QSAR concepts for nanomaterials

Rationale, methods, pitfalls

Dave Winkler | Modelling Team Leader, Adjunct Professor, Monash University, Melbourne

COST MODENA Summer School, August 2013

CMSE CLAYTON
www.csiro.au

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

MONASH University

CSIRO

CSIRO today: a snapshot

- Australia's national science agency
- One of the largest & most diverse in the world
- 6500+ staff over 55 locations
- Ranked in top 1% in 14 research fields
- 20+ spin-off companies in six years
- 160+ active licences of CSIRO innovation
- Building national prosperity and wellbeing

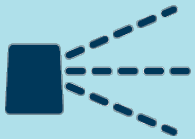
CSIRO's top 10 successes

1. WLAN
Wireless Local Area Network

2. POLYMER BANKNOTES

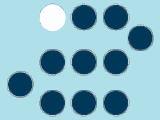
3. RELENZA FLU TREATMENT

4. EXTENDED WEAR CONTACTS



5. AEROGARD

6. TOTAL WELLBEING DIET



7. RAFT POLYMERISATION

8. BARLEYMAX

9. SELF TWISTING YARN

10. GMO COTTON

Nanosafety Research

The program can be split into four main areas of research:

- Nanoparticle detection, characterisation and quantification of nano-objects in aerosols.
- Health effects on human and other mammalian systems upon exposure to nanoparticles in the workplace and upon exposure to nanoparticles in products.
- Information on the fate and transport of nanomaterials released to the natural environment, and the effects of these on ecosystems.
- Computational modelling of nanoparticle toxicity.

A practical human *in vivo* study

Volunteers at North Curl Curl beach, Sydney, March 2009

B. Gulson, M. McCall, M. Korsch, L. Gomez, P. Casey, Y.Oytam, A. Taylor, L. Kinsley & G. Greenoak (2010) *Toxicological Sciences* (in press). “Small amounts of zinc from zinc oxide particles in sunscreens applied outdoors are absorbed through the skin.”

Do Zinc Oxide Nanoparticles from Sunscreen Penetrate Hairless Mouse Skin?

ZnO particles >100nm
"bulk"

ZnO particles ~20nm
"nano"

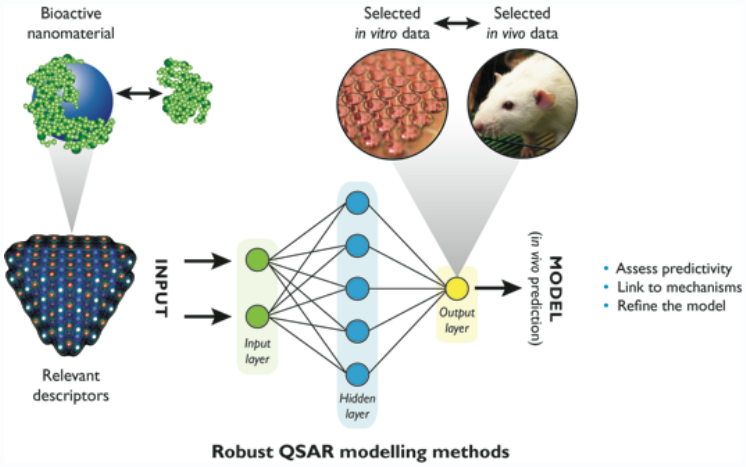
Predicting biological effects of nanomaterials

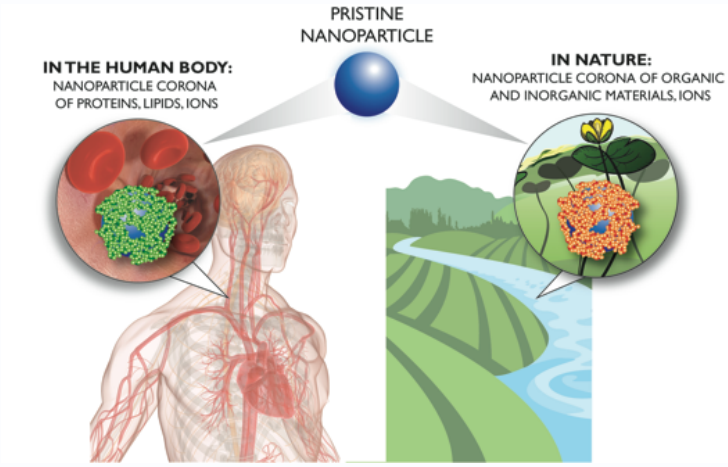
Opportunity: The novel properties of nanomaterials have seen rapid incorporation into products (50,000 product types by 2015). However, their adverse biological effects in humans and the environment are not known

Technology: Computational modelling of nanomaterials properties allows prediction of biological effects. Proprietary feature selection and machine learning methods are robust and widely applicable

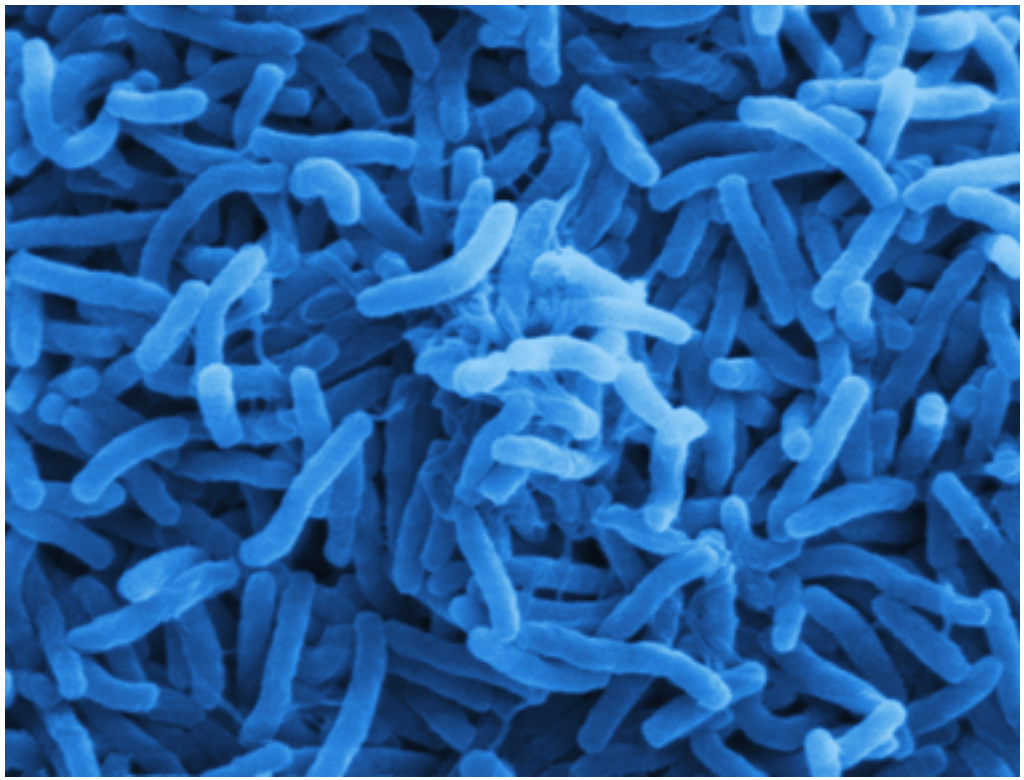
Advantages: Computational models are very fast and can make predictions on materials not yet used or even synthesized. Key publications in high impact journals

Value: Promises tools to help governments regulate nanomaterials safely without stifling commerce

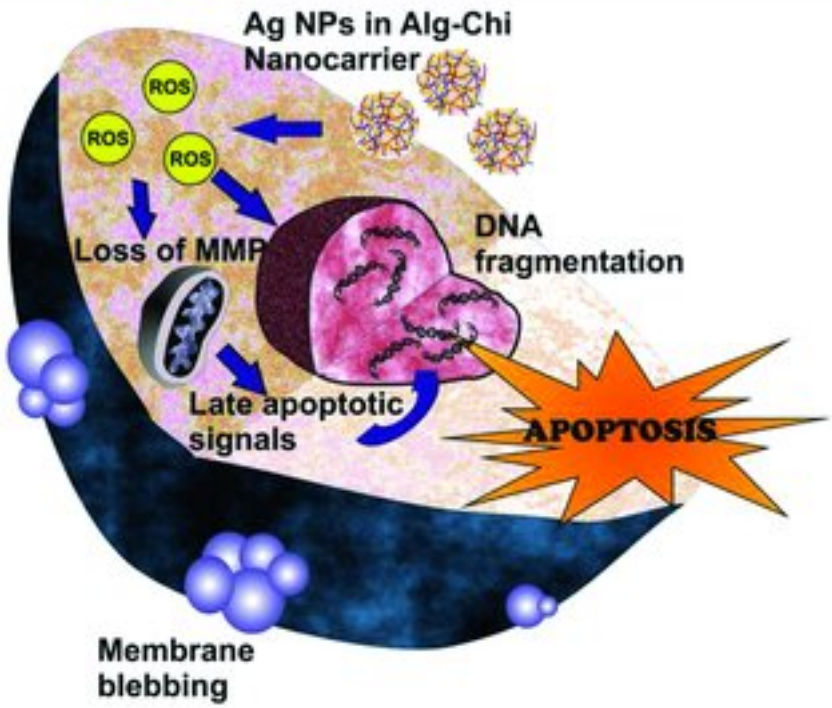




Which members of this polymer library will not allow bacteria to adhere & grow?

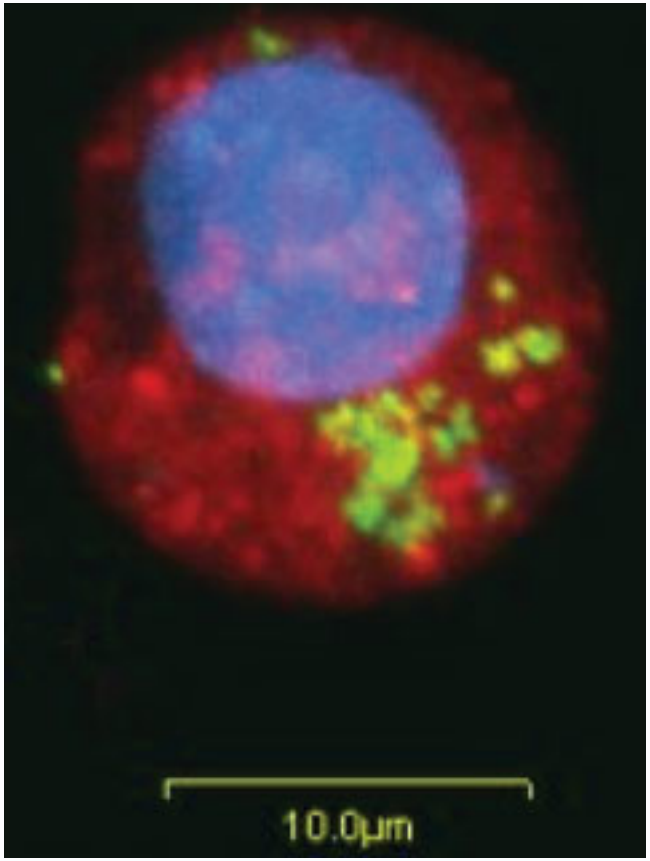


Will this nanoparticle cause damage to cells?

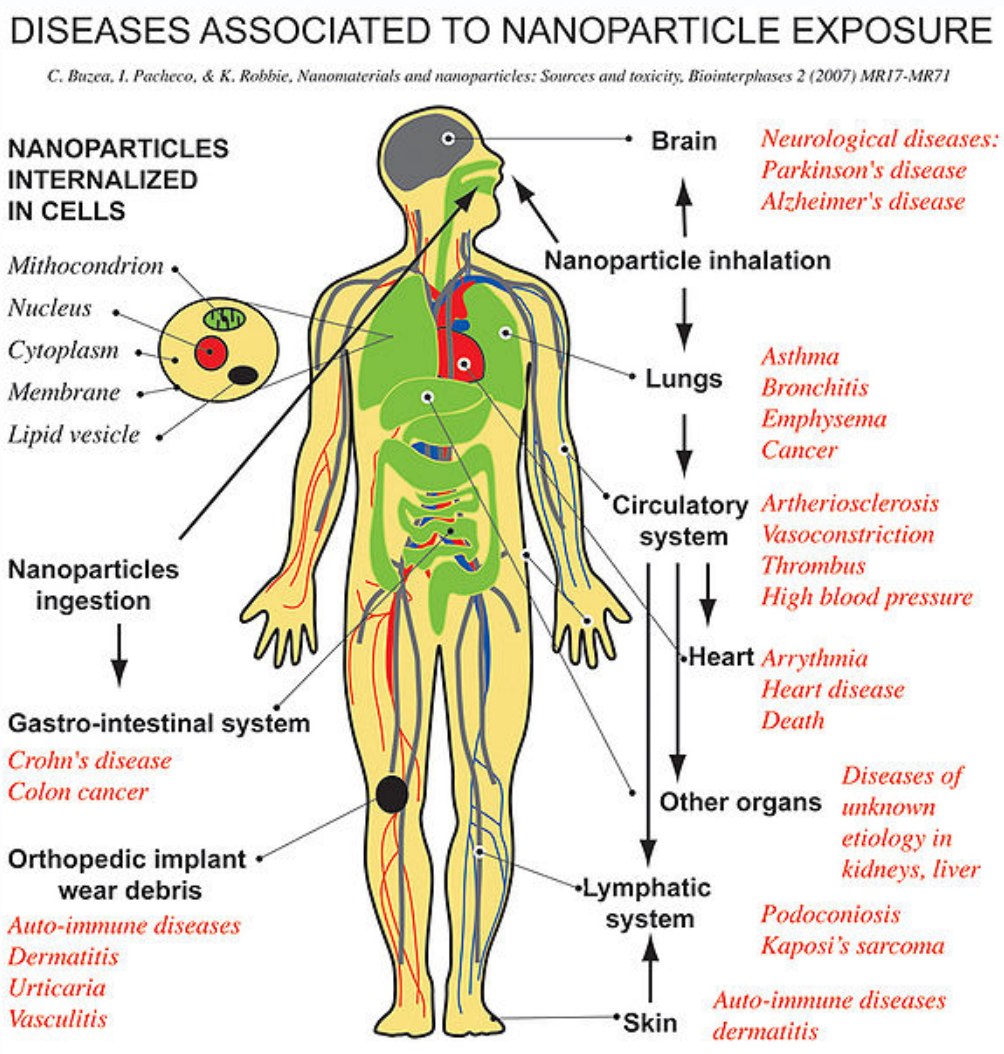


almutairi.ucsd.edu

Will this nanoparticle be taken up by macrophages?



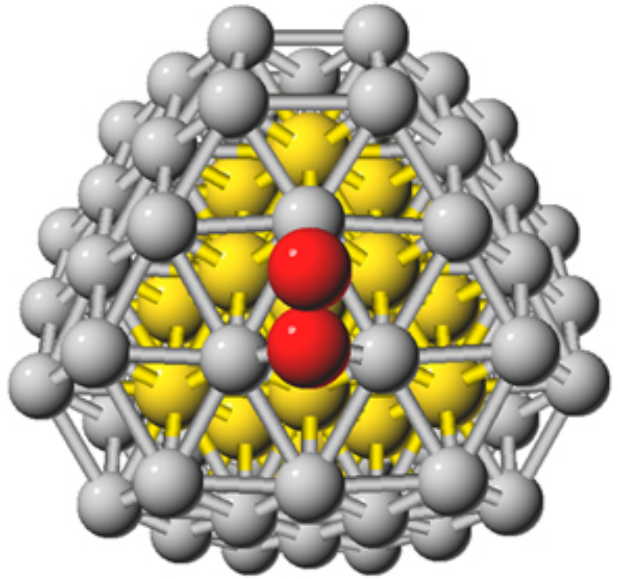
Diseases associated with nanoparticles



Wikipedia commons

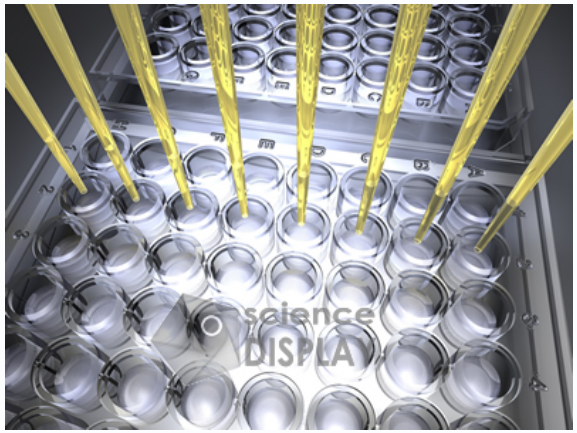
Why computational modelling?

- Experimental testing of nanomaterials for physicochemical, toxicological and environmental properties is time-consuming and expensive ~50,000 nano products by 2015
- Increasing pressure to reduce or discontinue animal testing.
- Computational methods like QSAR are becoming increasingly useful and reliable. Regulators use them for industrial chemicals
- Such tools will help regulators make decisions about the risk nano-materials may pose
- Computational modelling will complement, not replace the need for experimental assessment of the biological effects of nanoparticles



Need for predictive methods

- High throughput synthesis and testing methods used to discover many new materials must also be applied to testing adverse properties of nanomaterials
- Non-testing alternatives like statistical and machine learning modelling can play a a big role in materials discovery, optimization, and nanosafety. Such methods have been used successfully in other fields. They have been effective in drug discovery and can be applied to predicting biological effects of nanomaterials
- The size of materials space is so large that we cannot explore it all. Interactions of nanomaterials with biology are multifactorial, complex, and largely unknown



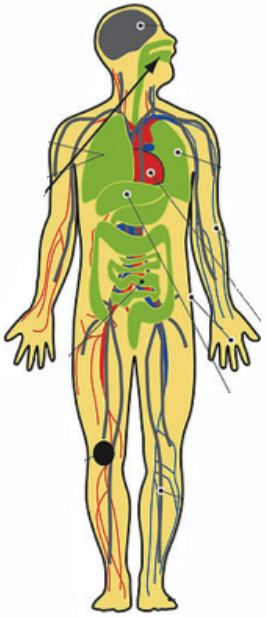
Problems we face – can we do anything?

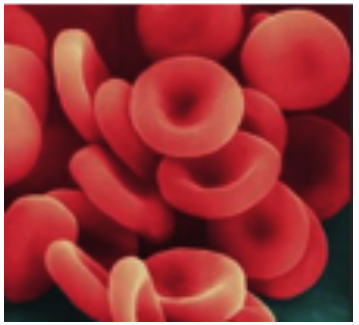
Nanoparticle: intrinsic physical and molecular properties

?? Complex, poorly understood processes: ingestion, uptake, interactions with proteins, transport, cell processes, light, dissolution etc

Regime of QSAR methods

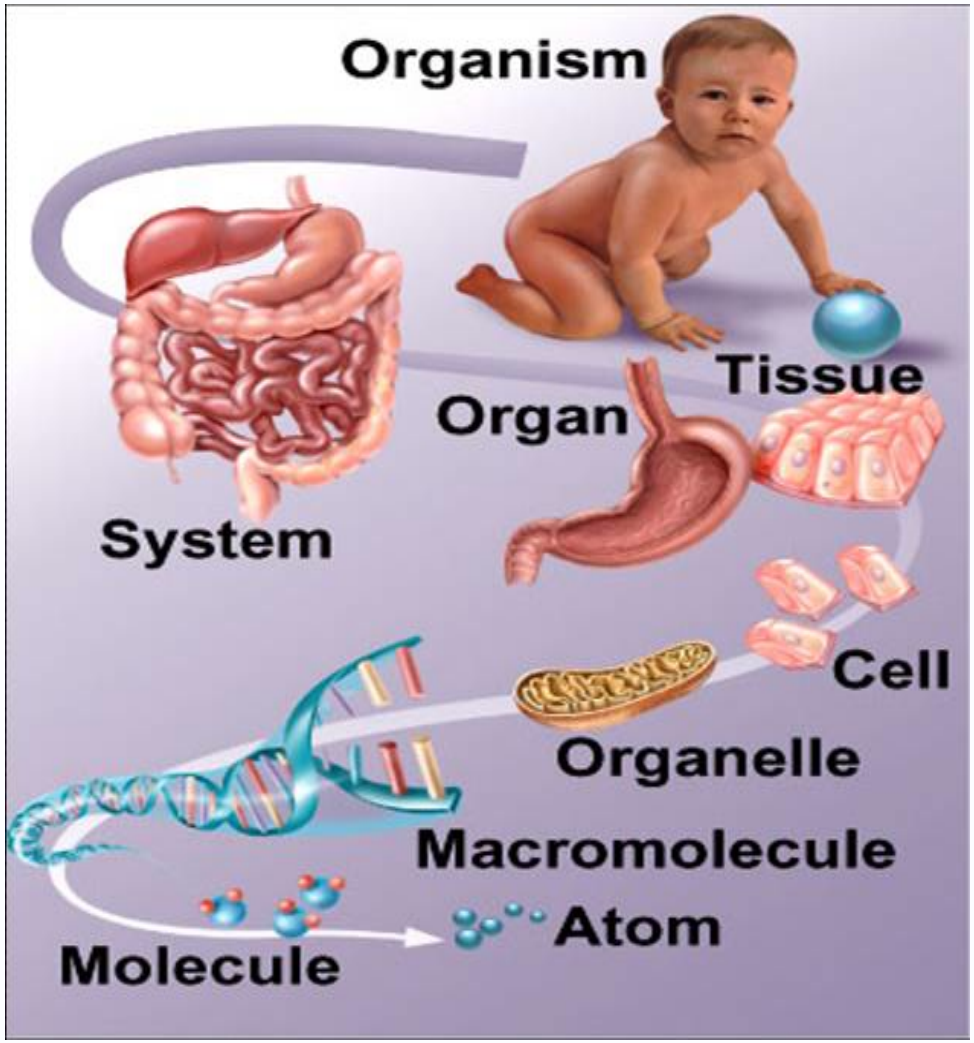
Potential detrimental effects on organisms





Cell-based experiments

Emergence and complex systems



Emergent properties of a complex system arise from interactions between lower level components. These properties do not exist in the components.

Consistent Concepts of Self-organization and Self-Assembly, Halley, JD, Winkler, DA, Complexity, 14(2), 10-17 (2008).

Classification of emergence and its relation to self-organization, Halley, JD, Winkler, DA, Complexity 13, 10-15 (2008).

Regime of QSAR methods

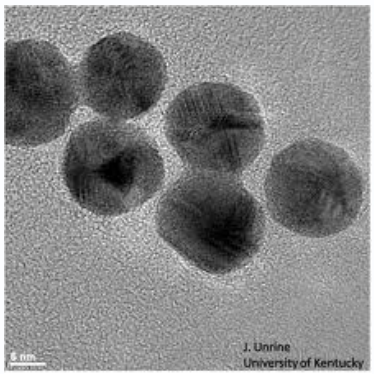
Image credit: P. Finkenstadt's class, via http://www.pc.maricopa.edu/.

