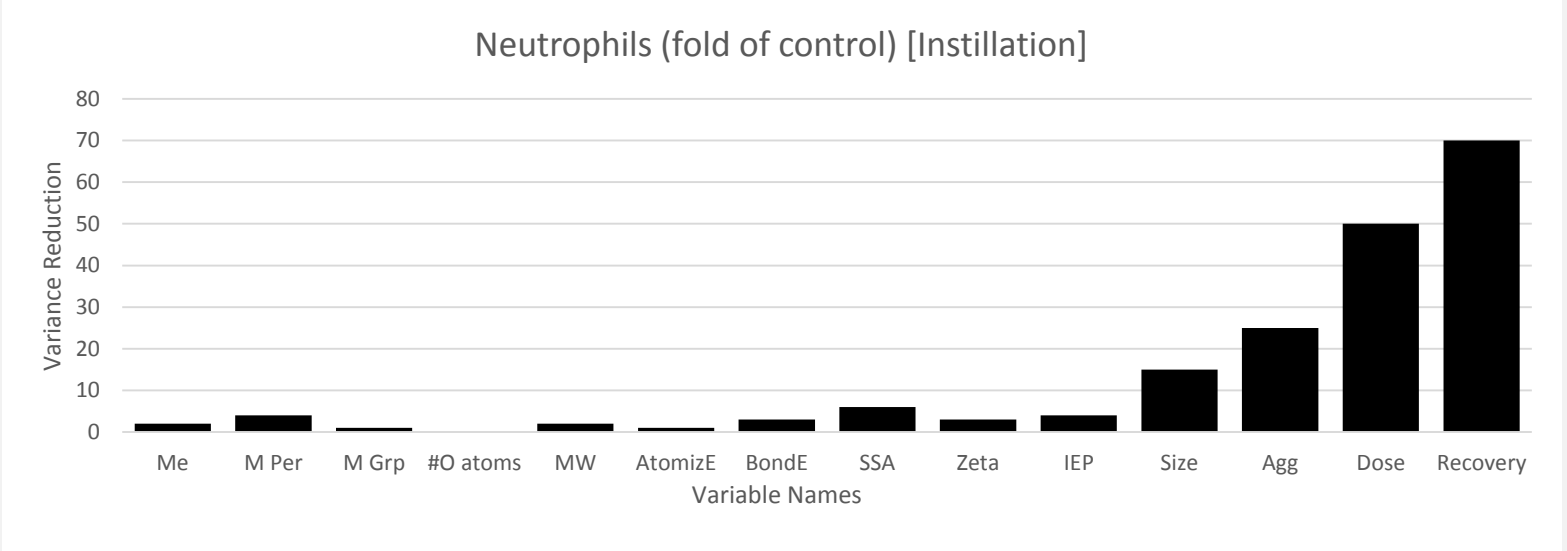
- At first glance, many of the chemical descriptors of metal oxide nanoparticles do not seem to help the model predict pulmonary toxicity in rodents
- Their true value could be conditional on another variable not yet in the model (e.g. biological or environmental prevalence)

