Integrating Data and Modelling

Haralambos Sarimveis,

National Technical University of Athens

eNanoMapper

Challenges

- **Model development**
  - Modelling services should have easy access to available data
  - Integrating and linking heterogeneous data from diverse resources and formulating them for modelling purposes
  - Characterisation of nanoparticles (physico-chemical, biological identity-nano/bio interactions)
  - Preprocessing and analysing raw data (for example images, omics data, spectral information)
  - Integration with theoretical means of describing nanoparticles (quantum mechanical descriptors)
  - Use state-of-the-art machine learning and methods and statistical analysis to produced accurate, well validated models with definition of the domain of applicability
  - Understanding of mechanisms of actions/ pathway analysis

- **Serving the community**
  - Easy access of the community to data, modelling tools and well-validated (published) public models (model repository)
  - Provide means of collaboration among modellers and experimentalists (optimal experimental design, inter-lab testing)
  - Cross-platform transform and transfer of produced models
  - Annotation of models and produced results using an ontology

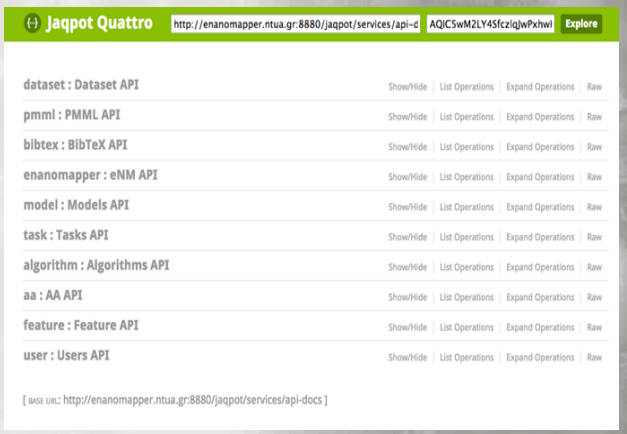
eNanoMapper

eNanoMapper computational infrastructure

- **OpenTox API Adjustments and Extensions**

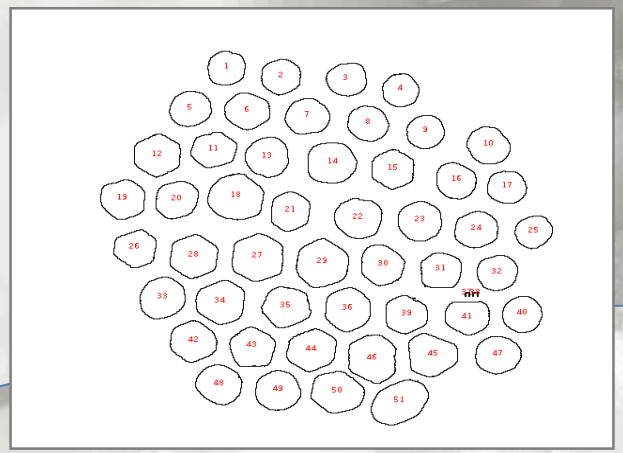
(documented through swagger, http://enanomapper.ntua.gr:8080/jaqpot/swagger/)

  - Introduction of PMML support for descriptor definition and model reporting: allows ***seamless cross-platform transfer*** of the models produced.
  - **One algorithm call** for both data preprocessing procedures (scaling, normalization, missing value handling) and calculation of domain of applicability to **increase efficiency** and avoid creation of intermediate data sets