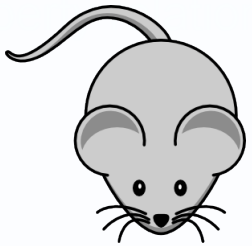
How do we model properties?

Modelling approaches can be categorized in many ways.

An artificial but useful classification is explicit (component level) and implicit (system level).

- Explicit methods attempt to predict or simulate system properties by describing all of the interactions of all components of the system ("bottom up"). Examples of explicit modelling methods include quantum mechanics, molecular mechanics and dynamics.



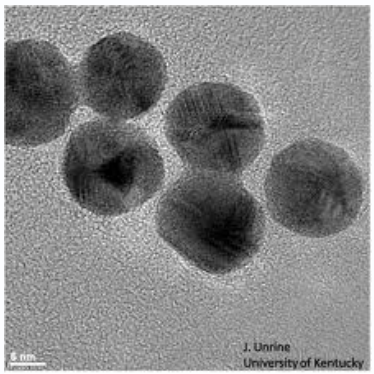


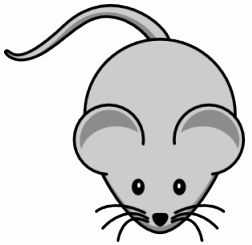
How do we model properties?

Modelling approaches can be categorized in many ways.

An artificial but useful classification is explicit (component level) and implicit (system level).

- Implicit methods predict the (emergent) properties of the system from the properties of the components without necessarily understanding all of their interactions (“top down”). Examples of implicit modelling methods include quantitative structure-property relationships (QSPR), pattern recognition, artificial intelligence.





Implicit modelling - QSAR/QSPR

- Biological systems and many materials have structures and interactions that are extremely complex
- Explicit modelling is extremely difficult or impossible in these cases.
- Implicit methods strip away all properties that are not relevant to the problem at hand. This allows us to focus on a limited set of properties in isolation.
- QSPR involves modelling of a higher level, or system property of a biological system or material using a mathematical representation of a molecular (microscopic, component) property and a very flexible pattern recognition algorithm.

"The whole is more than the sum of the parts." — *Aristotle*

Implicit modelling of materials: QSPR

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem." — John Tukey

However…

- oversimplifying throws away useful information
- overly complex models are poorly predictive and sensitive to noise, require too much data to train or validate, and are hard to interpret
- often problems are grossly underdetermined, with many independent variables and few dependent variable data points. e.g.
  - microarray experiments
  - QSPR models
  - docking
- Bayesian methods can provide solutions to many of these problems

Quantitative structure-activity modelling

Machine or statistical modelling methods in QSAR/QSPR/TNTR etc are **data driven**.

They work best for a **wide range** of different nanoparticles with **one or a few** measured biological properties.

They can generate useful models **without requiring all of the immense complexity** of the interactions of nanomaterials with biology being understood (model emergent properties)

Data from experiments measuring a wide range of biological, physicochemical etc properties for a **single** material need **other types of models** e.g. kinetic, cellular uptake, protein binding (molecular dynamics), redox properties (quantum chemistry calculations) etc

"All models are wrong, but some are useful"

Quantitative structure-activity modelling

Quantitative structure-activity relationships modelling (QSAR) was developed by Hansch and Fujita in the early 1960s to model physicochemical and biological properties of drugs.

In essence the method is *deceptively simple*. It is a supervised modelling method that describes the complex relationships between the molecular (microscopic) and physicochemical properties of materials and their biological (macroscopic) effects

Biological response (BR) = f(molecular properties)

The method involves finding relevant mathematical descriptions (descriptors) for the microscopic (molecular) properties and the optimum form for the (nonlinear) function $f = g_{corona} \times h_{uptake} \times j_{mechanism} \times k_{bioprocessing} \cdots$

It is essentially a kind of complex pattern recognition process. It can accommodate complex 'models within models'

Quantitative structure-activity modelling

Machine or statistical modelling methods in QSAR/QSPR/TNTR etc are **data driven**.

They work best for a **wide range** of different nanoparticles with **one or a few** measured biological properties.

They can generate useful models **without requiring all of the immense complexity** of the interactions of nanomaterials with biology being understood (model emergent properties)

Data from experiments measuring a wide range of biological, physicochemical etc properties for a **single** material need **other types of models** e.g. kinetic, cellular uptake, protein binding (molecular dynamics), redox properties (quantum chemistry calculations) etc

"All models are wrong, but some are useful"


Ceci n'est pas une pipe.

Good, bad, and useful models

Main steps in QSAR modelling

QSPR is a supervised learning method that needs a data set of materials and their biological properties. There are four steps…

1 Generate descriptors

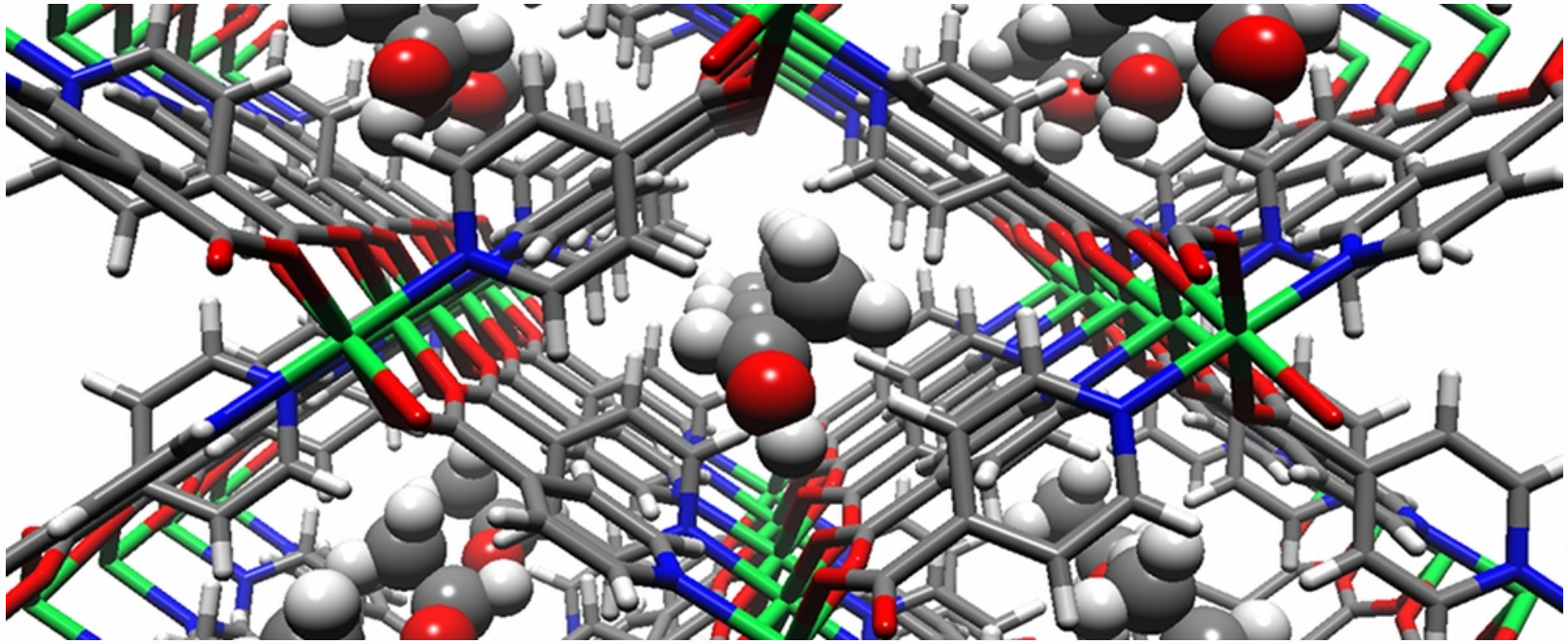
2 Select a sparse subset of descriptors in a context-dependent way

3 Deduce the potentially complex and nonlinear relationship between the descriptors and the property

4 Validate the model in terms of its robustness, prediction ability, and domain of applicability

The model can then be used to estimate the biological properties of new molecules where these data are not known

At the molecular scale materials can be viewed as a 3D distribution of properties



Clearly, structure and composition modulate the useful properties of materials- there are many ways of describing molecules mathematically

Main steps in QSAR modelling

QSPR is a supervised learning method that needs a data set of materials and their biological properties. There are four steps…

1. Generate descriptors
2. Select a sparse subset of descriptors in a context-dependent way
3. Deduce the potentially complex and nonlinear relationship between the descriptors and the property
4. Validate the model in terms of its robustness, prediction ability, and domain of applicability

The model can then be used to estimate the biological properties of new molecules where these data are not known

Descriptors

- Descriptors are mathematical representations of properties of molecular components (ligands, drugs, polymer repeat units, materials components)
- They can also be measured or calculated physicochemical properties of components (octanol-water partition coefficients, dipole moment, polarizability, HOMO-LUMO gap)
- There are thousands of different ways of describing molecules, just as there are many ways of describing a person, scene, or organism.
- Consequently, it is important to choose descriptors that are information-rich and relevant to the property to be modelled.
- There may be a 'universal' set of molecular descriptions, but this is still not a well-solved problem for QSPR
- **Development of specific descriptors for nanomaterials is a major area of research need in QSPR modelling**

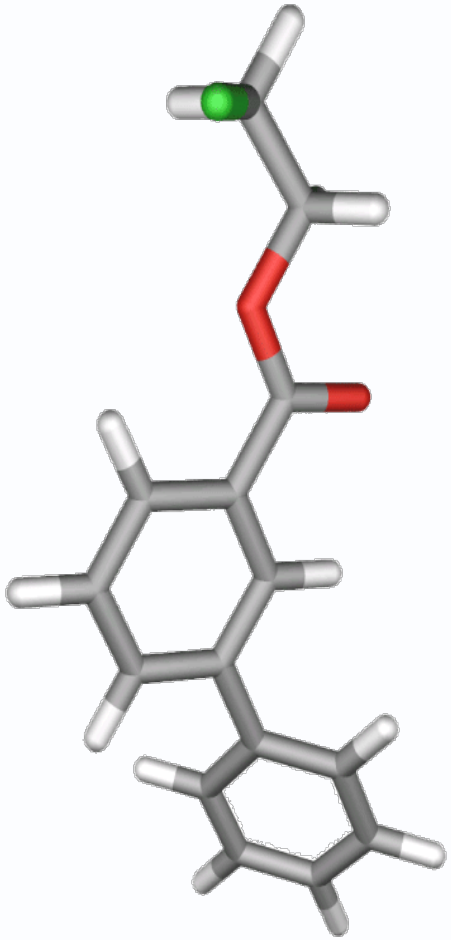
How do you describe a person?

- Size
- Build/shape
- Intelligence
- Odour
- Job
- Flexibility
- Volatility
- Weight
- Strength
- Colour
- Skills
- Handedness

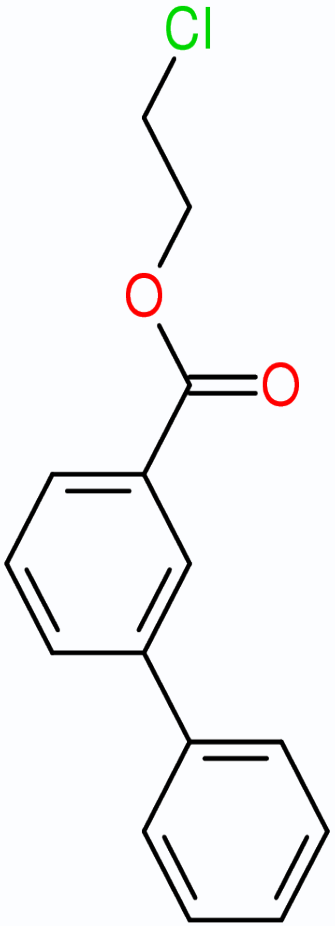
http://www.fanpop.com

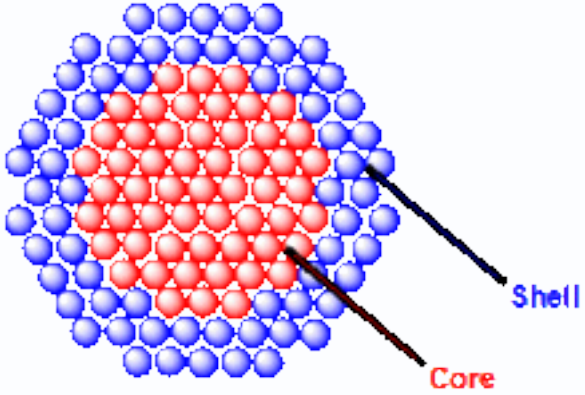
How do you describe a molecule?

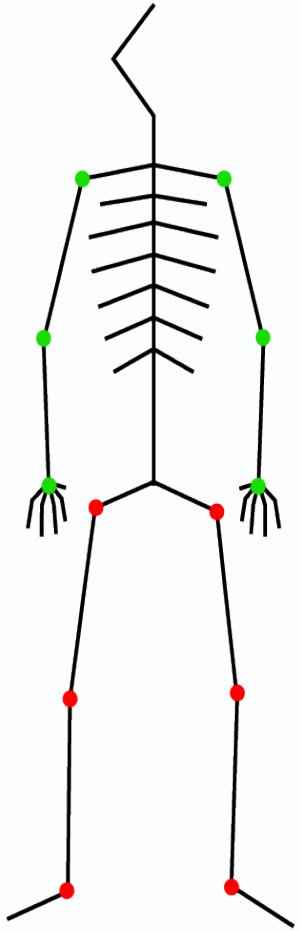
- Length
- Width
- Shape
- Flexibility
- Colour
- Aromaticity
- Complexity
- Chirality
- Reactivity
- Lipid solubility
- Volatility
- Molecular weight
- Functionality



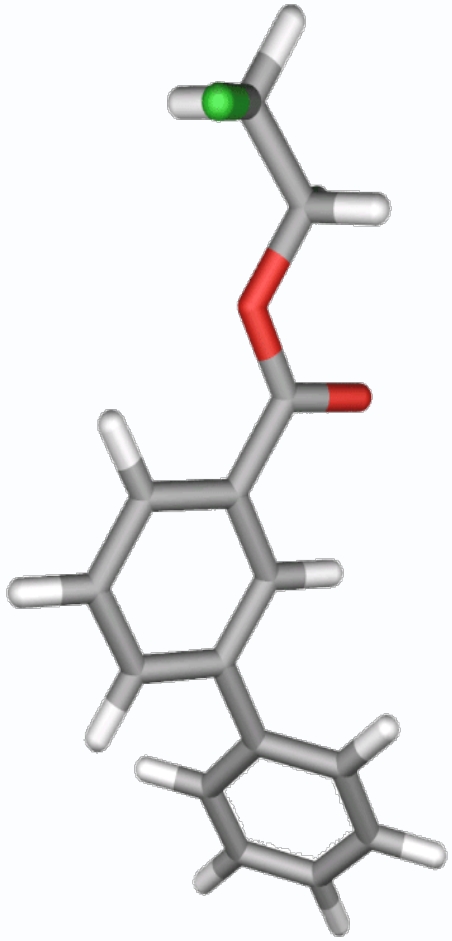
Molecules/nanoparticles have a topology/scaffold

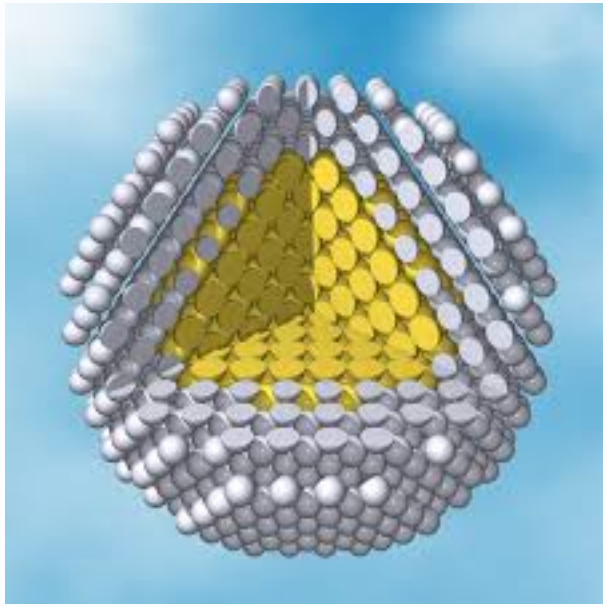


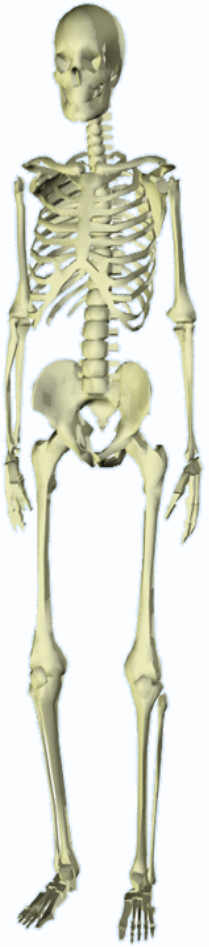




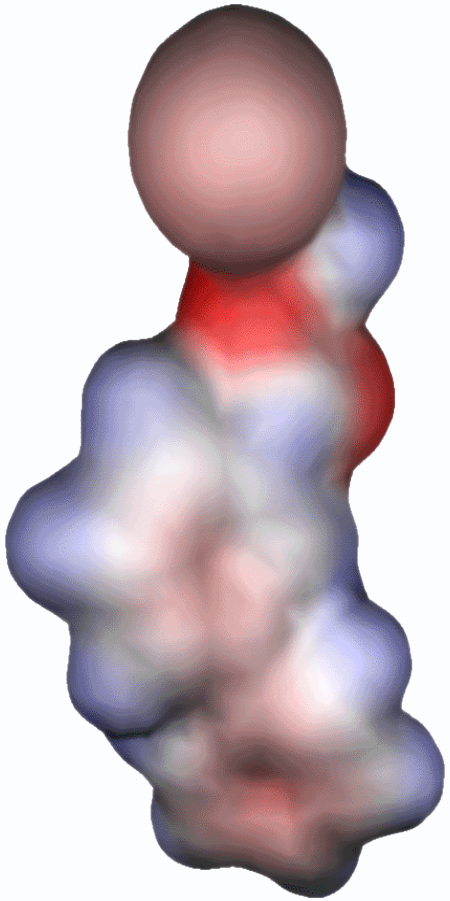
Molecules & nanoparticles are 3D objects

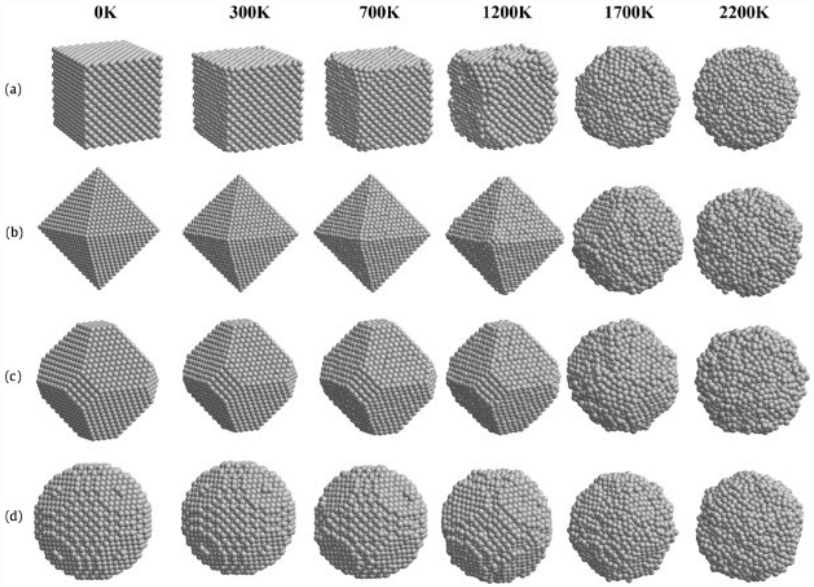






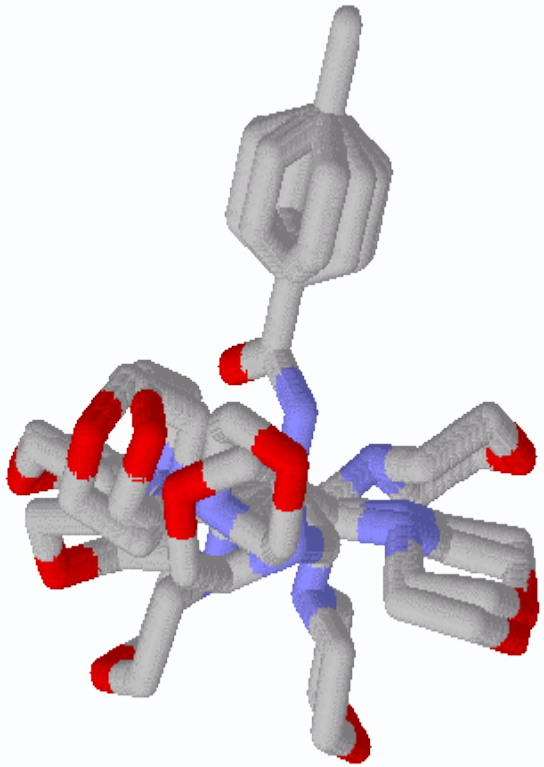
Molecules & nanoparticles have shapes and surfaces





Wen et al, Phys. Lett. A 2009

Molecules & nanoparticles are flexible or transform dynamically



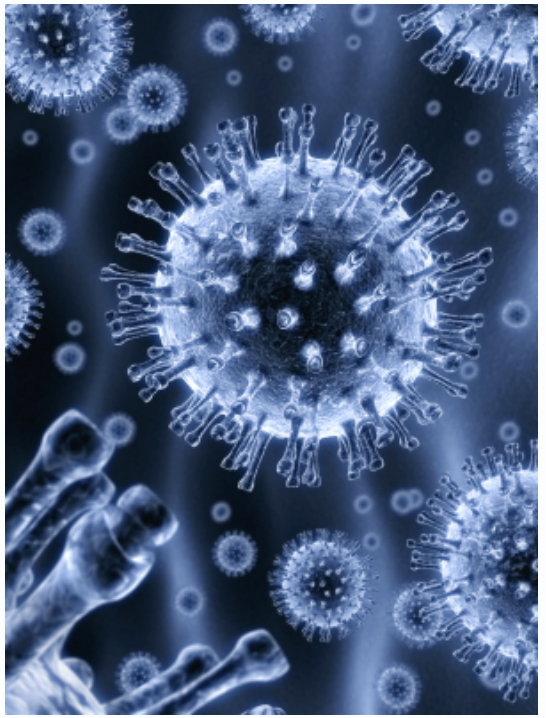
How do you describe a nanomaterial?