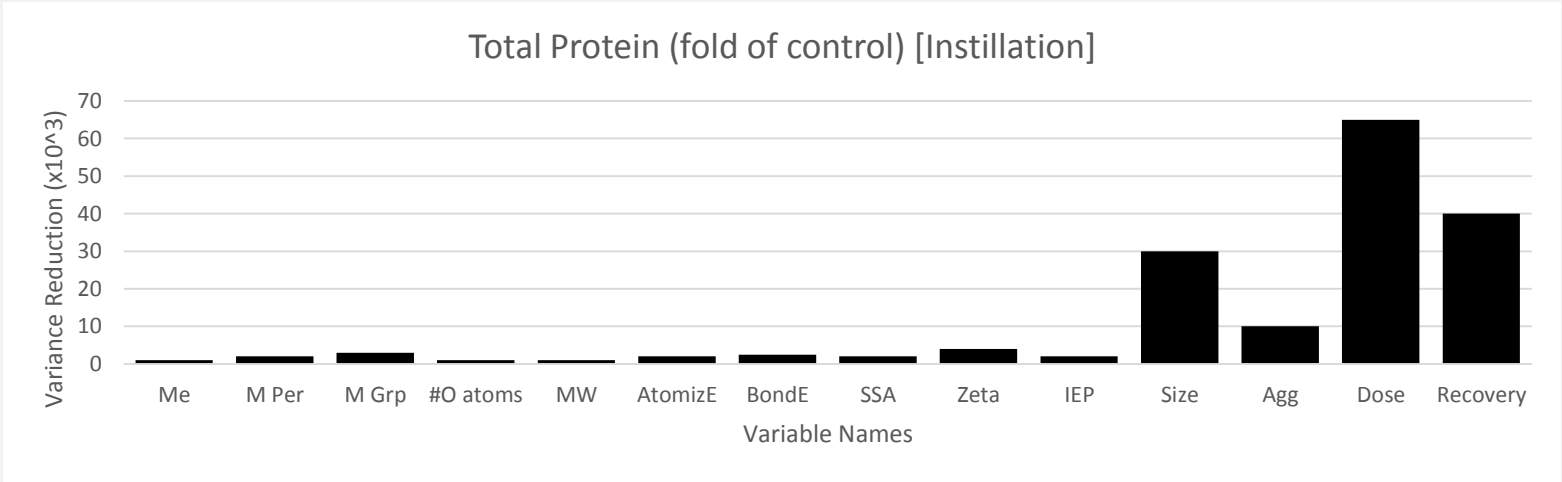




What is the value to QSAR descriptors for metal oxides when considered as a class

- It seems unlikely that none of these chemical properties are important in some way
- Combinations of descriptors need to be tested
- But, perhaps we would benefit from a new method of measuring importance





Development of a new algorithm to better reflect the expectation of dose-response shape

- Seems odd to consider dose or animal recovery time as fundamentally similar concepts to a nanoparticle property in the data mining exercise
- Requires a modified regression tree algorithm designed not to predict a constant value in the leaf nodes, but a function that incorporates our knowledge of the shape dose-response relationships