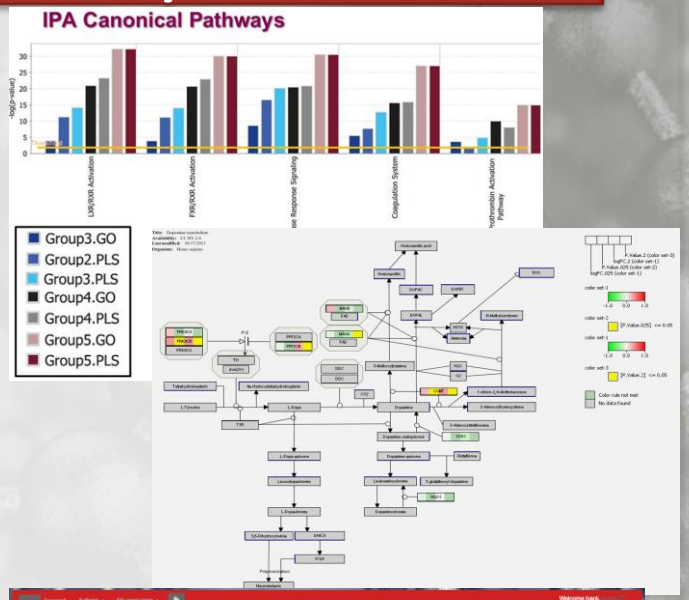
- **Descriptor Calculation Algorithms and Methods**
  - Development of **web tool for image descriptor** calculations. Source code: https://github.com/enanomapper/imageAnalysis

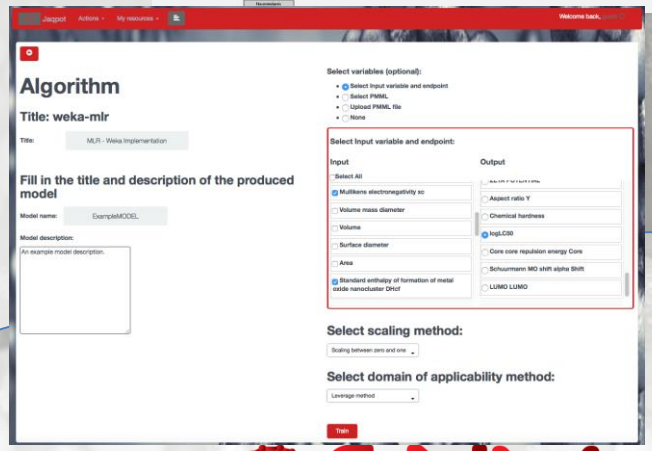
  First prototype: http://enanomapper.ntua.gr:8880/imageAnalysis/
  - **Gene Ontology (GO) descriptors**: Clustering of proteomics data based on Gene Ontology information, implemented in R language.
  - Utilization of **MOPAC OpenTox service** for developing Quantum mechanical descriptors for metal oxides
  - Extended the Java-based **Chemistry Development Kit** (CDK) with nanomaterial descriptors



eNanoMapper computational infrastructure (cont.)

- **Algorithm and modelling services**
  - **Extensions and updates of algorithm and modelling services**, compatible with API extensions and and fully integrated with the eNanoMapper Database.
  - Development of the **Conjoiner service** that performs the task of transforming the experimental data into a **modelling-friendly format**, and producing **standardized datasets**.
  - **Integration of third party services**: R language (OpenCPU), Python, WEKA
  - Implementation of **statistical and machine learning algorithms** (regression, classification, clustering) as **web services**
  - Development of **R tool for the creation of optimal QSAR models**: RRegrs, https://github.com/enanomapper/RRegrs
  - Provide services for **optimal experimental design** and **inter-laboratory comparison**
  - **Enrichment and Pathways Analysis**: Using many different approaches, like PathVisio, Cytoscape, Ingenuity Pathway Analysis (IPA) Chipster, GeneOntology, KEGG database.
  - Support specific needs of the community: **Collaborating with SUN project and RIVM** on the development of a **web-service for dose-response modelling** implementing the benchmark dose (BMD) method.
  - Developing a **modelling user interface** for easy access and use of modelling services, fully integrated with the eNanoMapper database, which can be additionally used for hosting public models.





Future perspectives

- **More Data – Big Data** (high throughput omics data –kinetics data)
- **New algorithms and modelling approaches** (handle sparse data, big data)
- **Integration of modelling approaches** (ab-initio modeling, atomistic scale with statistical approaches)
- **Hierarchical models** integrating systems toxicology with PBPK modelling for detailed multilevel organism/nanoparticle simulation as function of time

eNanoMapper

FP7-eNanoMapper

"eNanoMapper - A Database and Ontology Framework for Nanomaterials Design and Safety Assessment"

- Grant Agreement: 604134
- Duration: 36 months (1 Feb 2014 – 31 Jan 2017)