"A nanomaterial is a substance that, when injected into a rat, generates a paper"

How do you describe a nanoparticle?

- Atomization energy
- Volume fraction in solution
- Relaxivities R1 and R2
- Zeta potential
- Surface charge
- Surface charge density
- Size, and size distribution
- Shape/aspect ratio
- Surface area
- Surface structure
- Chemical composition
- Crystallinity
- Porosity
- Solubility
- Aggregation / Agglomeration
- Surface chemistry (functionalized nanoparticles)

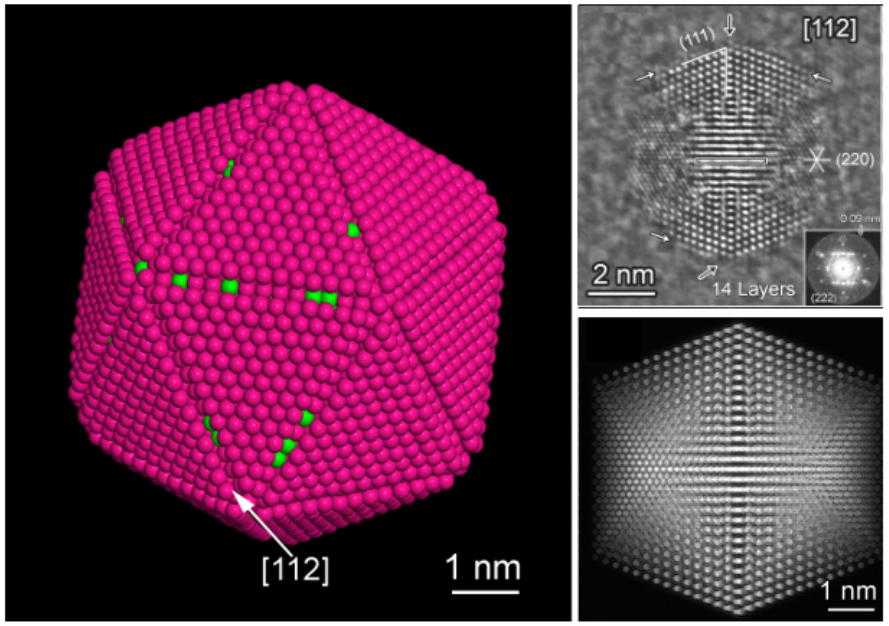
This is critically important



Liu et al. SMALL 2011 ASAP

Molecular description

Clearly, the usefulness of a description of a person, nanomaterial, or a molecule is *context-dependent*.
Descriptors used to construct QSPR models must similarly be chosen in a context-dependent way.
Descriptors must capture information relevant to the material property being modelled.
Nanoparticles with chemically modified surfaces are easier to model as we know how to describe small molecules



Quantum mechanical descriptors

Partial atomic charges

- Mulliken population analysis, Bader or similar
- Electrostatic potential-derived charges

Dipole moment

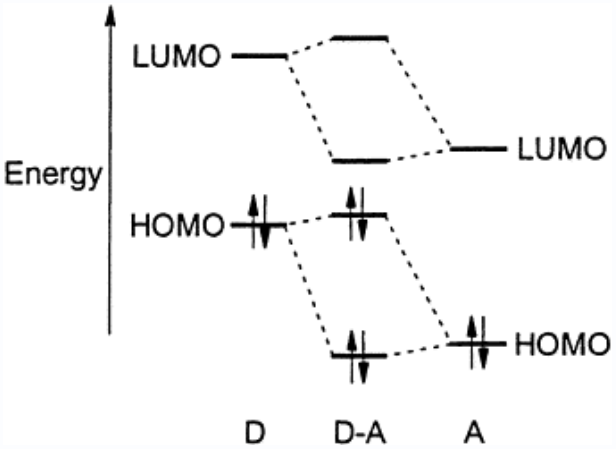
- Strength and orientation of behaviour of molecules in an electrostatic field

Frontier orbitals

- HOMO – energy of highest occupied molecular orbital (nuclephilicity)
- LUMO – energy of the lowest unoccupied molecular orbital (electrophilicity)

Superdelocalizability

- Estimate relative reactivity of atoms/ functional groups in molecules



Descriptor generation packages

Thousands of 1D, 2D, and 3D (even 4D and 5D!) descriptors have been invented.

Common molecular descriptor programs are: -

- DRAGON
- ADRIANA
- CODESSA
- SYBYL (3D molecular fields)
- We have our own in-house (ABC) descriptors

Main steps in QSAR modelling

QSPR is a supervised learning method that needs a data set of materials and their biological properties. There are four steps…

1 Generate descriptors

2 Select a sparse subset of descriptors in a context-dependent way

3 Deduce the potentially complex and nonlinear relationship between the descriptors and the property

4 Validate the model in terms of its robustness, prediction ability, and domain of applicability

The model can then be used to estimate the biological properties of new molecules where these data are not known

Why select descriptors?

We cannot and should not use all available descriptors because

- The system will be over determined (not enough data to determine the coefficients of each descriptor)
- The model may be overfitted
- No model may be obtained because of confounding effects of irrelevant variables (even when PLS and PCA used)
- Poor selection of descriptors from a large pool of possibilities can lead to chance correlations
- Selection of a sparse subset of descriptors needs to be done in a context (property) dependent way.

Feature Selection (FS) Problem

Modern software packages allow you to generate hundreds or even thousands features—not all of them are useful!

Why FS?

- Reduces dimensions, facilitating data visualization and interpretation;
- Reduces measurement and storage requirement;
- Reduces training and utilization times;
- Generously improves prediction performance;
- Removes irrelevant information or noise from a model.

Descriptor selection

- We can generate many descriptors for a given structure/material.
- We need a small set of relevant descriptors to optimize predictivity.
- Principal components analysis (generating a smaller number of orthogonal latent variables from a large number of descriptors) is often used but has disadvantages.
- Good, statistically sound methods of ‘feature detection’ or descriptor selection have been developed.
- An ideal feature selection method should completely remove uninformative descriptors and retain only those that are relevant to the problem, making the interpretation of the model easier.
- Feature selection aims to choose a small number of the most informative descriptors in a context-dependent way.

FS methods: wrappers and filters

Wrapper: utilizes the choice of prediction method to score subsets of features according to their predictive power;

Seem to be a "brute force" method but greedy search strategies are available (best-first, branch-and-bound, simulated annealing, genetic algorithms, etc.)

Two flavors:

- **Forward selection**: features are progressively incorporated into larger and larger subsets;
- **Backward elimination**: starting with the set of all features and progressively eliminates the least promising ones.

Example: Recursive feature elimination (RFE).

Feature Selection (FS)

Earlier, we briefly described linear dimension reduction (e.g. PCA) and feature selection (e.g. forward selection) methods

There are more complex and efficient methods for achieving this, some of which do nonlinear feature selection.

Removing uninformative weights

Principal Component Analysis and Partial Least Squares are commonly used to reduce the dimensionality of model but do not remove uninformative weights from the regression.

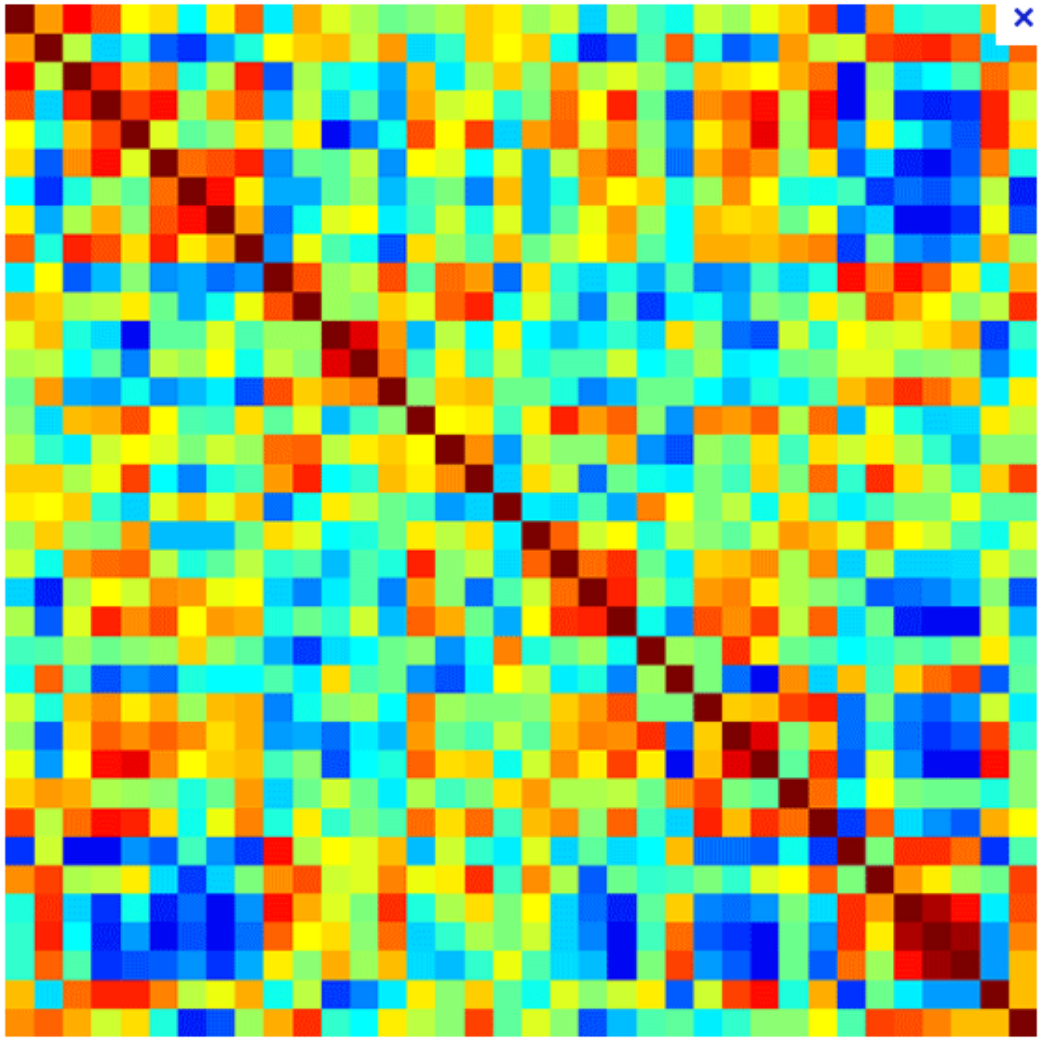
Clark and Cramer's work has shown that retaining uninformative descriptors can seriously degrade the performance of models

M. Clark, R. D. Cramer, Quant Struct-Act Rel 1993, 12(2), 137-145.

Removing uninformative weights

- Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci.*, 2000, 40 (5), pp 1160–1168. This method is limited to linear regression.
- The use of Automatic Relevance Determination in QSAR Studies using Bayesian Neural Networks. *Burden, F.R.; Ford, M.; Whitley, D.; Winkler, D.A.,J. Chem. Inf. Comput. Sci. 2000, 40, 1423-1430*.This non-linear method proved to be suitable only for models with a limited a limited number of variables
- Optimum QSAR Feature Selection using Sparse Bayesian Methods, Burden, FR, Winkler DA *QSAR Comb Sci*. 28, 645-653, (2009). Suitable for removing less relevant variables from very large sets (30,000 variables)

Correlation matrix



QSPR sparse feature selection-advantages

- A small number of most relevant descriptors to be chosen from a potentially large pool of possibilities in a supervised, objective way that obviates chance correlations
- Generation of QSAR models using robust, nonlinear methods such as neural networks that are also interpretable (if interpretable descriptors used)
- Limitations are that the method currently uses a linear (MLR) method to find sparse features, as do methods such as PCA. This limitation can be removed with difficulty.

What are Bayesian methods?

Bayes' theorem relates conditional and marginal probabilities of events A and B, where B has a non-vanishing probability:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- P(A) is the prior probability or marginal probability of A. It is "prior" – it does not take into account any information about B.
- P(A|B) is the conditional probability of A, given B. It is also called the posterior probability because it depends on the specified value of B.
- P(B|A) is the conditional probability of B given A.
- P(B) is the prior or marginal probability of B – a normalizing constant.

Intuitively, Bayes' theorem quantifies how beliefs about observing 'A' (prior knowledge) are updated by having observed 'B'.

Regression by expectation maximization

It is important to choose a small number of variables relevant to the problem. We use sparse Bayesian feature selection methods based on an expectation (or likelihood) maximization(EM) algorithm.
Regular Multiple Linear Regression(MLR) uses a Gaussian prior

$$p(w \mid \alpha) = \prod_{i=1}^{N_V} \frac{\alpha}{2} \exp(-\alpha w_i^2) = \left(\frac{\alpha}{2}\right)^{N_V} \exp(-\alpha \parallel w_i \parallel_2)$$

$$\parallel w_i \parallel_1 = \sum_i \mid w_i \mid$$

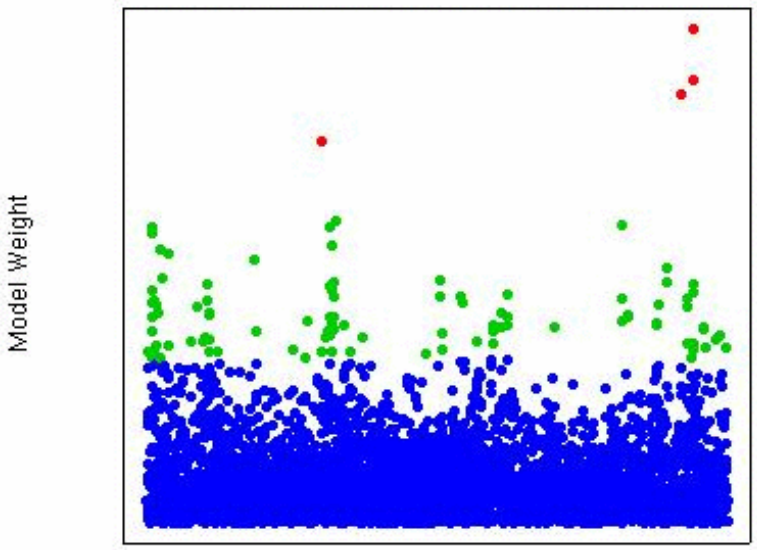
where the **w** are the MLR coefficients.
Multiple Linear Regression with expectation maximisation (MLREM) uses a Laplacian prior with well-known sparse properties

$$p(w \mid \alpha) = \prod_{i=1}^{N_V} \frac{\alpha}{2} \exp(-\alpha \mid w_i \mid) = \left(\frac{\alpha}{2}\right)^{N_V} \exp(-\alpha \parallel w_i \parallel_1)$$

We have modified this to provide tuneable sparsity to match data

Feature selection using expectation maximization

These sparse Bayesian feature selection methods can very effectively deliver a relatively small number of relevant features very efficiently.



*Figeuiredo, IEEE Trans Patt Anal Mach Intell , **25**, 1150 (2003)*
*Burden, Winkler, QSAR Comb Sci. **28**, 645-653, (2009)*

Advantages

- Optimal (tunable) parsimonious models- controls overfitting and maximizes prediction accuracy
- No composite variables – models more easily interpreted.
- Supervised and objective - selects features according to Bayesian evidence - minimizes or eliminates chance correlations.
- Specifically developed to analyze very large datasets, and grossly underdetermined systems – very efficient.
- Robust and repeatable –offers a solution to one of the most vexing problems in QSPR, selecting the most relevant descriptors for use in models

Main steps in QSAR modelling

QSPR is a supervised learning method that needs a data set of materials and their biological properties. There are four steps…

1 Generate descriptors

2 Select a sparse subset of descriptors in a context-dependent way

3 Deduce the potentially complex and nonlinear relationship between the descriptors and the property

4 Validate the model in terms of its robustness, prediction ability, and domain of applicability

The model can then be used to estimate the biological properties of new molecules where these data are not known

Quantitative structure-property modelling

Quantitative structure-activity relationships modelling (QSAR) was developed by Hansch and Fujita in the early 1960s to model physicochemical and biological properties of drugs. When applied to nonbiological properties it is called quantitative structure-property relationships (QSPR)

The method is *deceptively simple*. It is a supervised modelling method that describes the complex relationships between the molecular (microscopic) and physicochemical properties of molecules and their biological (macroscopic) effects

Biological response (BR) = $\mathcal{F}$(molecular properties)

The method involves finding relevant mathematical descriptions (descriptors) encoding microscopic (molecular) properties and the optimum form for the (nonlinear) function $\mathcal{F}$

It is essentially a kind of complex pattern recognition process

Structure-property mapping

- Finding the relationship between the microscopic descriptors and the macroscopic property can be done in many ways.
- Common ways are multiple linear and polynomial regression and related methods.
- More recently, machine learning methods like neural networks and support vector machines have been used. These are more efficient and remove subjectivity from the modelling process.
- However, regression is an “ill-posed” problem and instability or artefacts can arise.
- The mapping process needs to find the best balance between bias and variance.
- Regularization helps as it penalizes overly complex models by use of a weight penalty, but finding the correct regularization constants is difficult.

Finding structure-activity relationships

There are many methods of varying sophistication

Simple linear statistical regression methods like multiple linear regression

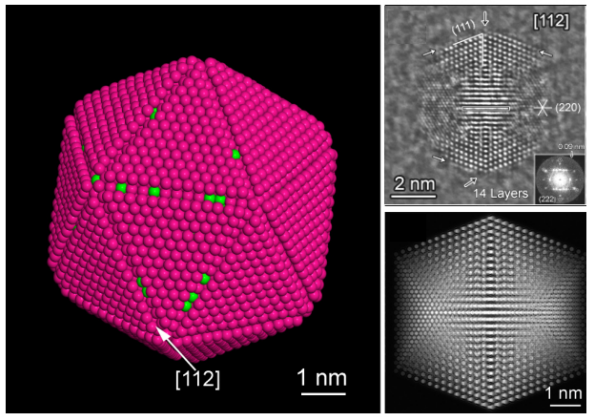
$$BR = a + bx1 + cx2 + \ldots$$

Nonlinear regression methods using polynomials or kernel functions (e.g. Gaussians) (lecture 3)

$$BR = a + bx + cx^2 + dx^3 + \ldots.$$

$$BR = a + b\phi_1 + c\phi_2 + d\phi_3 + \ldots.$$

Nonlinear machine learning methods like neural nets



Quantitative Structure-Property Relationship Modeling of Diverse Materials, Le, T.C., et al. Chem. Rev. ***112,*** *2889 (2012).*

Linear models – Classical QSAR

Hansch and Fujita are recognized as having first introduced MLR into QSAR in the form of the following Hansch equation:

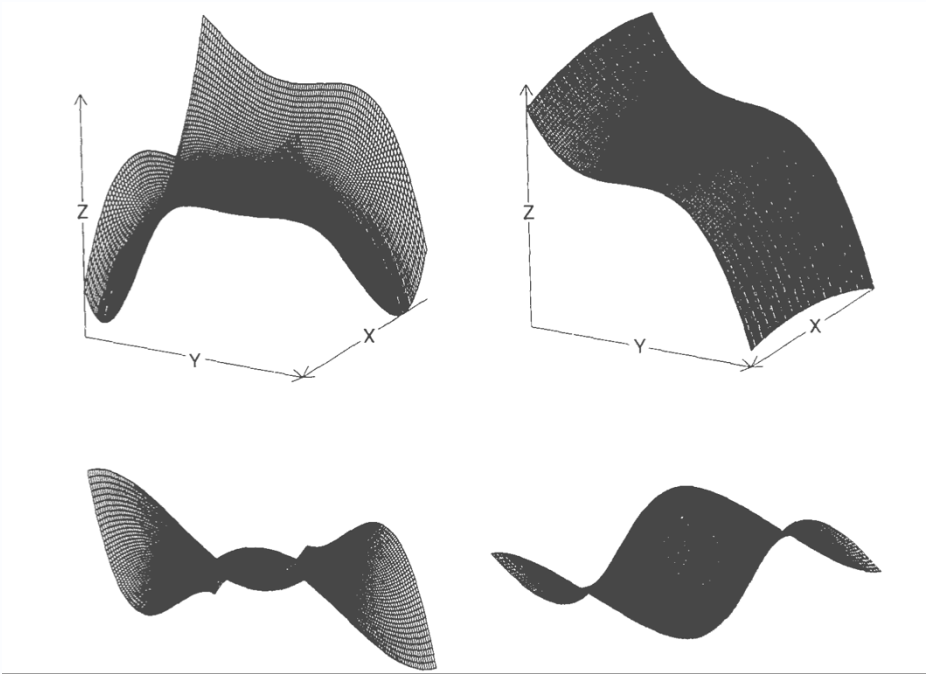
$$\log (1/C) = a(\log P)^2 + b(\log P) + C\sigma + dE_S + \ldots + e$$

a-e are constants determined for a particular reaction or biological activity by Multiple Linear Regression analysis.

logP, π, σ, F,R, E_S etc, are independent variables whose values are obtained directly from experiment or from tabulations; other parameters than those shown may also be included.

Complex, Nonlinear, Multidimensional Response Surface

Property= F (Structure, physicochemical Properties)



Modelling complex, nonlinear relations

Linear methods (e.g. MLR) can generate good models.

However, in many cases, the structure-activity relationship is *nonlinear*.

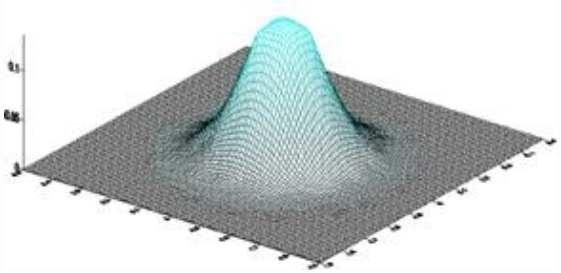
Polynomial regression methods, nonlinear kernel methods, and neural network are then the methods of choice for QSPR modelling.

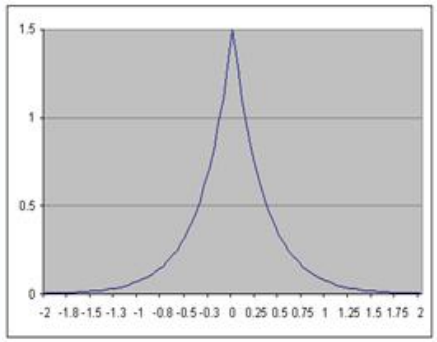
Neural networks are very useful because they are nonlinear universal approximators, but can generates poor models if care is not taken.

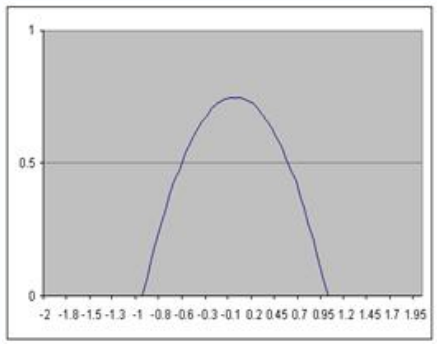
Neural networks can also be overtrained, becoming better and better at predicting (memorizing) training data, and worse at predicting new data. Techniques exist to avoid overtraining.