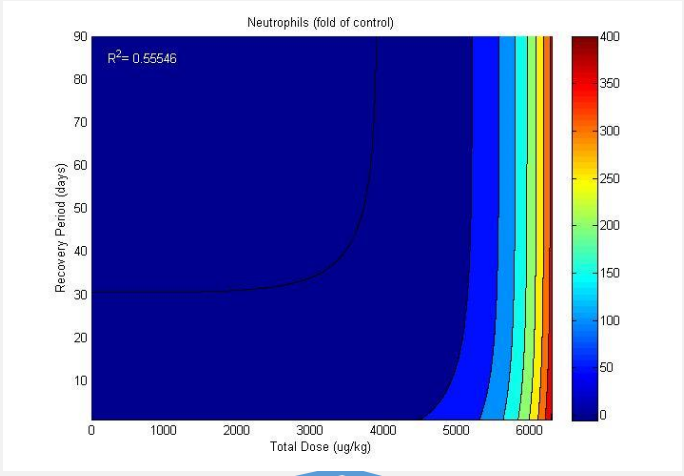
$$Outcome = A + Ce^{-Bx} - Fe^{-Dt}$$

Where,

x is the dose or exposure metric

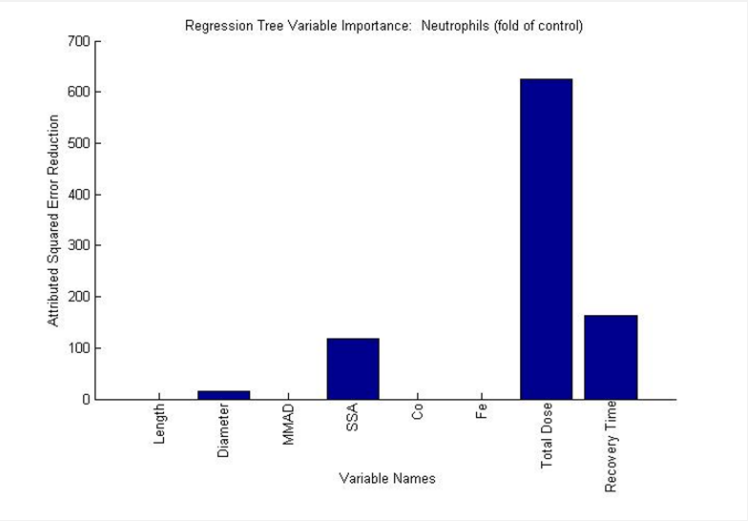
t is the recovery period

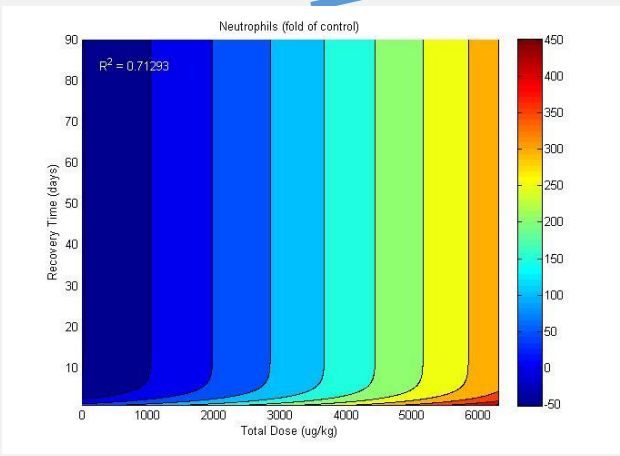
The model contour surfaces show how dose-response and recovery shift with changes in particle properties