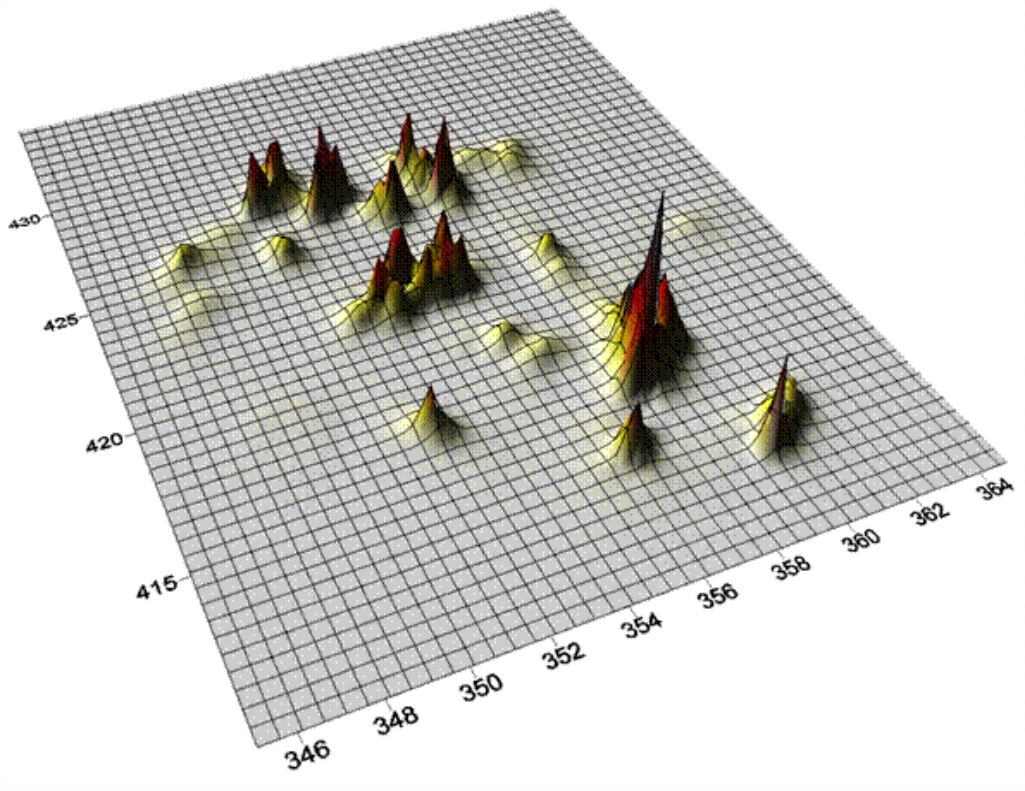
Bayesian regularized neural nets automatically choose the optimum complexity of a QSAR model – achieving the best balance between bias and variance

Kernel regression









http://www.spatialanalysisonline.com/output/html/Pointdensity.html

Ridge regression and regularization

The least squares cost function has an extra regularising term

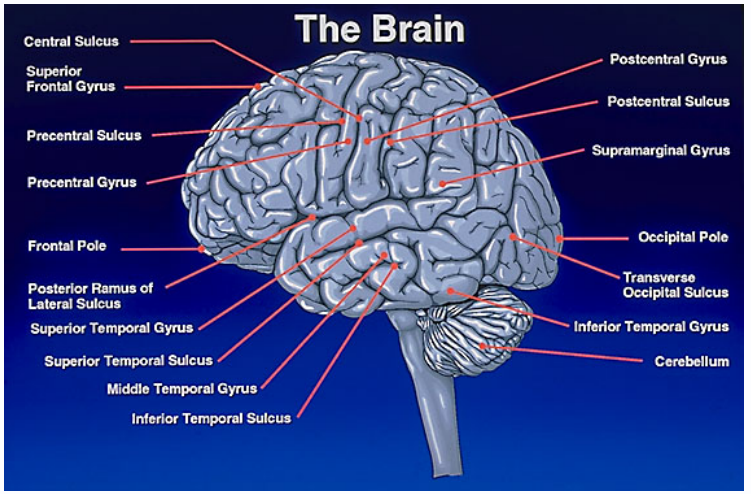
$$\mathrm{CF(w)} = \sum_i \left(y_i - \hat{y}_i\right)^2 + \lambda \sum_i w_i^{\,2}$$

The second term has the effect of producing a model with small weights making the model more robust and predictive.

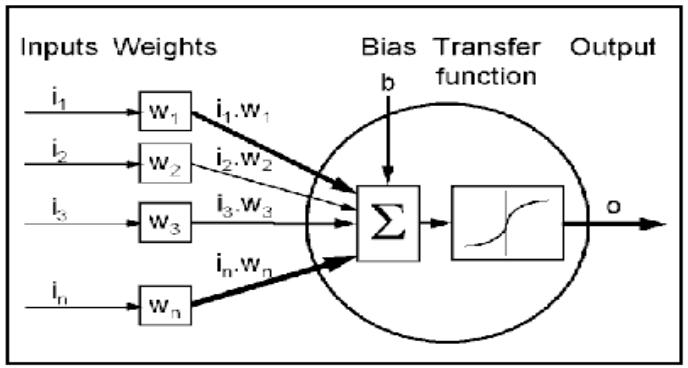
More complicated regularizers can also be used.

Neural nets

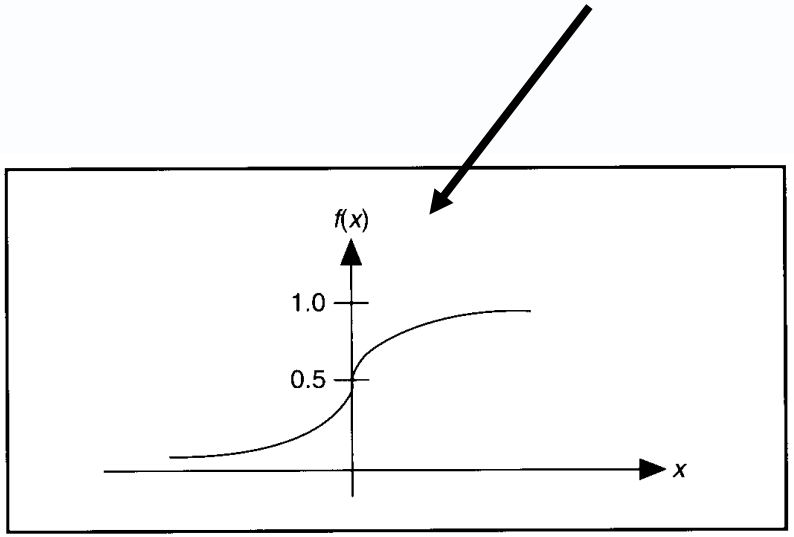
- Neural networks are a relatively recent way of finding complex relationships between structure and activity
- They are modelled on the human neural system and learn relationships between data in a similar way to humans
- They are ‘universal approximators’ capable of learning any complex nonlinear relationship given enough training data



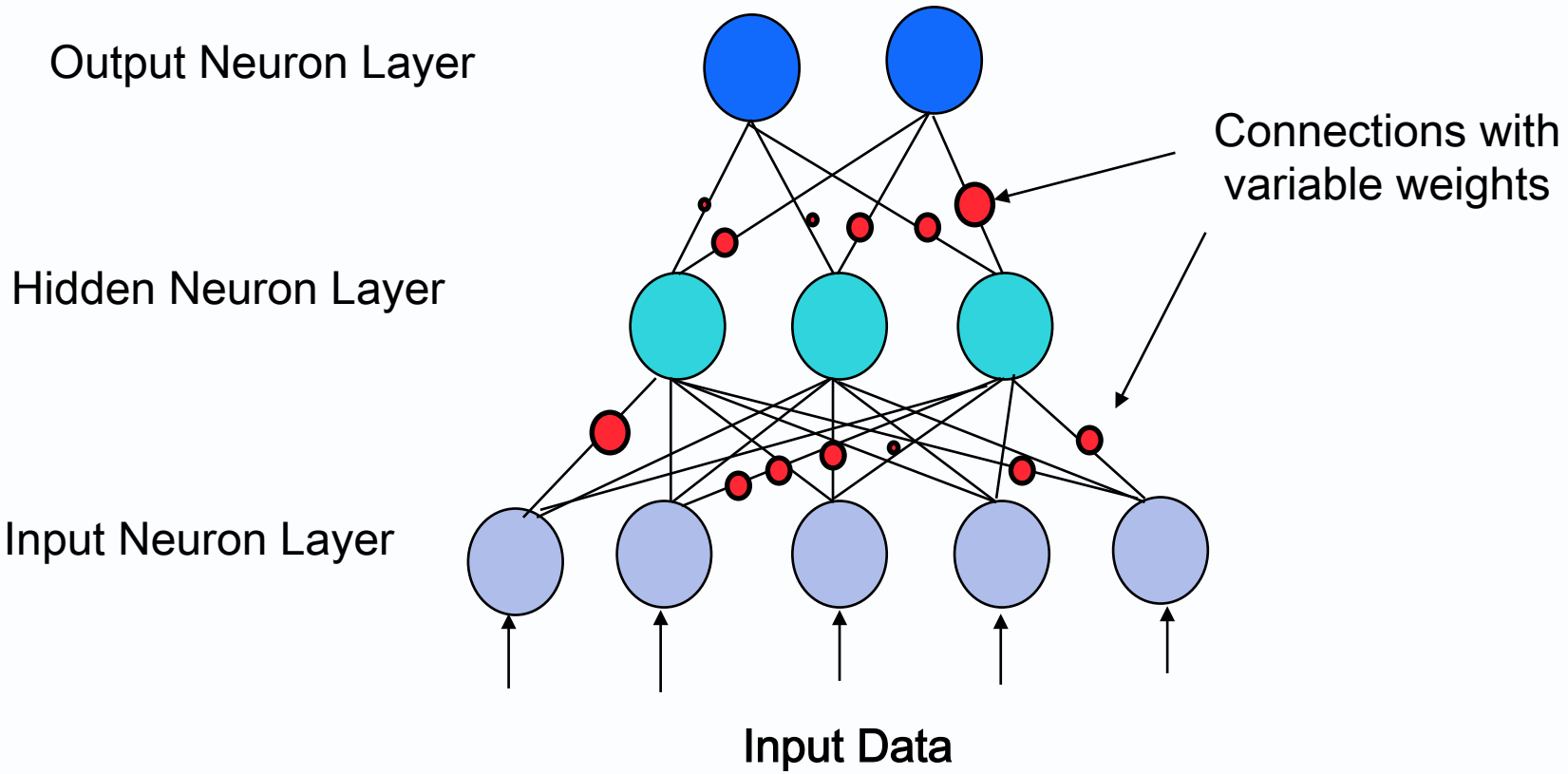
An artificial neuron or neurode



Sigmoid



Neural Network



Back-propagation

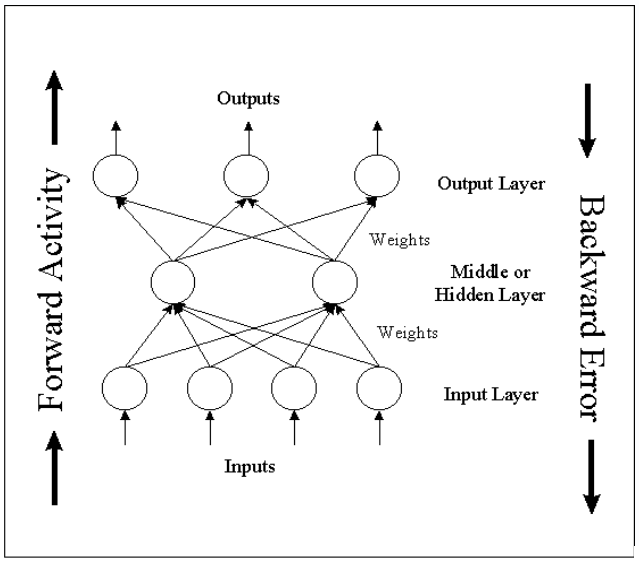
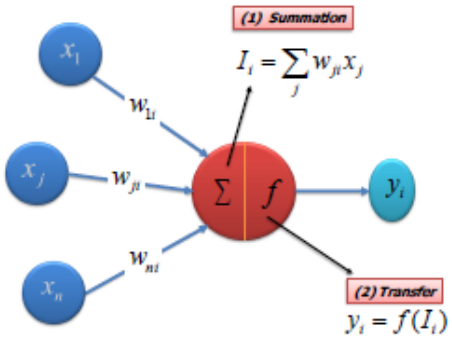
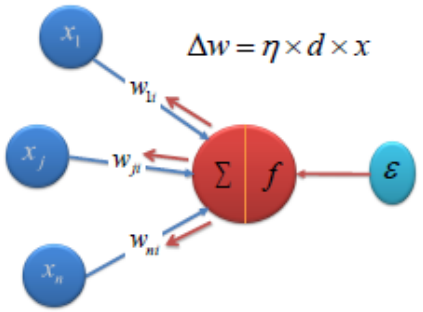


FIGURE 4.1. Typical Artificial Neural Network Setup (Caudill and Butler, 1992a).

Feedforward Input Data

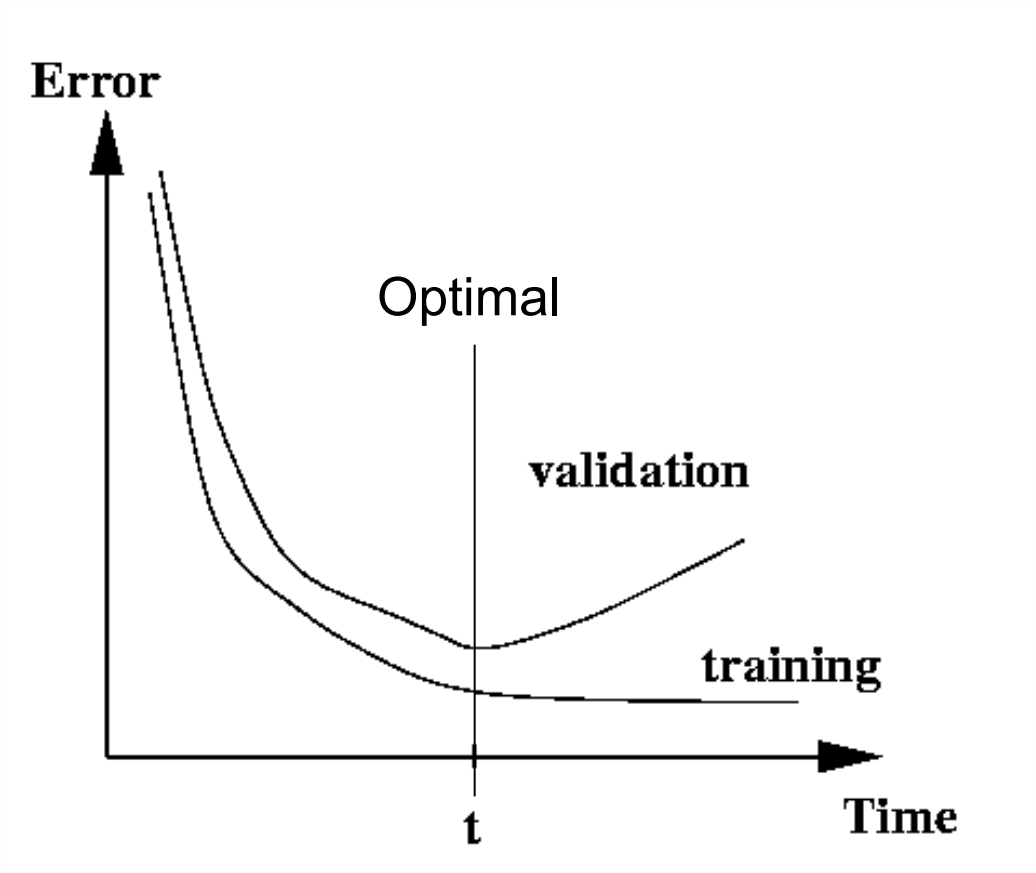


Backward Error Propagation



Avoiding neural network overtraining

Early stopping using validation set:



Advantages of neural nets

- Model free nonlinear mapping system
- Generalising ability
- Subjective decisions on SAR not needed
- Can recognise patterns
- Can deal with fuzzy, noisy or missing data
- Bayesian methods make them very robust
- May deal with complex systems well

Disadvantages of neural nets

- Can overfit data by generating too complex model
- Can be overtrained so that they memorize training data and can't predict so well
- Hard to optimize net architecture
- Hard to interpret model
- These problems largely overcome by Bayesian methods

Bayesian Regularized Artificial Neural Network with a Gaussian Prior

This is a back-propagation network with one hidden layer and a single dependent variable.

The cost function (regularizer) is given by

$$CF(\mathbf{w}) = \beta \sum_{i=1}^{n} (\mathbf{y}_i - f(\mathbf{x}_i))^2 + \alpha \sum_{j=1}^{p} w_j^2$$

where α and β are hyper-parameters adjusted to maximise the log(likelihood) (≡ expectation maximisation)

This method has proved to be robust and reliable as well as providing an estimate of the number of effective parameters (weights) used in the model.

*Burden, F.R. and Winkler, D.A. J. Med. Chem., **42**(16); 3183-3187 (1999).*
*Winkler,D.A., and Burden,F.R. Mol. Simul.2000, **24**, 243-258.*

Optimal self-pruning neural network and nonlinear descriptor selection

Again, using a Laplacian prior (LP) in a backpropagation artificial neural network we must minimise.

$$M(\mathbf{w}) = \beta E \; + \alpha E_W \;\; = \beta \sum_{i=1}^{N_D} (\mathbf{y}_i - f(\mathbf{x}_i))^2 + \alpha \sum_{j=1}^{N_W} |w_j|$$

By assigning non-informative priors to α and β and integrating them out we are left with maximising the loss function *L* .

$$L = \frac{1}{2} N_D Log E_D + N_W Log\, E_W$$

which was employed by our Bayesian regularised artificial neural network to generate BRANNLP.

Unnecessary weights are driven to zero and if all the weights associated with a particular descriptor are driven to zero then the descriptor is discounted in the model.

Classification

Sometimes biological data are not available as continuous variables, rather as classes or categories. It is still possible to build models using these types of data (qSPR)

Categories could include things like: effective/ineffective, inactive/ weak/moderate/potency etc

The common QSPR methods can be used if the category data are converted into numbers e.g. active/inactive could be represented by 1/0 or -1/1, inactive/weak/moderate/potent could be represented as 1/3/5/7.

Classification

qSPR models can then give predictions as to the likelihood (probability) that a molecule belongs to a given class.

Other methods such as cluster analysis, self-organizing maps can also be used

Recently a set of very powerful mathematical methods called support vector machines (SVM) have been developed that can do classification very well. We are applying a sparse form of SVM, the Relevance Vector Machine (RVM) to QSPR problems. Detailed description of these methods is beyond the scope of this course

Modules to carry out these types of analyses are available in KNIME that we will learn about later

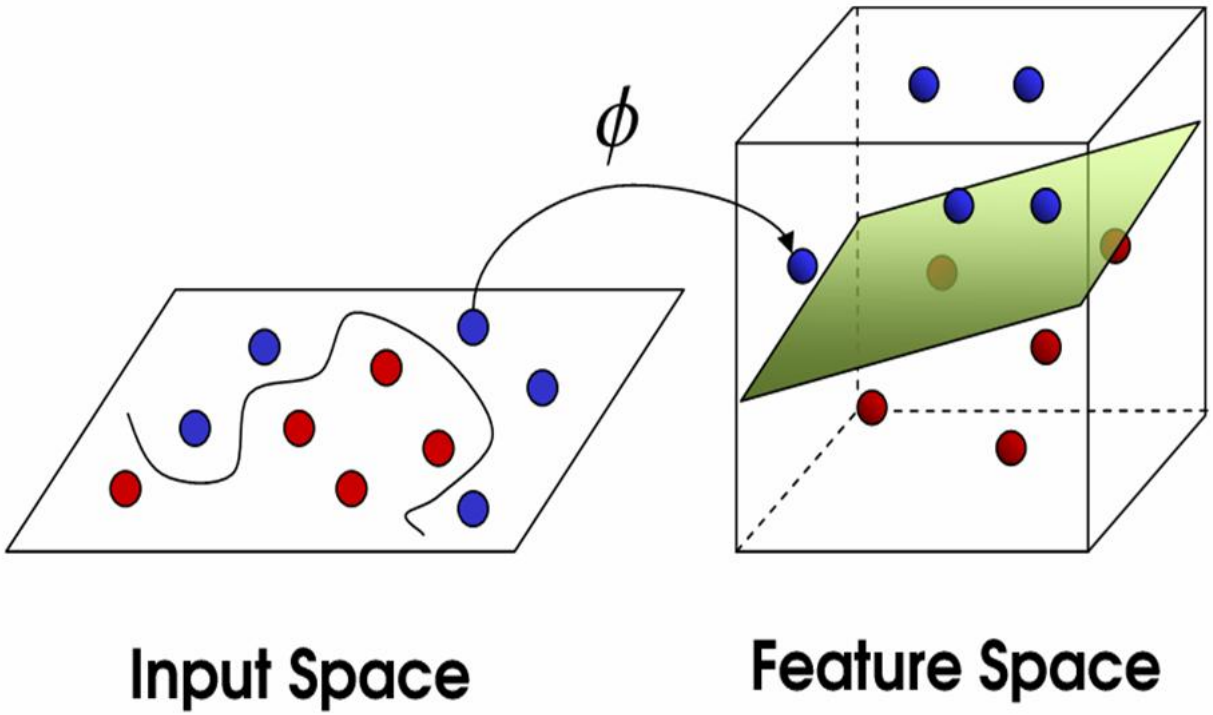
Support vector machines

Some classification problems cannot cleanly divide a set of materials into two classes e.g. protein fouling versus non-fouling.

A support vector machine constructs a hyperplane or set of hyperplanes in a high- dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Modules to carry out these types of analyses are available in KNIME that we will learn about later

Support vector machines



http://www.imtech.res.in/raghava/rbpred/svm.jpg

Main steps in QSAR modelling

QSPR is a supervised learning method that needs a data set of materials and their biological properties. There are four steps…

1 Generate descriptors

2 Select a sparse subset of descriptors in a context-dependent way

3 Deduce the potentially complex and nonlinear relationship between the descriptors and the property

4 Validate the model in terms of its robustness, prediction ability, and domain of applicability

The model can then be used to estimate the biological properties of new molecules where these data are not known

Validation

- Models need to be validated, that is, assessed for how well they model predict new data.
- The domain of applicability of the model also needs to be understood so that reasonable extrapolation can be done.
- Bootstrapping, jack-knifing, and cross validation methods like “leave-one-out” are commonly employed to achieve this.
- However, Kubinyi, and Tropsha (among others) have shown that cross validated statistics do not correlate well with those from independent test sets (“Beware of q2”, Golbraikh & Tropsha, J. Mol. Graph. Mol. 20, 269 (2002)).
- An independent test set of data never used to develop the model is the “gold standard” for validation.
- These concepts are well understood but still poorly applied.

Validating QSPR models

- Cross validation methods involve leaving one or more data points out of the training data, building the model, and predicting the property for the omitted data. This is done multiple times until all data in the set has appeared in the training and cross validation sets at least once.
- It has been shown that leave-one-out (LOO) cross validation and external test set metrics do not correlate significantly (Tropsha).
- Bootstrapping methods involve taking the original data set and sampling from it to form a new sample (called a 'resample' or bootstrap sample) of the same size. The bootstrap sample is not identical to the original data set as it is taken from the original using sampling with replacement. This process is repeated typically 1000 or 10,000 times.
- Although wide used, these methods have been shown to give an overly optimistic estimate of the predictive power of the model.

Validating QSPR models

- The gold standard for estimating model predictive power is to predict properties for new materials that were not used in building the model. This is approximated by partitioning the data set into training and test sets.
- The training set is used to generate the model (usually 50-80% of the data set)
- The properties of the test set are predicted by the model. Data from the test set are never used to build the model, so are ‘independent’.
- Test sets may be chosen randomly (best for large data sets) or by some kind of cluster analysis (better for smaller data sets)
- The best validation is to predict something completely unknown, synthesize it, and test the prediction.

Quality metrics for QSPR models

Many metrics are in the literature. Common metrics are…

r^2, q^2. Coefficient of determination – square of the correlation coefficient for the prediction of the training, cross-validation or test sets. Describes how much of the variance in the data can be explained by the model. Depends on the number of descriptors and fitted parameters in the model, so potentially less useful for comparing models. 1=perfect model, 0=no model. For good models these two are similar.

SEE, SEP. Standard error of estimation (training set) and prediction (test set). This is a measure of the estimated error in the predictions. They are independent of the details of the modelling process, so a better basis for comparing models. Should be as low as possible, but not lower than the error in the measured data. For good models these two are similar.

QSPR modelling traps & pitfalls

- Uninformative descriptors
- Overfitting and grossly underdetermined systems
- Poor descriptor selection and chance correlations
- Modelling complex, nonlinear structure property relationships with linear models
- Incorrectly validating QSPR models
- Not understanding the domain of applicability of models
- Overtraining neural network models

Uninformative descriptors

Descriptors must contain information *relevant* to the biological property being modelled. This is an especially important issue for nanomaterials.

They may be uninformative for two main reasons.

- may *not contain relevant information*, making the construction of a useful model impossible. Usually uninformative descriptors of this type are not problematic, as it is clear when the model is poor.
- may contain information relevant to the property being modelled but may be *obscure or arcane properties* derived from quantum chemical calculations or topological (connectivity) properties of the structure. Although such descriptors can generate successful and useful models, it is extremely hard to understand how the microscopic properties influence the macroscopic (measured) properties in a useful way.

Overfitting

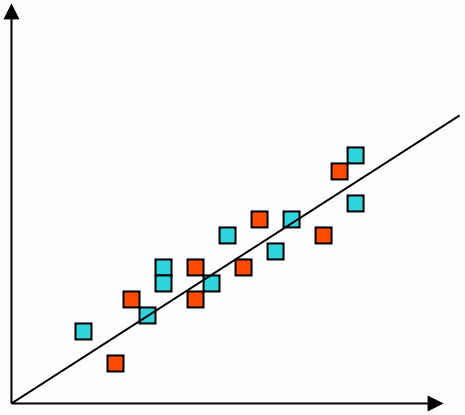
QSAR models are *overfitted* if the number of adjustable parameters in the model (e.g. coefficients in an MLR model, or weights in a neural network) exceeds the number of data points available to be modelled. In statistical terms, these are called grossly underdetermined systems.

Overfitted models are very good at predicting the training data, but *very bad* at predicting new data.

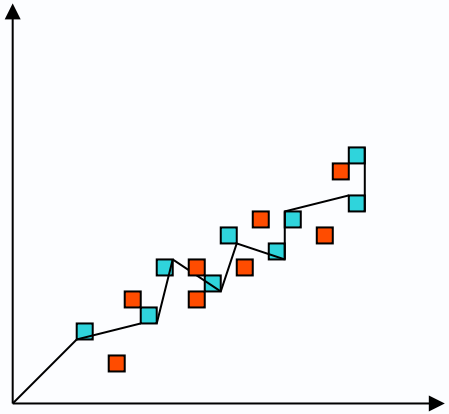
Parsimonious models with a simple structure and small number of descriptors generally have *higher predictive power* than more complex models.

Overfitting can be minimized by ensures the number of fitted parameters in the model (coefficients in a regression or weights in a neural network) are much less than the number of examples (materials or molecules) in the training set.

Overfitting Problem



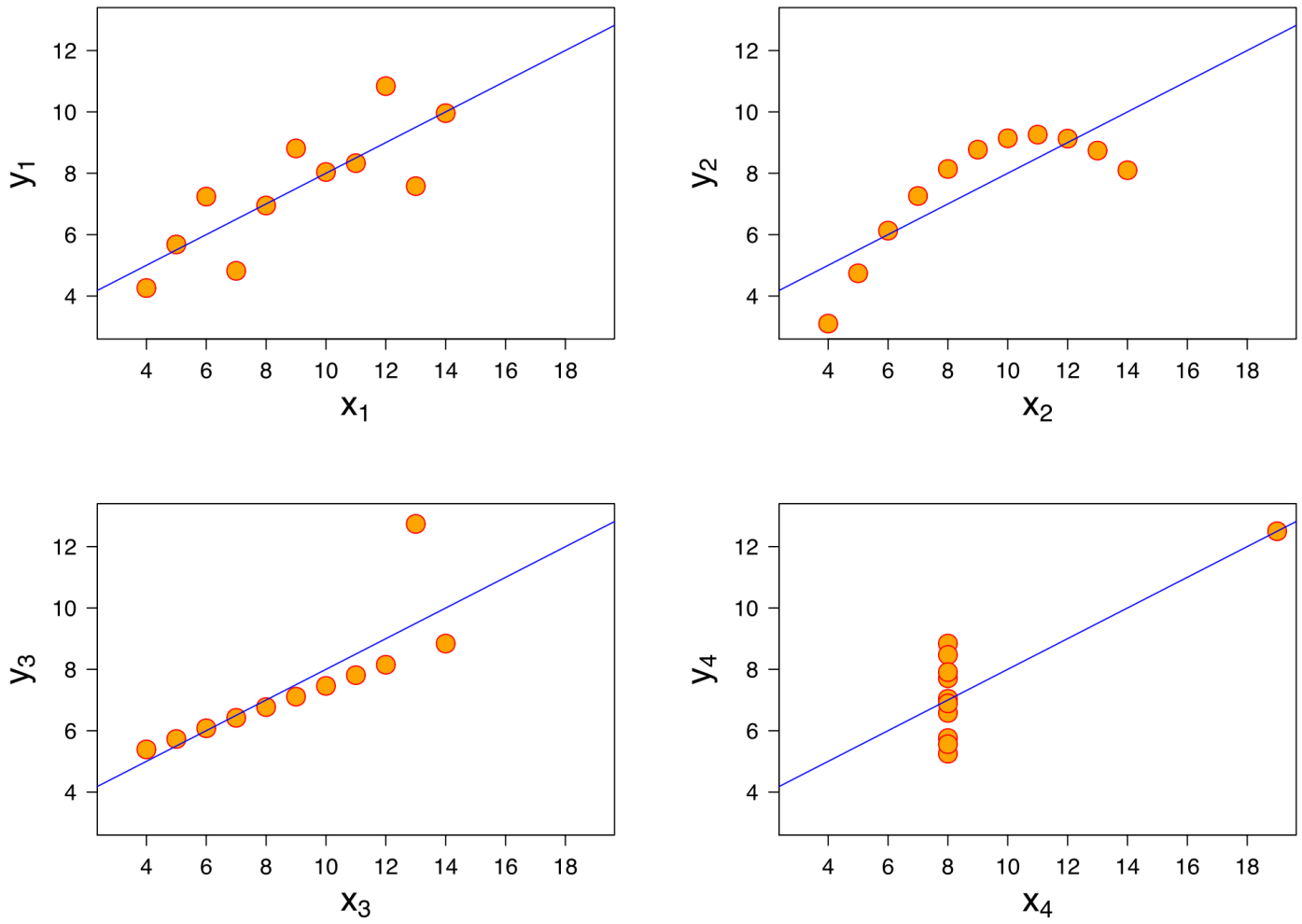
Simple model with training error but giving good prediction.



Complex model with no training error but giving poor prediction.

A model almost perfectly predicting training data may be not good, or even useless for prediction!

Four plots that have the same stats



http://en.wikipedia.org/wiki/Anscombe%27s_quartet

Four plots that have the same stats

Property	Value
Mean of *x* in each case	9 (exact)
Variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

http://en.wikipedia.org/wiki/Anscombe%27s_quartet

Descriptor selection, chance correlations

- We can generate many descriptors for a given structure.
- To avoid overfitting, we need a small set of relevant descriptors.
- If small subsets of relevant descriptors or not chosen carefully, *chance correlations* can arise.
- Chance correlations are apparently reasonable models that can arise even when using random numbers as descriptors
- Good, statistically sound methods of ‘feature detection’ or descriptor selection have been developed.
- As mentioned before, sparse, less complex models generally have the best ability to predict the properties of new compounds.

Topliss JG, Edwards RP. J Med Chem. 22, 1238 (1979)

Chance correlations

Topliss showed that if you generate a large number of random numbers and select small subsets of these as descriptors for a QSAR model, there is a non trivial chance that a statistically reasonable-looking model can appear. This is called a chance correlation.

This can happen when choosing lots of subsets of descriptors from a large pool

Domain of applicability of models

- Models are trained using data within a range of property values and for a specified range of molecular types (the applicability domain of the model).
- Like any form of extrapolation, predicting properties of materials outside of the molecular or property space used to develop the model must be done with care.
- Molecular similarity based on the descriptors used to develop the model can be a guide to how far outside the model domain the prediction lies.
- Some probabilistic modelling methods such as Bayesian regularized neural network that generate a distribution of weights rather than a single set of weights can also be used to estimate the domain of applicability of the model.

Domain of applicability

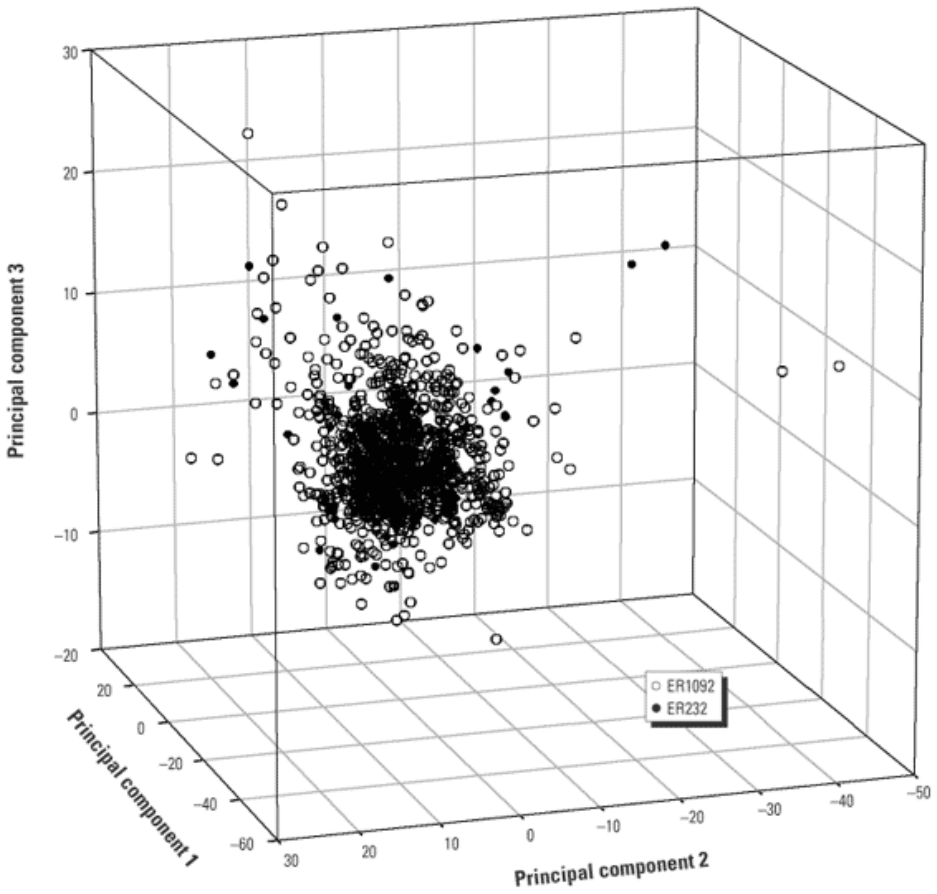
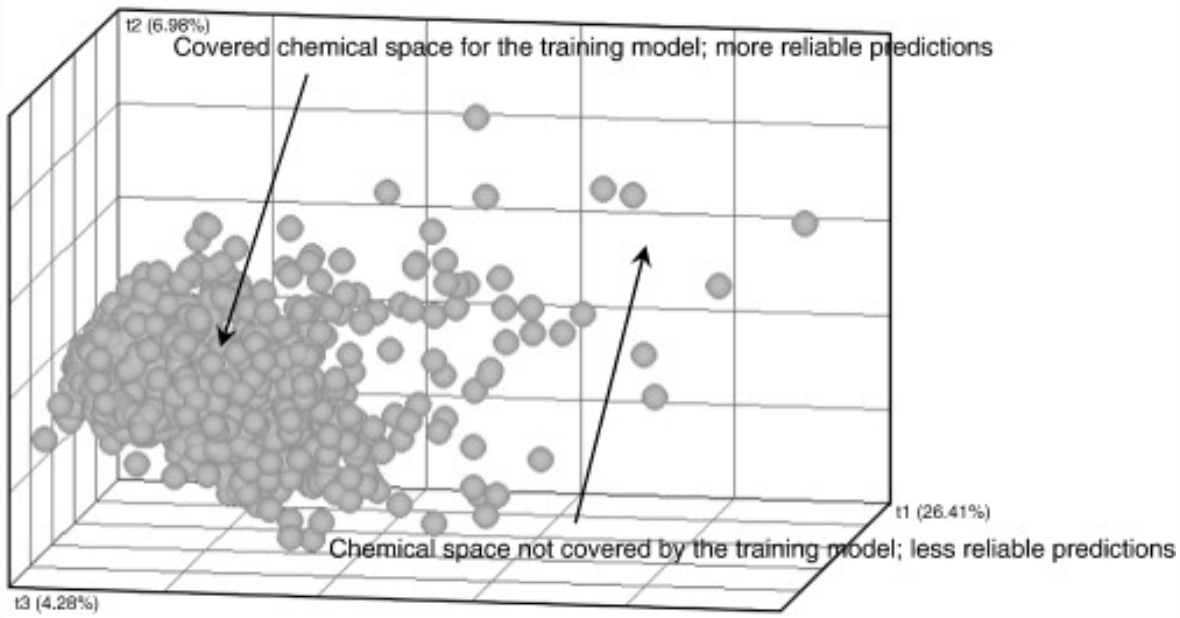


Figure 1. Comparison of structural diversity of ER232 and ER1092 in a chemistry space defined by three principal components of over 270 2D structural descriptors.

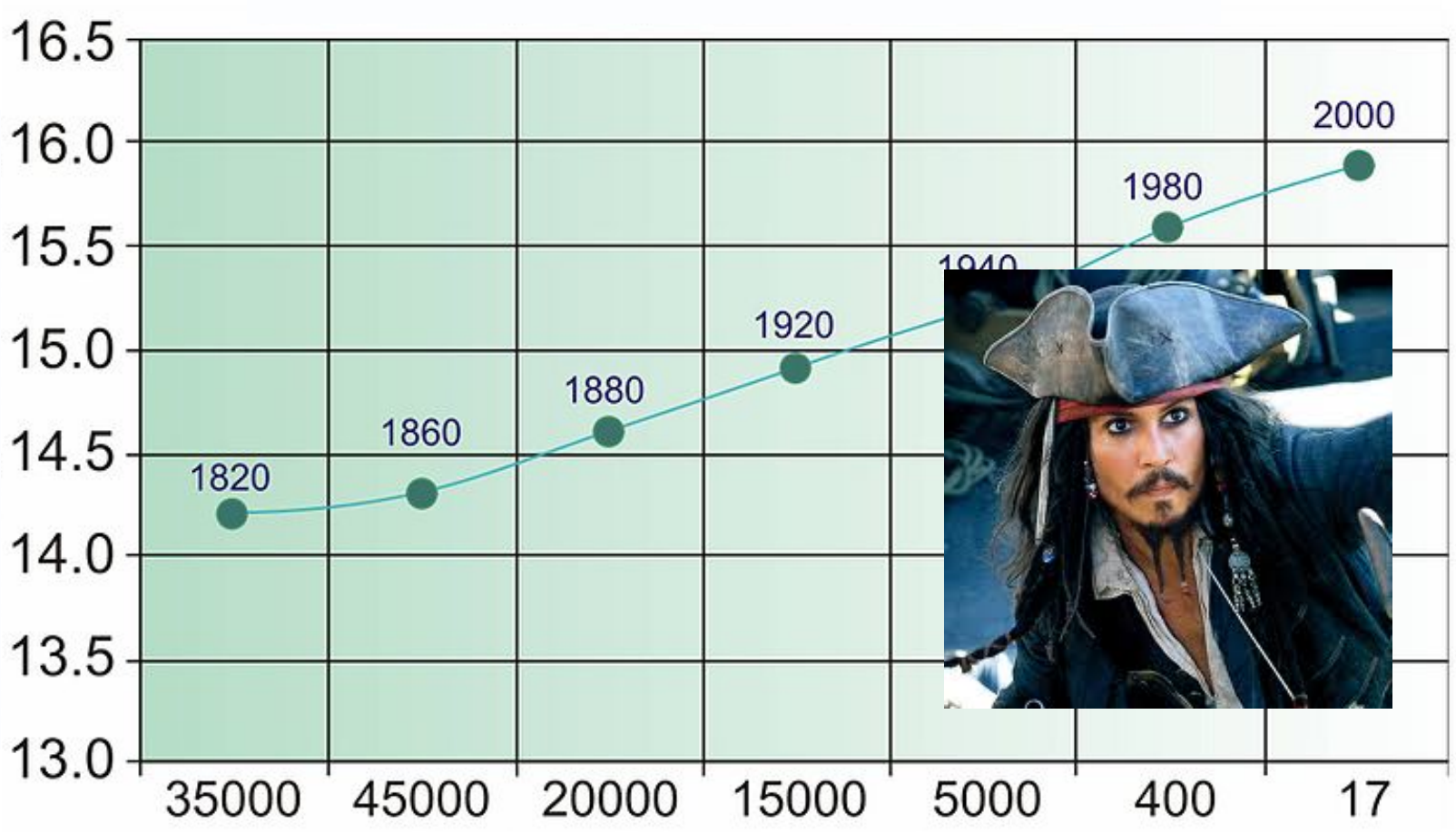
Tong et al. Environ. Health Persp. 112, 1249-1254

Domain of applicability



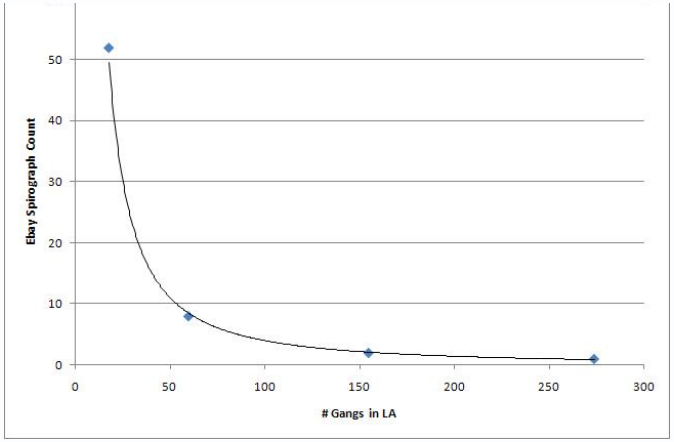
Valerio, Toxicol. Appl. Pharmacol. 241, 2009, 356

Beware of correlation vs. causation

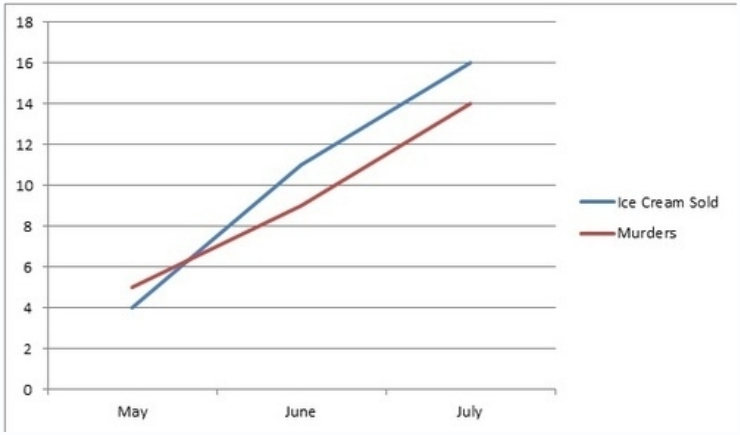


Other fun correlations

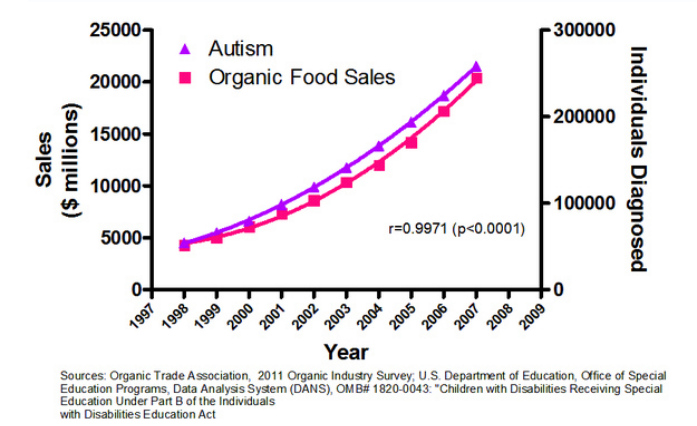
Not buying spirographs increases gang membership