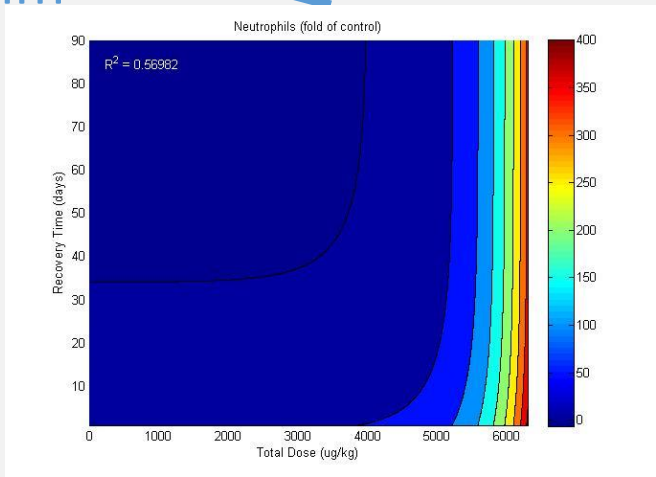


TRUE

Diameter < 5 nm

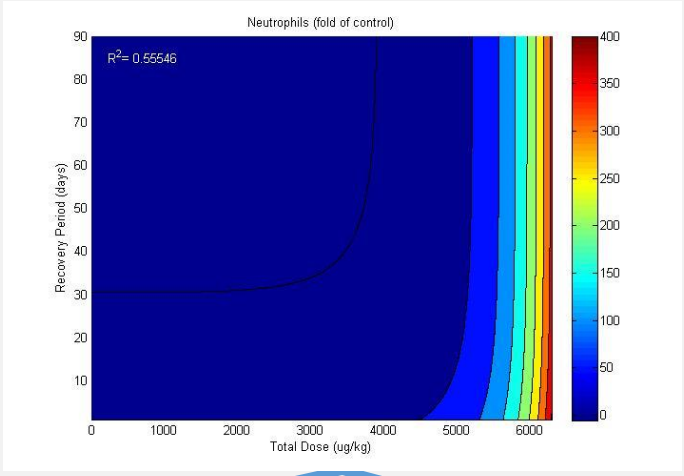






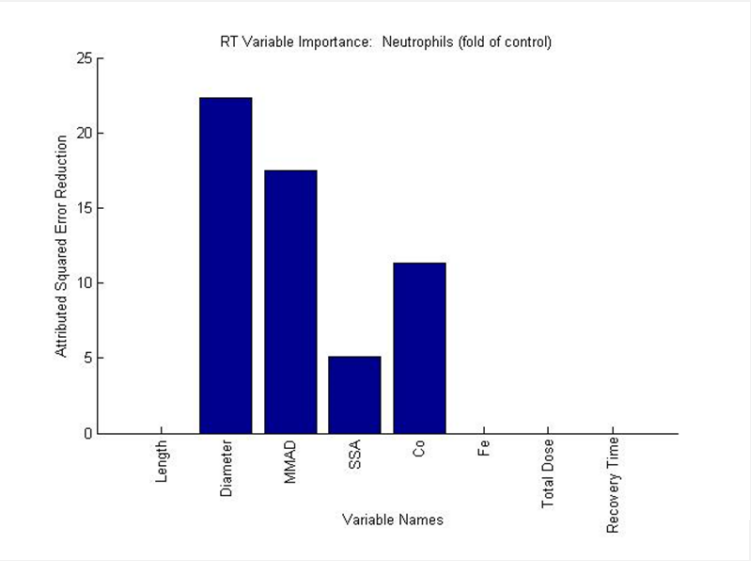
Variable importance from traditional regression tree model

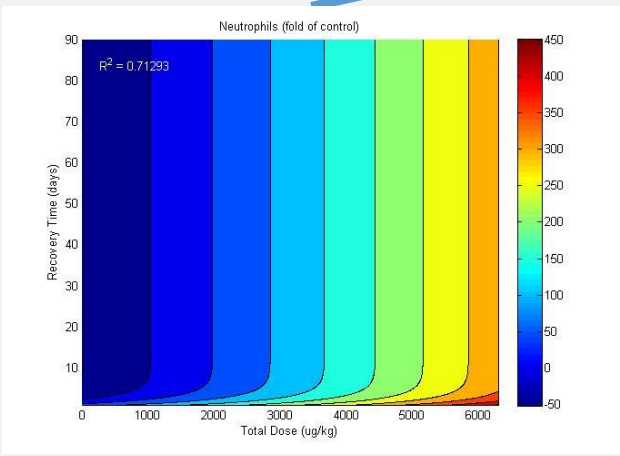
Now particle properties can be analyzed for their effects on dose-response rather than considered alongside dose

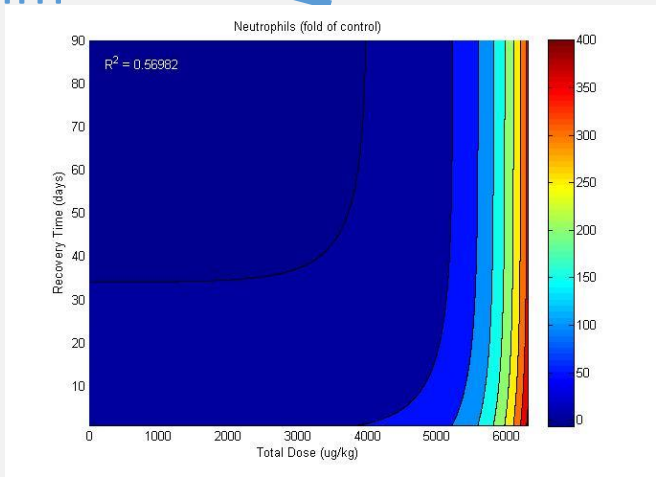


TRUE

Diameter < 5 nm







Variable importance from new 2-D exponential regression tree model

This approach shows promise for better quantifying knowledge in the field

- The large number of independent studies in nanotoxicology should be incorporated into QSAR modeling and evaluation as much as possible
- This process is one way of doing that and ensuring that we do not ignore lingering sources of uncertainty in our knowledge base

- In the future…
  - Complete testing of possible descriptor parameters including those that are valid beyond the list of metal oxides
  - Test and validate the QSAR descriptors in the new treed exponential regression tree model for information content
  - Expand data set to environmentally relevant exposure studies in other organisms and investigate the effect of particle properties and QSAR descriptors

Acknowledgements

Vignesh Ramchandran (Penn State)

Elizabeth Casman (Carnegie Mellon)

Jacob Borst (Penn State)

Steve Edinger (Penn State)

This work has been supported by:

National Science Foundation (NSF) and the Environmental Protection Agency (EPA) under NSF Cooperative Agreement EF-0830093, Center for the Environmental Implications of NanoTechnology (CEINT)