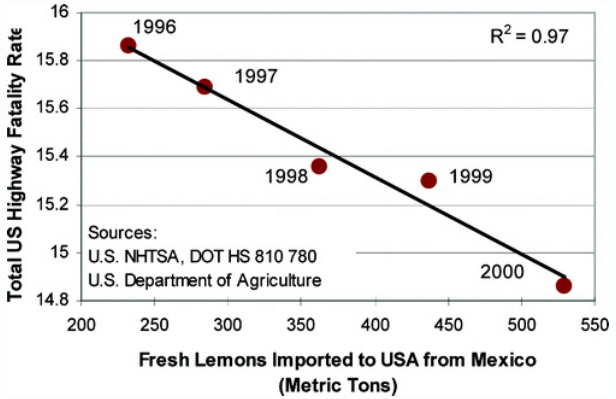
Buying ice cream causes murders



Organic food causes autism



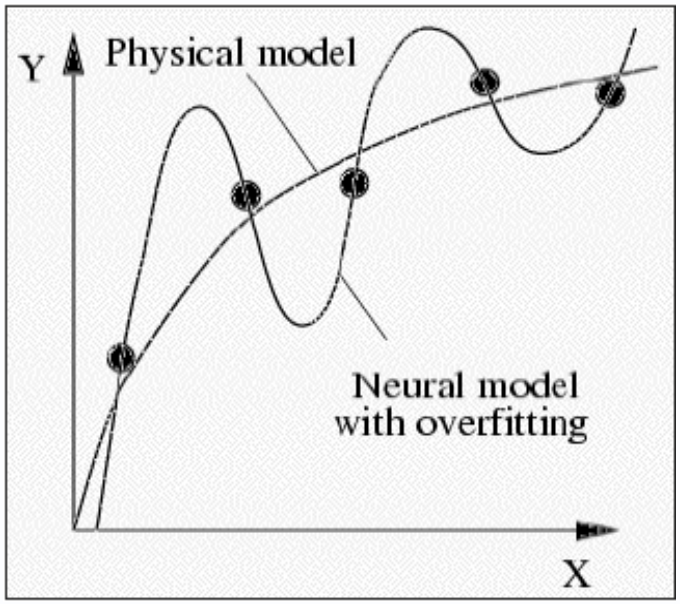
Importing lemons causes road deaths

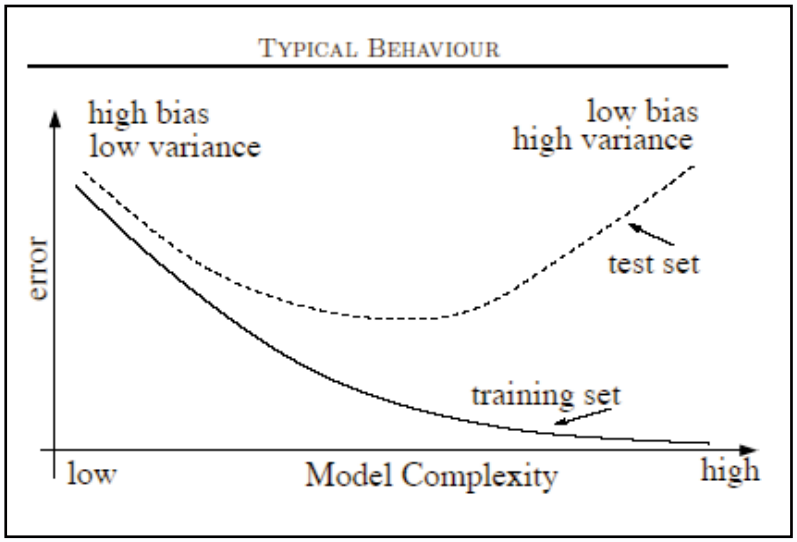


http://www.buzzfeed.com/kjh2110/the-10-most-bizarre-correlations

Overtraining neural networks

If neural networks are trained for too long they get worse at predicting new data





Effect of sparsity on generalization

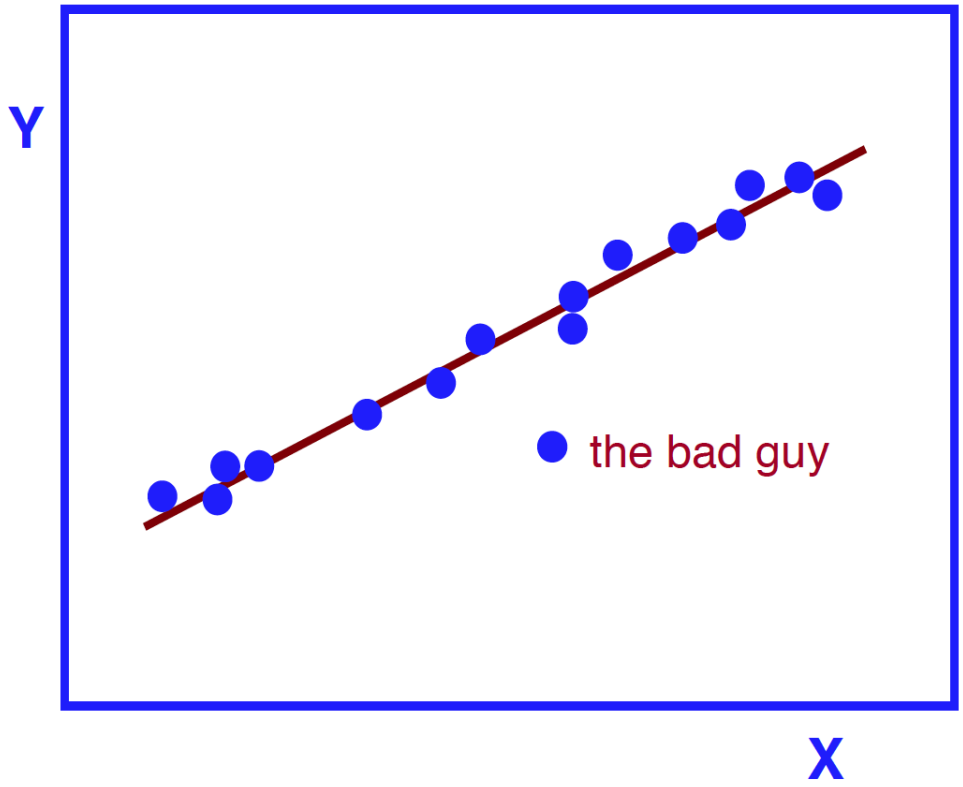
Sparsity	MLR	0	1	3	**5**	6	7
N_{descr}	53	53	53	49	**37**	24	9
SEE	**0.56**	**0.35**	**0.34**	**0.33**	**0.34**	**0.53**	**0.71**
r^2 train	0.77	0.89	0.89	0.89	**0.90**	0.75	0.60
SEP	**0.58**	**0.40**	**0.40**	**0.40**	**0.46**	**0.57**	**0.74**
r^2 test	0.76	0.87	0.87	0.85	**0.80**	0.72	0.57
$N_{eff\ wts}$	54	111	100	76	**54**	34	11

BRANNLP with 2 hidden layer nodes, scCO2 solubility data
Predictivity improves as sparsity increases, until model becomes too sparse

Dealing with outliers

Hugo Kubinyi, www.kubinyi.de

„Good“ and „Bad“ Guys in Regression Analysis



outlier in the test set:
r^2, Q^2 good
r^2_{pred} poor

outlier in the training set:
r^2, Q^2 poor
r^2_{pred} good

Purposes of QSPR Modelling

Interpretive modelling

- Often congeneric series of molecules (common core structure)
- Often smaller data sets that are more carefully measured
- Descriptors chosen to be interpretable
- Can help understand the molecular interactions of molecules at the biochemical target
- Define structures as targets for further synthesis and testing

Predictive modelling

- Often very large diverse data sets
- Data often noisy or incomplete
- Descriptors chosen for computational efficiency and predictive power
- Use the model to screen large databases of virtual molecules for those likely to be novel and active

Simple examples of QSPR

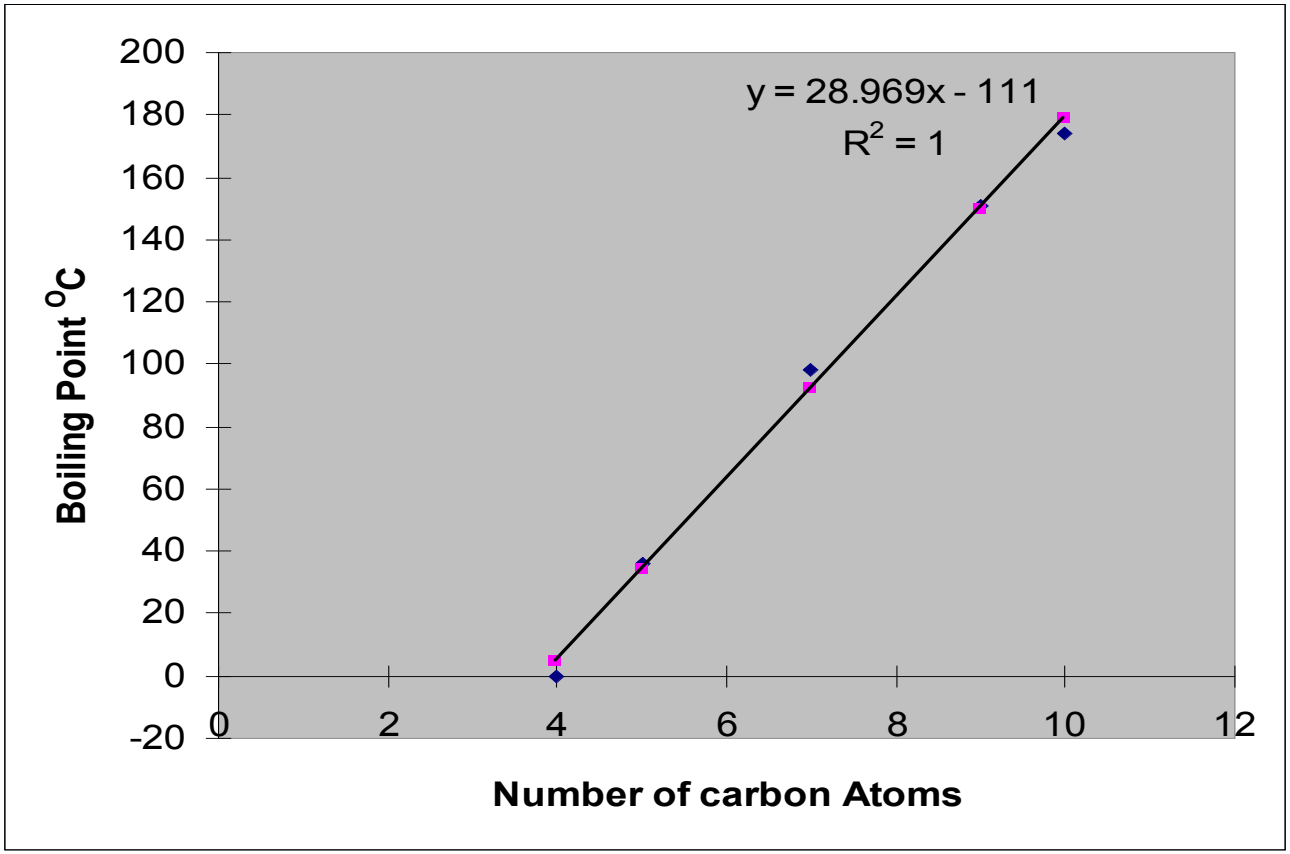
Structural measure	Property modelled
Molecular Mass	Boiling Points
Functional group	Solubility
Molecular Volume	Partition Coefficient
Number of atoms	Drug activity
Number of double bonds	Toxicology

A simple example: alkane boiling points

Training set of measured values

Butane	-0.5˚
Pentane	36.1˚
Heptane	98.4˚
Nonane	150.8˚
Decane	174.1˚

Regression using number of carbon atoms



Some alkane boiling points (^{O}C)

Alkane	Measured	Predicted	Error
Butane	-0.5	5	-5.5
Pentane	36.1	34	2.1
Hexane	68.7	63	5.7
2,2 – dimethylbutane	49.7	63	-13.3
Cyclohexane	80.7	63	17.7
Heptane	98.4	92	6.4
Octane	125.7	121.0	4.7
Nonane	150.8	150	0.8
Decane	174.1	179	-4.9

How well does it work? Examples

1. **MIO nanoparticle cellular uptake**
   (Shaw/Weissleder, Harvard)
2. Functionalized gold nanoparticle interaction with proteins (Yan, St Jude's/Shandong)

Selective uptake of surface modified MION

nature biotechnology

Cell-specific targeting of nanoparticles by multivalent attachment of small molecules

Ralph Weissleder[1], Kimberly Kelly[1,2], Eric Yi Sun[1,2], Timur Shtatland[1] & Lee Josephson[1]

Nanomaterials with precise biological functions have considerable potential for use in biomedical applications. Here we investigate whether multivalent attachment of small molecules can increase specific binding affinity and reveal new biological properties of such nanomaterials. We describe the parallel synthesis of a library comprising 146 nanoparticles decorated with different synthetic small molecules. Using fluorescent magnetic nanoparticles, we rapidly screened the library against different cell lines and discovered a series of nanoparticles with high specificity for endothelial cells, activated human macrophages or pancreatic cancer cells. Hits from the last-mentioned screen were shown to target pancreatic cancer *in vivo*. The method and described materials could facilitate development of functional nanomaterials for applications such as differentiating cell lines, detecting distinct cellular states and targeting specific cell types.

One of the emerging goals of nanotechnology is to functionalize inert and biocompatible materials to impart precise biological functions. Several novel materials have recently been described for diagnostic or therapeutic use[1–3], including quantum dots[4–6], polymers[7,8] and magnetofluorescent nanoparticles[9,10]. Considerable effort has been directed toward rational surface modifications and coatings to modulate pharmacokinetic properties (e.g., blood half-life, elimination and biodegradation), toxicity, immunogenicity and efficient targeting. Targeting has generally been achieved by conjugating nanoparticle surfaces to antibodies. Although this approach has succeeded for *in vitro* sensing[11,12], its *in vivo* application has proved more challenging because of cost, lim… potential immunogenicity after …ations[13]. Another targeting app… involves nanomaterials that discriminate among distinct cell types, or among different physiological states of a given cell type.

RESULTS

Synthesis of nanoparticle library

The first step towards creation of the nanoparticle library was to identify biologically and chemically suitable nanoparticles that could be detected by magnetic and fluorescent means and could be chemically modified. We used magnetofluorescent nanoparticles[9,10] as starting material because such preparations can be made with high ($R2 > 30\ mMsec^{-1}$) magnetic relaxivity, because related materials are biocompatible and in clinical use[16], and because aminated base materials facilitate conjugation of small molecules through sulfhydryl, carboxyl, amine and anhydride chemistries (**Fig. 1e**).

Ralph Weissleder, Harvard,

NANO LETTERS

Letter

pubs.acs.org/NanoLett

Modeling Biological Activities of Nanoparticles

V. Chandana Epa,[†] Frank R. Burden,[⊥] Carlos Tassa,[§] Ralph Weissleder,[§] Stanley Shaw,[§,∥] and David A. Winkler*[,⊥,‡]

[†]CSIRO Materials Science and Engineering, 343 Royal Parade, Parkville, Victoria 3052, Australia

[⊥]CSIRO Materials Science and Engineering, Bayview Avenue, Clayton, Victoria 3168, Australia

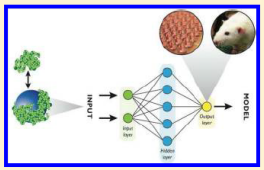
[‡]Monash Institute of Pharmaceutical Sciences, 381 Royal Parade, Parkville 30152, Australia

[§]Center for Systems Biology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States

[∥]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States

Supporting Information

ABSTRACT: Products are increasingly incorporating nanomaterials, but we have a poor understanding of their adverse effects. To assess risk, regulatory authorities need more experimental testing of nanoparticles. Computational models play a complementary role in allowing rapid prediction of potential toxicities of new and modified nanomaterials. We generated quantitative, predictive models of cellular uptake and apoptosis induced by nanoparticles for several cell types. We illustrate the potential of computational methods to make a contribution to nanosafety.



KEYWORDS: *Nanoparticle toxicity, model prediction, apoptosis, cellular uptake, Bayesian methods*

Many products are now exploiting the novel properties of nanomaterials, but their potential harmful effects are incompletely understood, a critical issue for regulatory authorities. Experimental testing of all potential nanomaterials is impractical; computational approaches such as machine learning methods can help assess potential risk of new and modified nanomaterials and prioritize nanomaterials for experimental testing. Puzyn et al.[1] and Fourches et al.[2] recently reported models of nanoparticle properties that demonstrated proof of concept for this approach. Here we report the use quantitative structure–activity relationship (QSAR) methods Methods). These data were used for our first computational modeling study. Of the possible combinations of biological assays and cell types, only the apoptosis assays exhibited a dose–response relationship. Of these, only the smooth muscle cell apoptosis assay generated statistical significant models. We initially investigated the dependence of the apoptosis response on the relaxivities (R1 and R2) and the zeta potential (available for 32 of the nanoparticles). We found a very significant relationship between the relaxivity R1 and the apoptosis assay results. However, as the relaxivities correlated almost completely with the type of iron oxide core, it is very likely

MIO nanoparticle cellular uptake

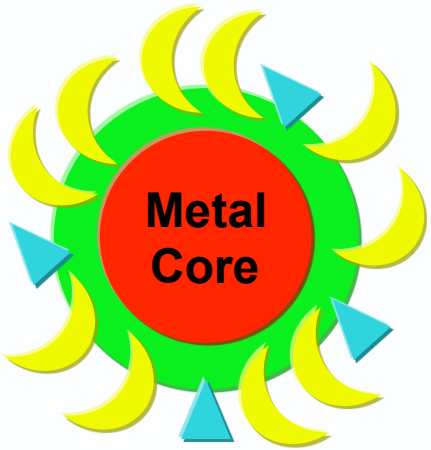
Weissleder et al. investigated whether the multivalent attachment of small organic molecules on a same NP can modify its binding affinity to certain cells.

- PaCa2: Pancreatic cancer cell
- HUVEC: human umbilical vein endothelial cell
- U937: Macrophage cell line
- GMCSF: Activated primary human macrophages
- RestMph: Resting primary human macrophages

109 nanoparticles with same core (CLIO) but different surface chemistries

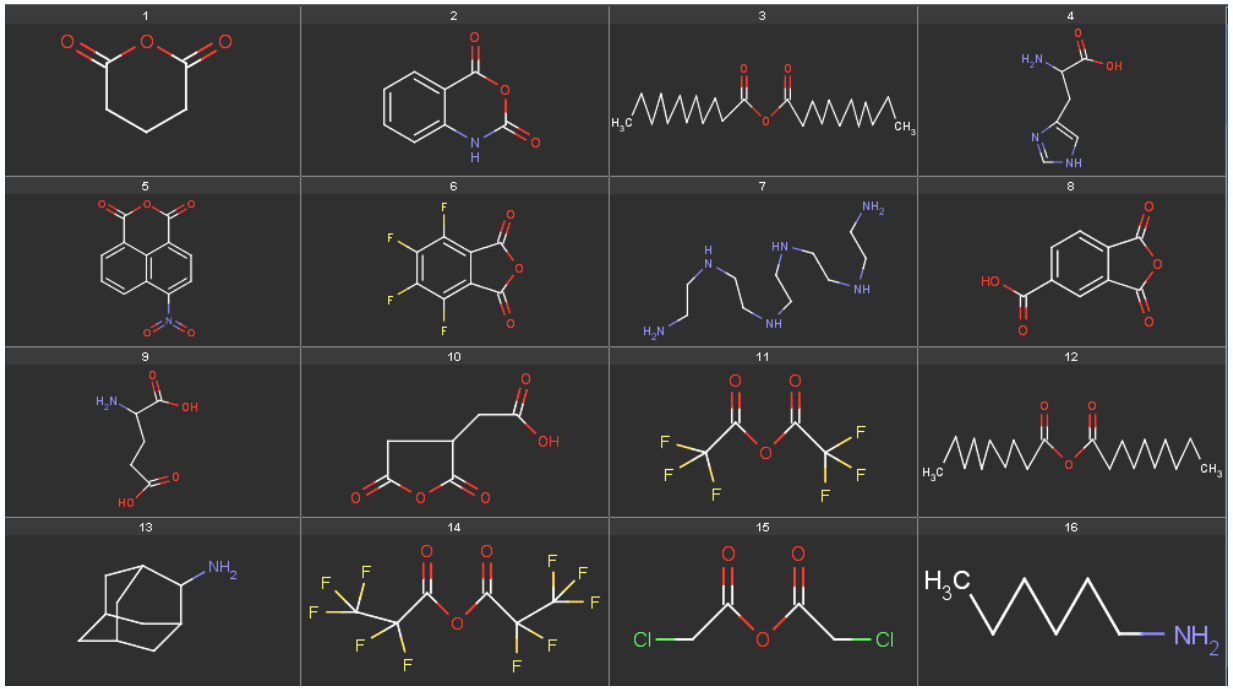
Weissleder et al. Nat. Biotechnol., 2005, 23, 1418

CLIO – cross-linked iron oxide core



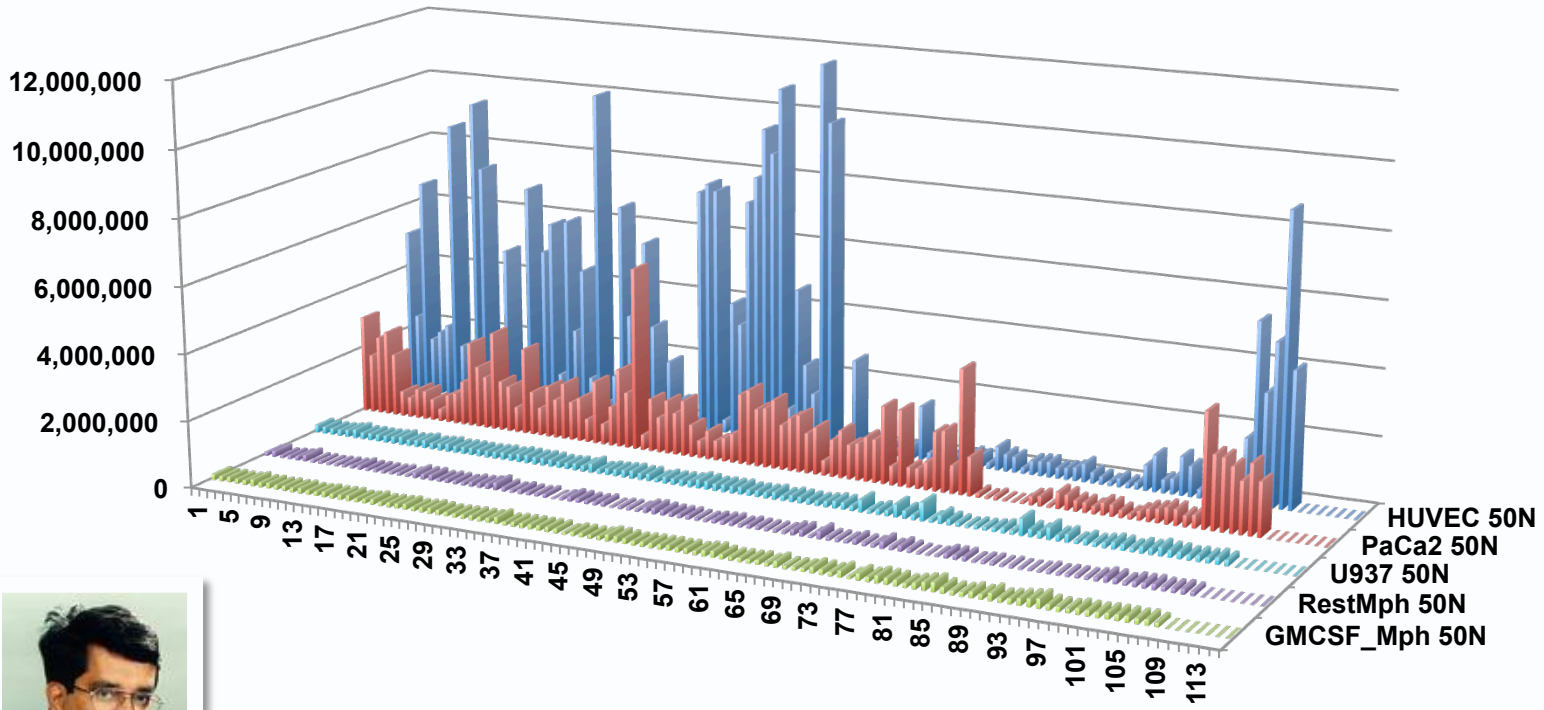
FITC (fluorescein isothiocyanate)

Small organic compounds



Slide adapted from Tropsha, UNC

Look at the data! Cellular uptake



Vidana Epa

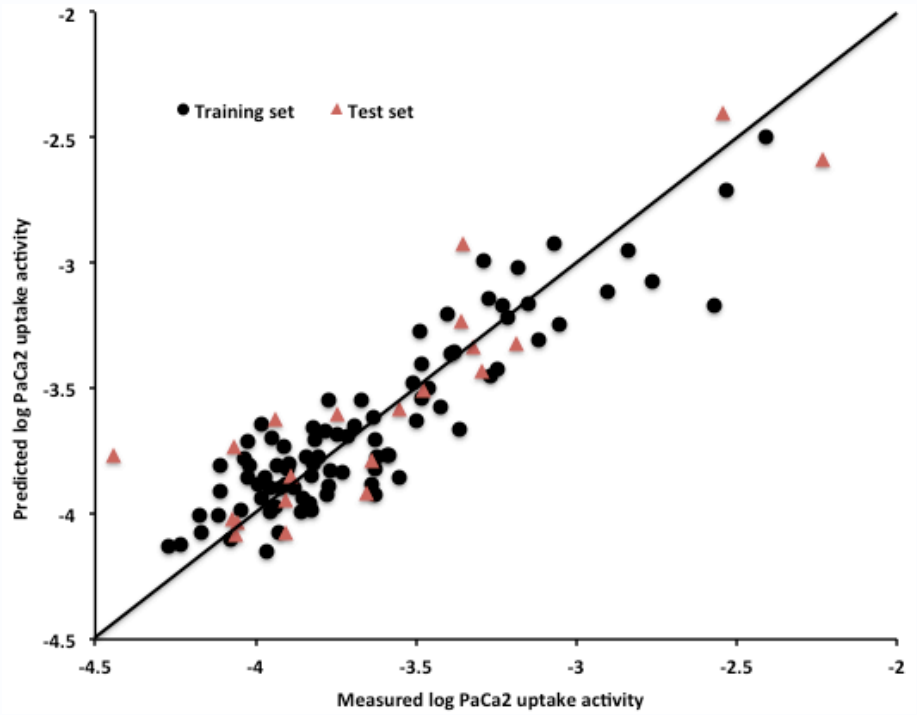
QNTR models of nanoparticle uptake

Cell Type	Model	Descriptors	r^2	SEE(scaled)	q^2	SEP(scaled)
HUVEC	MLREM	11	0.74	0.13	0.63	0.14
	BRANNGP	11	0.70	0.11	0.66	0.13
PaCa2	MLREM	19	0.76	0.10	0.79	0.13
	BRANNGP	19	0.77	0.07	0.54	0.14
U937	MLREM	7	0.42	0.11	0.25	0.14
GMCSF_Mph	MLREM	15	0.59	0.10	0.02	0.44
RestMph	MLREM	16	0.43	0.13	0.001	0.43

Only two cell types have uptake that is sensitive to the surface chemistry. The macrophages and macro-phage-like cell lines do not take up nanoparticles in a manner that is modulated by the surface functionalization. As these are 'universal phagocytes' perhaps this is not unexpected.

Nanoparticle cellular uptake

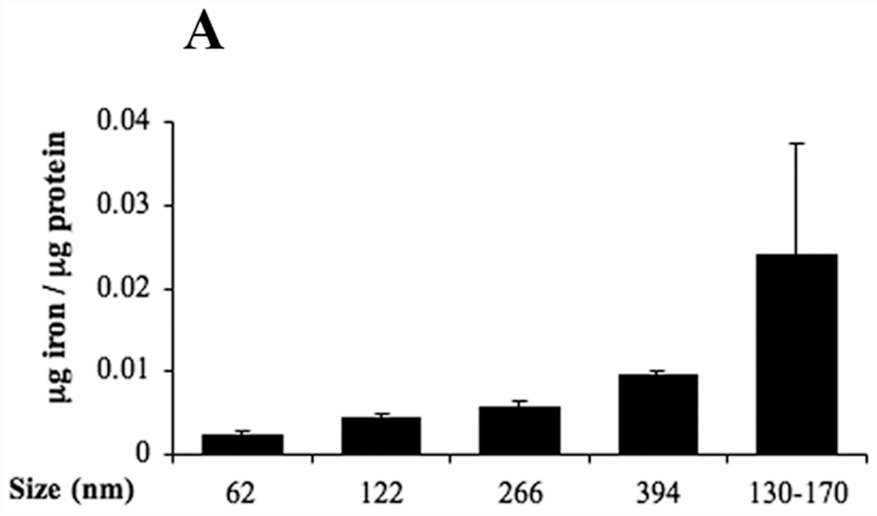
Performance of PaCa2 model for training set (black dots) and test set (red triangles). Each point represents a different surface-modified nanoparticle. r_{train}^2=0.77, r_{test}^2=0.79, SEE= 0.19, SEP=0.24 (logs)

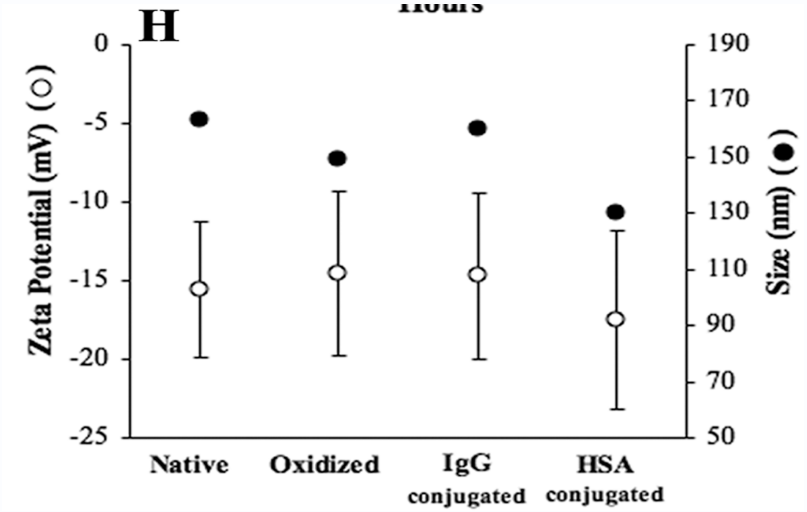


DRAGON and in-house ABC descriptors, sparse feature selection

Epa et al. Nano Lett 2012

Uptake on CLIO nanoparticles by macrophages





Uptake increases exponentially with nanoparticle size. Zeta potential in biological fluids is usually small and negative.

Beduneau A, Ma Z, Grotepas CB, Kabanov A, Rabinow BE, et al. 2009 Facilitated Monocyte-Macrophage Uptake and Tissue Distribution of Superparmagnetic Iron-Oxide Nanoparticles. PLoS ONE 4(2): e4343. doi: 10.1371/journal.pone.0004343

Uptake on CLIO nanoparticles by PaCa cells

Many types of nanoparticles are designed primarily to image tumours by preferential accumulation or cell specific targeting. Higher uptake by PaCa cells is therefore not unexpected.

Swiss Med Wkly. 2010;140:w13081

How well does it work? Examples

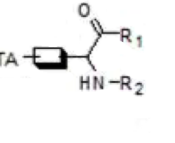
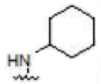
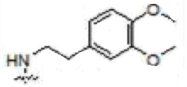
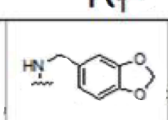
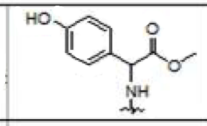
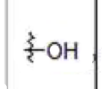
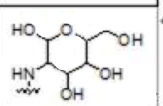
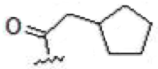
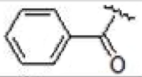
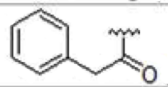
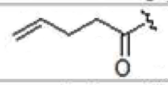
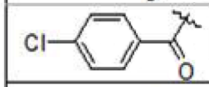
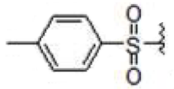
1. MIO nanoparticle cellular uptake (Shaw/Weissleder, Harvard)
2. **Functionalized gold nanoparticle protein interaction** (Yan, St Jude's/Shandong)

Protein binding to modified gold nanoparticles

- Acetylcholinesterase (AChE)
- Nonspecific protein binding to functionalized gold nanoparticles (f-GNPs)

Bing Yan St Jude's Hospital, Memphis now at Shandong University

Protein binding to f-GNPs

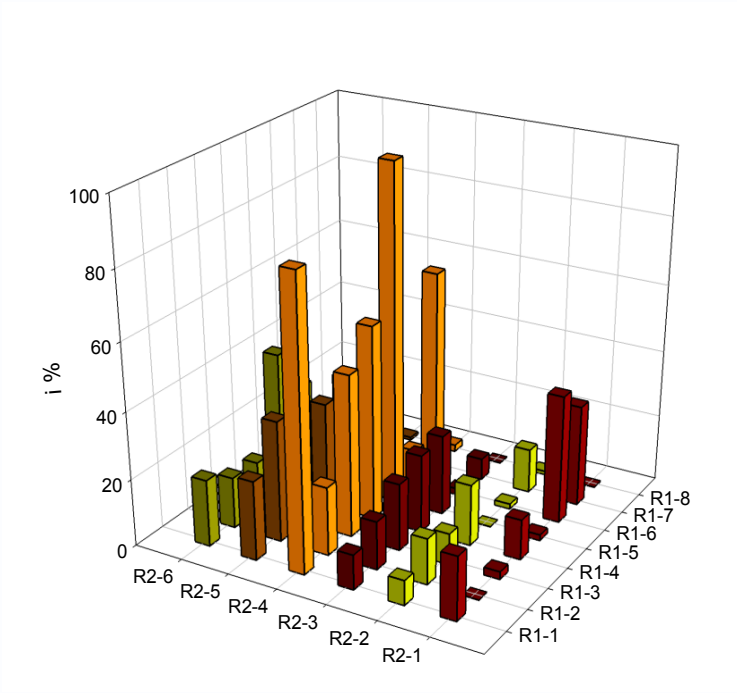
		R_1-							
									
R_2-		1	2	3	4	5	6	7	M1S
		8	9	10	11	12	13	14	M2S
		15	16	17	18	19	20	21	M3S
		22	23	24	25	26	27	28	M4S
		29	30	31	32	33	34	35	M5S
		36	37	38	39	40	41	42	

Bing Yan
Shandong
University

Surfaces of multiwall carbon nanotubes were chemically modified for tissue targeting. 47 GNPs with different surface chemistries

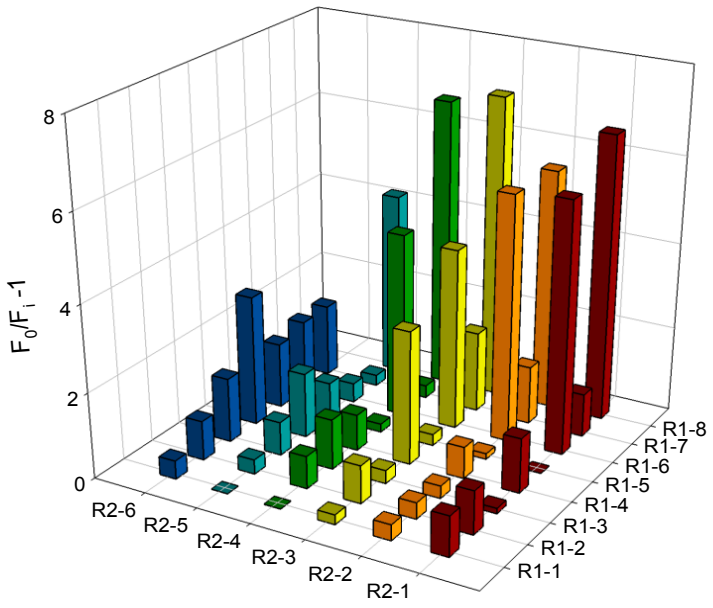
Look at the data!

Activity Inhibition Result



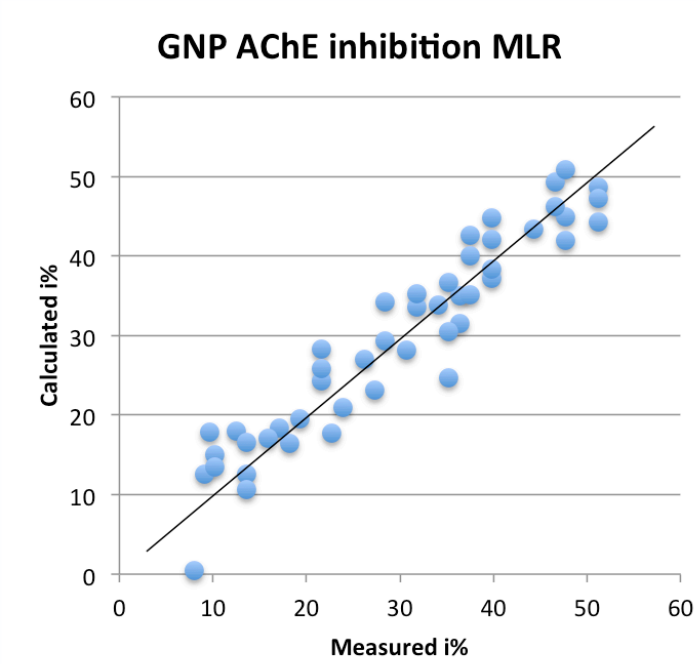
The inhibition of AChE's activity by GNPs. Experimental conditions: final conc. of Au was 3 μg/mL and the AChE was 0.3 nM .

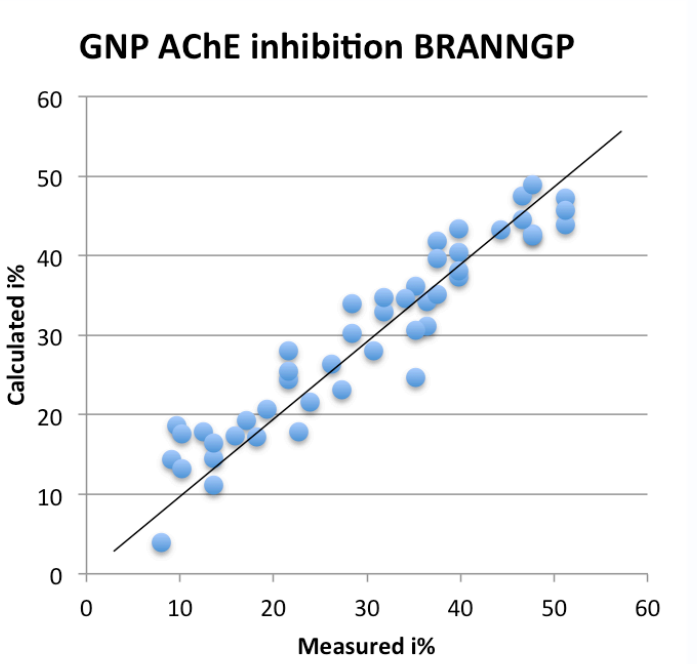
Fluorescence Quenching Result



The fluorescence quenching of AChE by GNPs. Experimental conditions: AChE 1.67 μM in 0.1M PBS (pH~8.0), final conc. of Au in AChE solution was 17 ppm.

AChE inhibition by f-GNPs – no test set

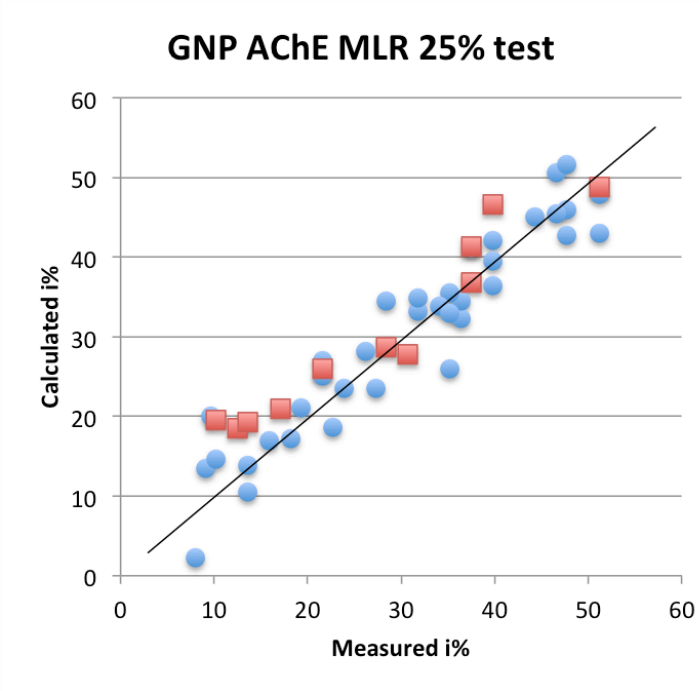


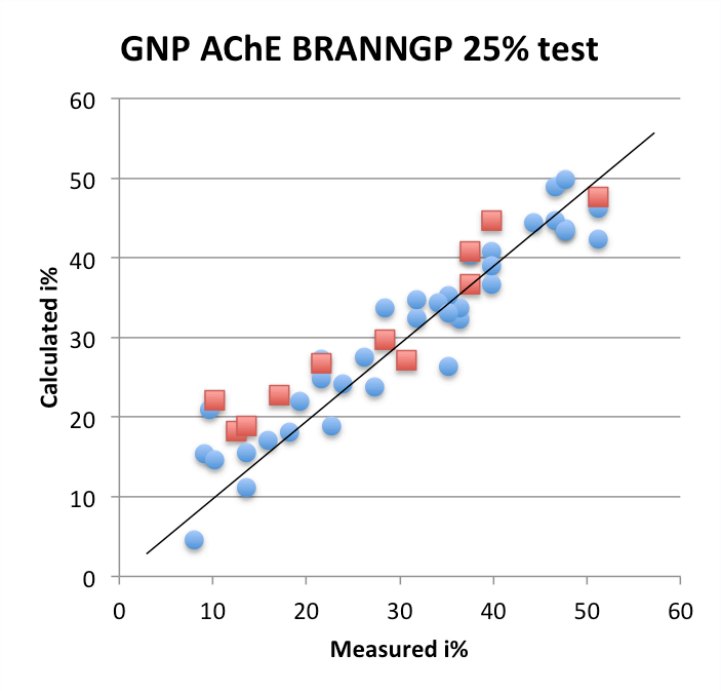


Linear and machine learning models could predict the effects of chemical modification of GNP surface with good accuracy.

Zhou et al, Nano Lett. 3, 859 (2008)

AChE inhibition by f-GNPs – 25% test set





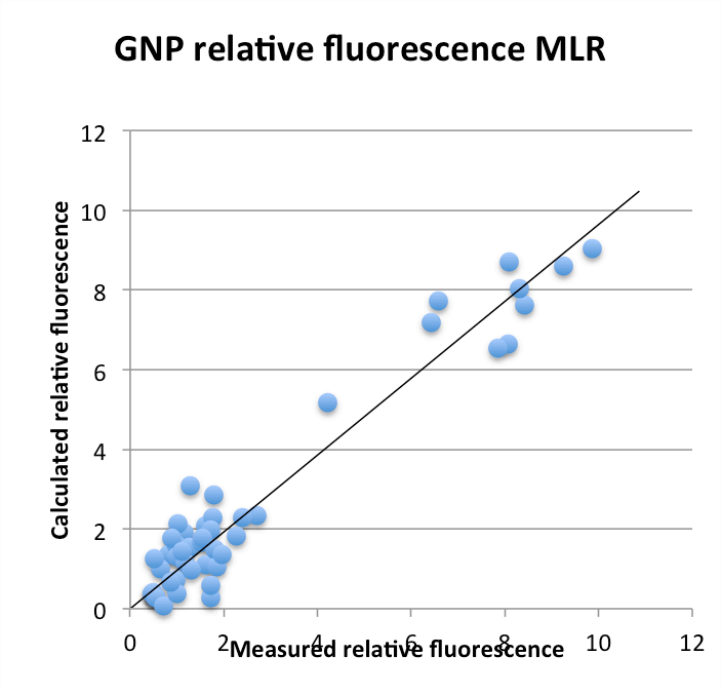
Linear and machine learning models could predict the effects of chemical modification of GNP surface with good accuracy.

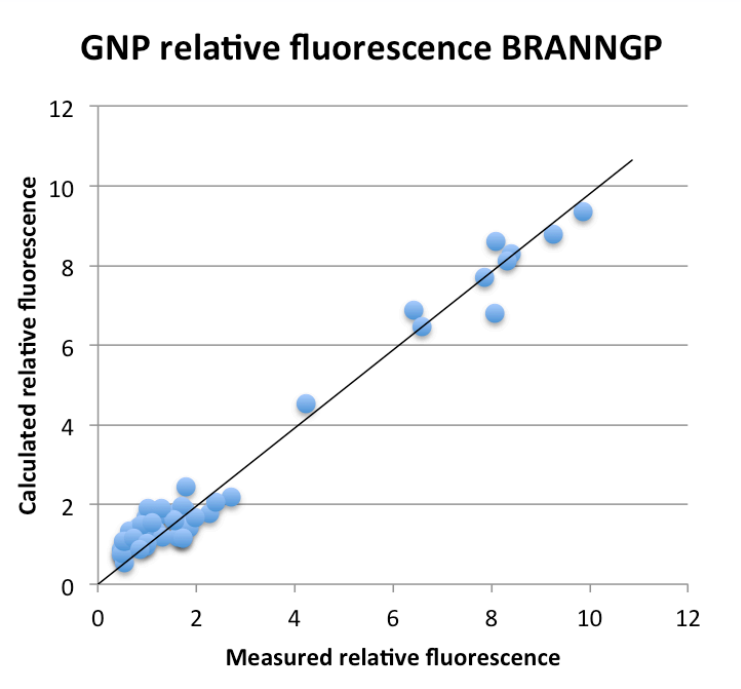
Zhou et al, Nano Lett. 3, 859 (2008)

AChE binding to f-GNPs - AChE inhibition

Model	β	N_{descr}	R^2 train	SEE	R^2 test	SEP	N_{eff}
MLREM	1.3	21	0.97	2.2			22
	1.5	17	0.91	3.4			18
	1.7	14	0.90	3.6			15
	1.7	14	0.90	3.6			15
	1.8	6	0.69	6.5			7
	2	6	0.58	7.4			7
MLR	0	14	0.90	4.9			15
BRANNGP 2nodes		14	0.85	3.8			14
BRANNGP 3 nodes		14	0.77	4			15
BRANNLP 2 nodes		14	0.90	4			33
MLR 20% test		14	0.91	5.1	0.81	5.6	15
BRANNGP 2 nodes 20% test		14	0.84	3.7	0.8	5.2	14
BRANNGP 3 nodes 20% test		14	0.87	3.7	0.81	5.2	15
BRANNLP 2 nodes 20% test		14	0.89	3.6	0.77	5.5	16
MLR 25% test		14	0.90	5.3	0.93	4.9	15
BRANNGP 2 nodes 25% test		14	0.85	3.9	0.91	5.1	14
BRANNGP 3 nodes 25% test		14	0.85	3.9	0.91	5.1	14
BRANNLP 2 nodes 25% test		14	0.88	3.7	0.87	5.5	16

Nonspecific protein binding by f-GNPs

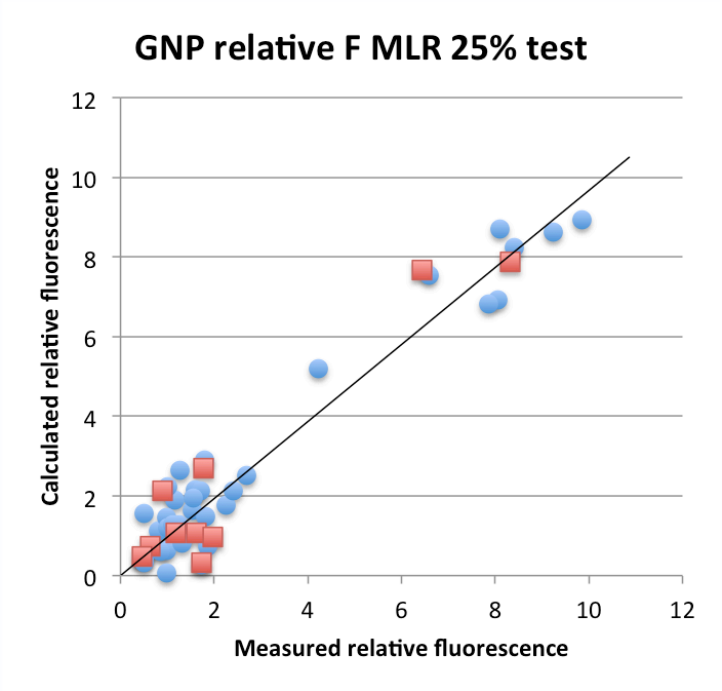


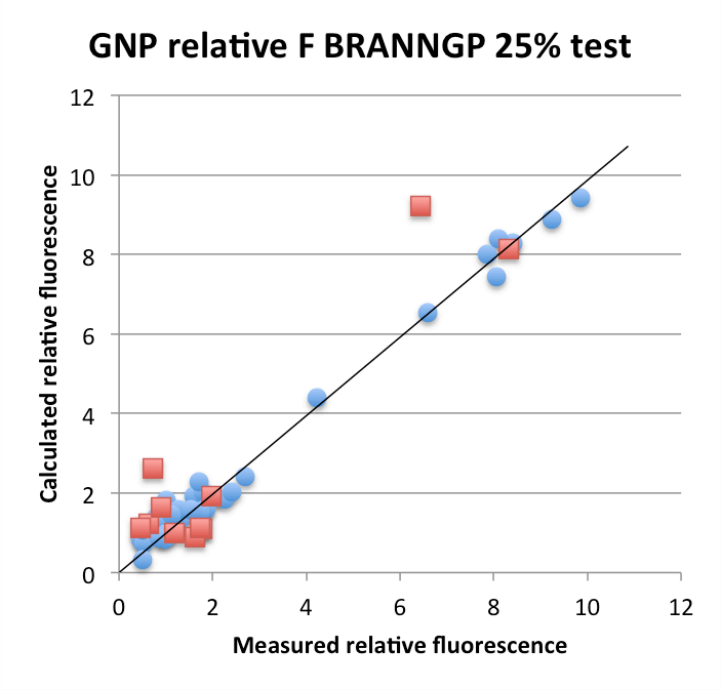


Model trained using all data in training set and no test set. Sparse linear and nonlinear models using DRAGON descriptors for surface chemistries.

Zhou et al, Nano Lett. 3, 859 (2008)

AChE inhibition by f-GNPs – 25% test set





Model trained using 75% of data in training set and 25% in test set. Sparse linear and nonlinear models using DRAGON descriptors for surface chemistries.

Zhou et al, Nano Lett. 3, 859 (2008)

Nonspecific protein binding to f-GNPs

Model	β	N_{descr}	R^2 train	SEE	R^2 test	SEP	N_{eff}
MLREM	6	11	0.91	0.84			12
MLR	0	10	0.93	0.82			11
BRANNGP 2nodes		10	0.96	0.42			16
BRANNGP 3 nodes		10	0.97	0.44			16
BRANNLP 2 nodes		10	0.96	0.50			15
MLR 20% test		10	0.93	0.88	0.93	0.75	11
BRANNGP 2 nodes 20% test		10	0.96	0.43	0.94	0.75	15
BRANNGP 3 nodes 20% test		10	0.97	0.39	0.91	0.95	16
BRANNLP 2 nodes 20% test		10	0.98	0.36	0.93	0.88	20
MLR 25% test		10	0.94	0.86	0.9	0.87	11
BRANNGP 2 nodes 25% test		10	0.98	0.31	0.86	1.12	18
BRANNGP 3 nodes 25% test		10	0.98	0.31	0.88	1.04	14
BRANNLP 2 nodes 25% test		10	0.97	0.46	0.91	0.84	12

Take home messages

- Modelling is hard because materials are complex.
- QSAR/QNTR is a simple method that can be very useful when used carefully. In the hands of a skilled practitioner it can yield very good results
- Models are easy to build but also very easy to get wrong. Many published QSAR studies have serious errors
- Data quality, quantity, diversity, range, relevance are paramount
- QSAR methods can capture complex relationships between structure and biological activity, even for multiple modes of action
- Descriptor generation and selection is the key step in QNTR
- New mathematical and machine learning methods have made model building more robust.
- The methods are very fast and can deal with very large data sets.

What can QSAR/QNTR not do?

- Replace the need for experimental measurement. Models are synergistic with measurements.
- Generate good predictive models without understanding modelling process and without remaining skeptical until models are validated.
- Build predictive models with very small data sets, poor quality data.
- Generate good models with bad descriptors or data sets with low diversity, or low dynamic range of biological activities.
- Make reliable predictions that are well outside the property space in which they are trained.
- Convince regulators and other science professionals that they are useful unless their predictivity is tested experimentally
- Molecular details of the mechanism of action are often not accessible from the model.

Acknowledgements

Our collaborators: Stan Shaw (MGH), Ralph Weissleder (Harvard), Bing Yang (St Judes/Shandong)

We would like to gratefully acknowledge support from:

- European COST office
- National Enabling Technologies Scheme
- Australian Stem Cell Centre (research grant).
- CSIRO Advanced Materials Platform.
- Chemical Structure Association (Dubois Award)
- CSIRO Newton Turner award for Exceptional Senior Scientists

Recent relevant publications

- Robust QSAR Models Using Bayesian Regularized Artificial Neural Networks, Burden FR, Winkler, DA, *J. Med. Chem*., **42**, 3183-3187 (1999).
- An optimal self-pruning neural network that performs nonlinear descriptor selection for QSAR, Burden, FR, Winkler, DA, *QSAR Comb. Sci*. **28**, 1092 – 1097 (2009).
- Optimum QSAR Feature Selection using Sparse Bayesian Methods, Burden, FR, Winkler DA, *QSAR Comb Sci*. **28**, 645-653, (2009).
- Modelling biological activities of nanoparticles. Epa, VC, Burden, FR, Tassa, C, Weissleder, R, Shaw, S, Winkler, DA *Nano Lett*., **12**, 5808−5812 (2012).
- Computational nanotoxicology, Epa VC, Winkler DA, Tran L, In *Adverse Effects of Engineered Nanoparticles*, Fadeel, Pietroiusti, and Shvedova (Eds.), Elsevier, Berlin 2011.
- *In silico* strategies for safe management of manufactured nanomaterials, Winkler DA, Mombelli E, Pietroiusti A, Tran L, Worth A, Fadeel B, McCall MJ, *Toxicol. (2012) ASAP*.
- **Towards predictive modelling of diverse materials properties, TC Le, VC Epa, FR Burden, DA Winkler. *Chem. Rev*. 112 (5), 2889–2919 (2012).**

Hands-on exercises

We will build QSPR models of a simple nanomaterials data set, uptake of metal ion oxide nanoparticles with chemically modified surfaces, using a simple modelling tool, KNIME.

We will compare the results of using a simple multiple linear regression models with nonlinear models generated by a polynomial regression, and a backpropagation neural network.

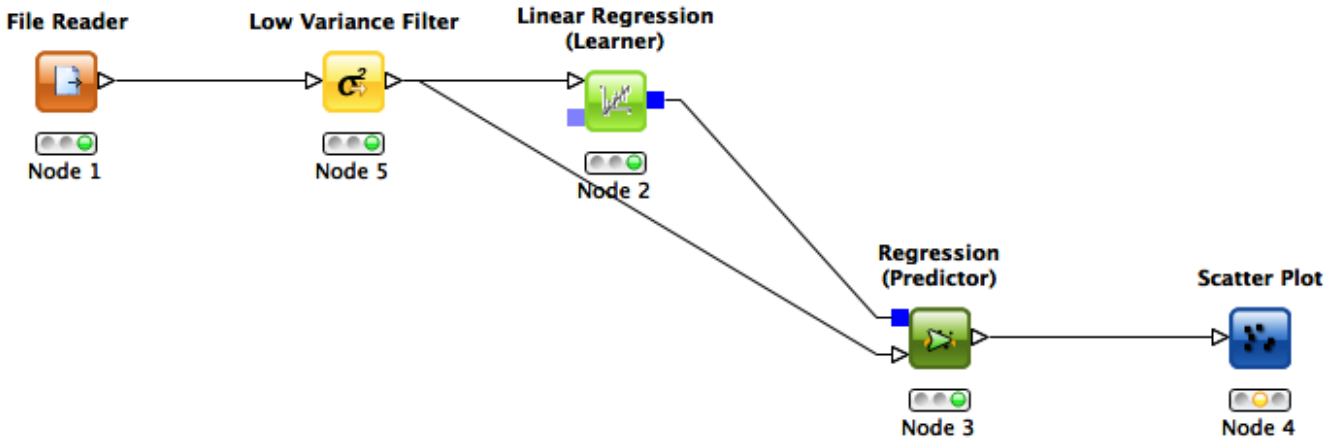
The software, workflows, key papers, and data are on your memory stick or have been downloaded from the MODENA web site or elsewhere.

You should have installed and tested your software prior to the working session

KNIME

KNIME is an object-oriented programming method (www.knime.org).

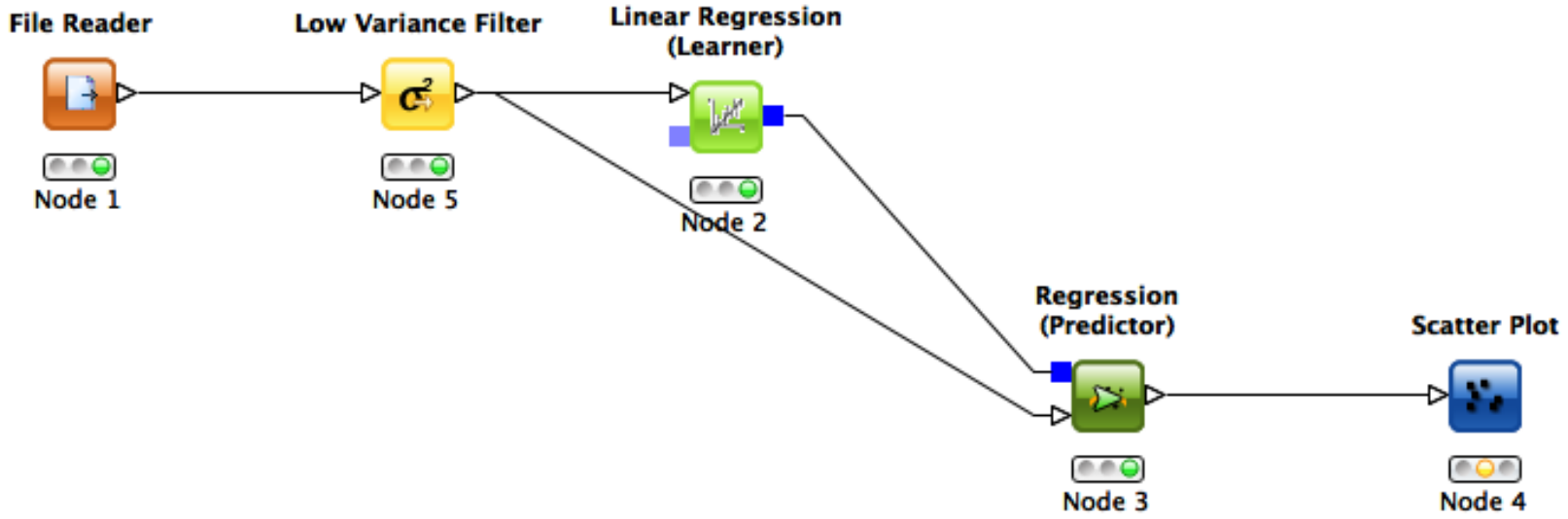
The workspace allows the placement and connection of different modules that perform specific functions.



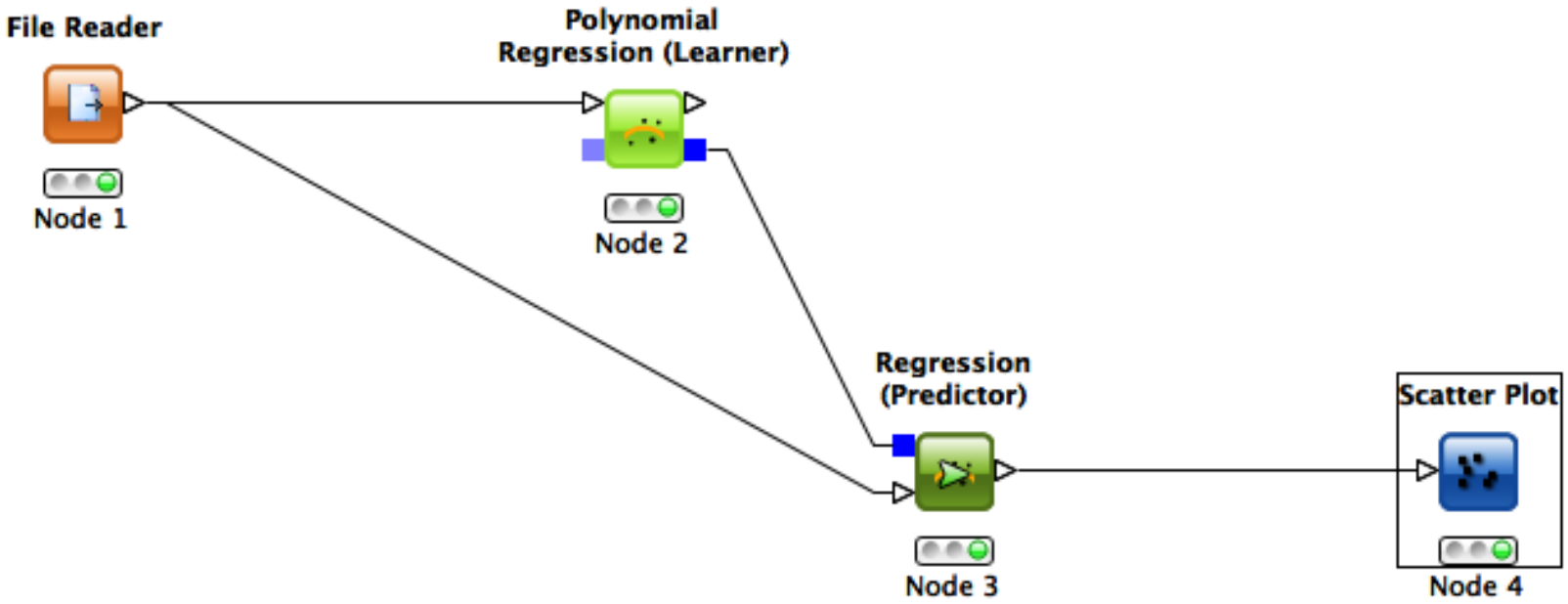
Example

- The data are from Weissleder et al. *Nature Biotech*. 2005, **23**, 1418.
- They provided data on the uptake of surface functionalized metal iron oxide nanoparticles (MION) by five types of cells: human umbilical vein endothelial cells (HUVEC), primary resting human macrophages, granulocyte macrophage colony stimulating factor (GM-CSF)–stimulated human macrophages, a U937 human macrophage-like cell line and human pancreatic ductal adenocarcinoma cells.
- We modelled these data using Bayesian regularized neural network and found the QSPR to be quite nonlinear. Only three of the cell types were found to take up the functionalized nanparticles to a significant extend which depended on the surface chemistry.

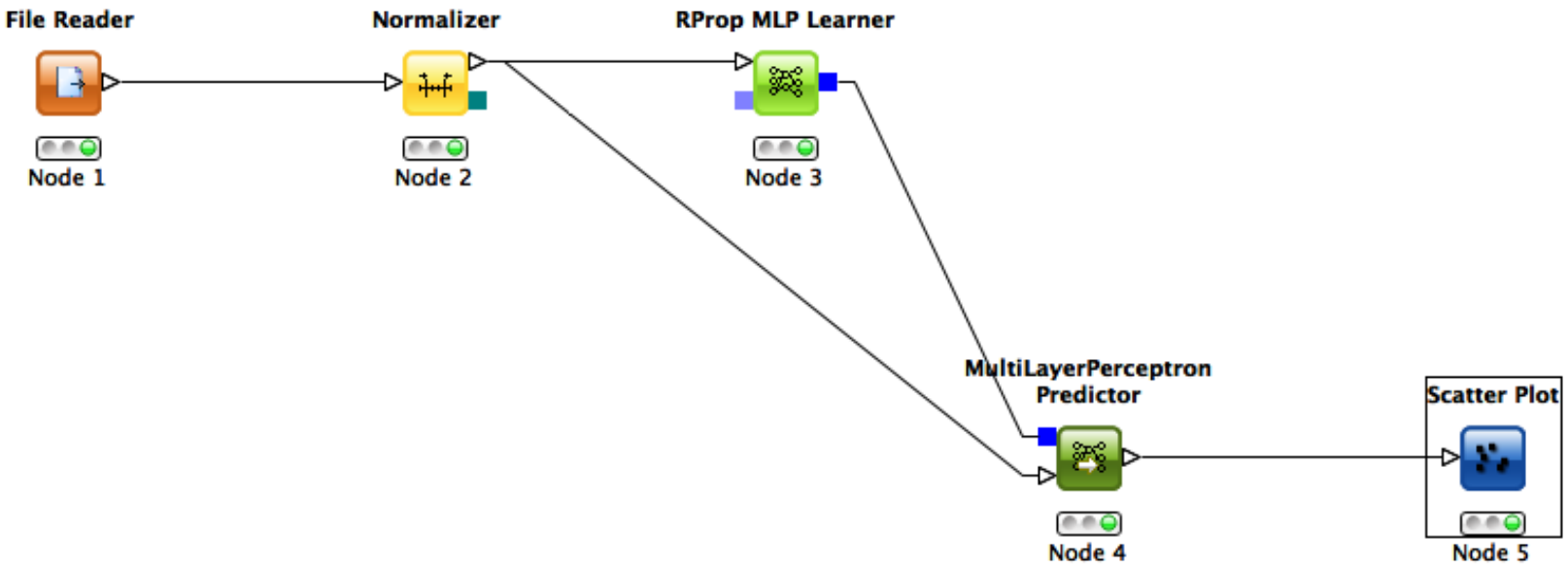
KNIME linear model workflow



KNIME polynomial model workflow



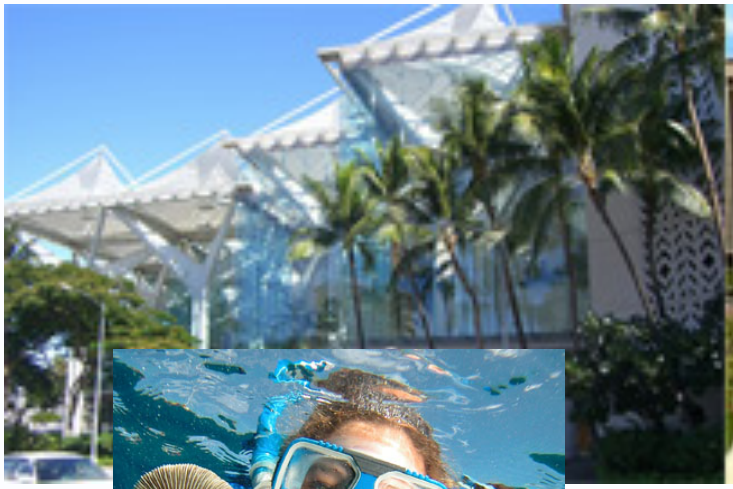
KNIME neural network workflow



Pacifichem 2015, 15-20 December, Waikiki, Hawaii
World's largest international chemistry congress

>300 cutting-edge symposia, many on materials, modelling, nanotechnology

http://www.pacifichem.org



Thank you

Prof. Dave Winkler
Monash Institute of Pharmaceutical Sciences

t +61 3 9545 2477
e dave.winkler@csiro.au@csiro.au
w www.csiro.au/CMSE

CSIRO Materials Science & Engineering
Modelling Research team Leader

twitter @drdavewinkler
Skype drdavewinkler