Enriching protein corona fingerprints using gene ontology information
An integration technique

Georgia Tsiliki,

NTUA

http://www.chemeng.ntua.gr/labs/control_lab

Overview

- eNanoMapper's computational infrastructure
- ENM descriptors
- GO descriptors
- Modelling and analysis tools for ENM predictive toxicology
  - RRegrs
- Application to protein corona data
  - Gene set enrichment analysis
  - Biological validation: Ingenuity Pathway Analysis

Computational Infrastructure

Computational Infrastructure

eNanoMapper's computational infrastructure is aiming to extract and analyse knowledge from diverse types of ENM-related theoretical descriptors, experimental data and associated metadata.

A number of modelling and analysis tools are being developed and implemented during the project, compliant to the OpenTox Application Programming Interface (API) and particularly tailored to the needs of ENM predictive toxicology. These include:

- Theoretical descriptors
- Modelling algorithms for correlating ENM properties with their biological and environmental impact
- Integrated analysis: experimental design, inter-laboratory testing, dose/response modelling

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)
  - Data **preprocessing** procedures (scaling, normalization, missing value handling) and calculation of domain of applicability through one algorithm call to increase efficiency and avoid creation of intermediate data sets
- Descriptor Calculation Algorithms and Methods
  - ImageJ: a web tool for **image descriptor calculations**. Source code: https://github.com/enanomapper/imageAnalysis , First prototype: http://enanomapper.ntua.gr:8880/imageAnalysis/
  - Utilization of **MOPAC** OpenTox service for developing Quantum mechanical descriptors for metal oxides
  - Extended Java-based **Chemistry Development Kit** (CDK) with nanomaterial descriptors
  - **Gene Ontology (GO) descriptors** (clustering of proteomics data based on Gene Ontology information, implemented in R language)

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)
  - Data **preprocessing** procedures (scaling, normalization, missing value handling) and calculation of domain of applicability through one algorithm call to increase efficiency and avoid creation of intermediate data sets
- Descriptor Calculation Algorithms and Methods
  - ImageJ: a web tool for **image descriptor calculations**. Source code: https://github.com/enanomapper/imageAnalysis , First prototype: http://enanomapper.ntua.gr:8880/imageAnalysis/
  - Utilization of **MOPAC** OpenTox service for developing Quantum mechanical descriptors for metal oxides
  - Extended Java-based **Chemistry Development Kit** (CDK) with nanomaterial descriptors
  - **Gene Ontology (GO) descriptors (clustering of proteomics data based on Gene Ontology information, implemented in R language)**

eNM eNanoMapper

eNanoMapper Framework

- NanoQSAR algorithm and modelling services
  - Extensions and updates of algorithm and modelling services to be compatible with API extensions and support of eNanoMapper Database (Access to algorithm and modelling services through swagger, http://enanomapper.ntua.gr:8080/jaqpot/swagger/#!/aa/login)
  - Integration of third party services: R language (OpenCPU), Python, WEKA
  - Implementation of statistical and machine learning algorithms (regression, clustering, classification) as web services
  - Development of R tool for the creation of optimal QSAR models (RRegrs, https://github.com/enanomapper/RRegrs/tree/master/RRegrs)
  - Creation of QSAR models for predicting cell association of gold nanoparticles using corona information
  - Pathway-based Analysis: Variable selection using GO descriptors/RRegrs and PLS/VIP methods on corona data. Enrichment Analysis using Ingenuity Pathway Analysis (IPA) software.

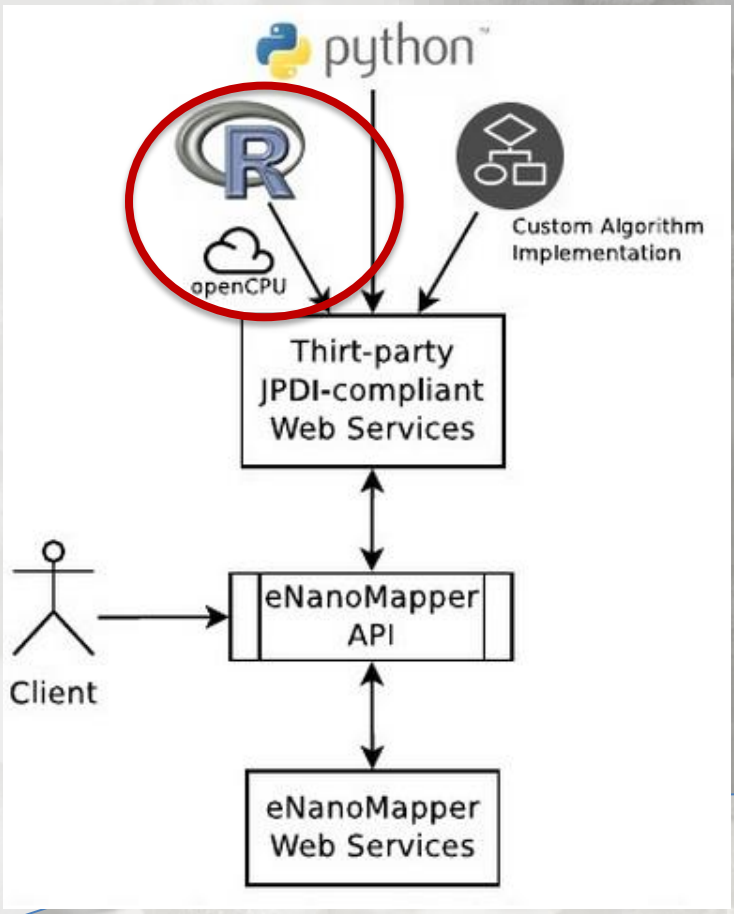
eNanoMapper Framework

- NanoQSAR algorithm and modelling services
  - Extensions and updates of algorithm and modelling services to be compatible with API extensions and support of eNanoMapper Database (Access to algorithm and modelling services through swagger, http://enanomapper.ntua.gr:8080/jaqpot/swagger/#!/aa/login)
  - Integration of third party services: R language (OpenCPU), Python, WEKA
  - Implementation of statistical and machine learning algorithms (regression, clustering, classification) as web services
  - **Development of R tool for the creation of optimal QSAR models (RRegrs,** https://github.com/enanomapper/RRegrs/tree/master/RRegrs)
  - **Creation of QSAR models for predicting cell association of gold nanoparticles using corona information**
  - **Pathway-based Analysis**: Variable selection using GO descriptors/RRegrs and PLS/VIP methods on corona data. Enrichment Analysis using Ingenuity Pathway Analysis (IPA) software.

eNanoMapper

Integration with third-party services

- The eNanoMapper framework already provides wrappers for **WEKA**, the **R** language, and the **Python** language
- To this end, eNanoMapper allows easy access to a **wealth of algorithms and methods**, as well as specially designed libraries for the analysis and interpretation of **–omics** and biological data
- Integration with R is made via OpenCPU system (https://www.opencpu.org/) which defines a HTTP API for embedded scientific computing based on R. OpenCPU acts as a wrapper for R to readily expose R functions as RESTful HTTP resources. This implementation uses forks of the R process to serve concurrent requests immediately with little performance overhead. By doing so it enables access to those functions on simple HTTP calls converting R from standalone application to a web service.



Statistical and machine learning algorithms exposed as web services

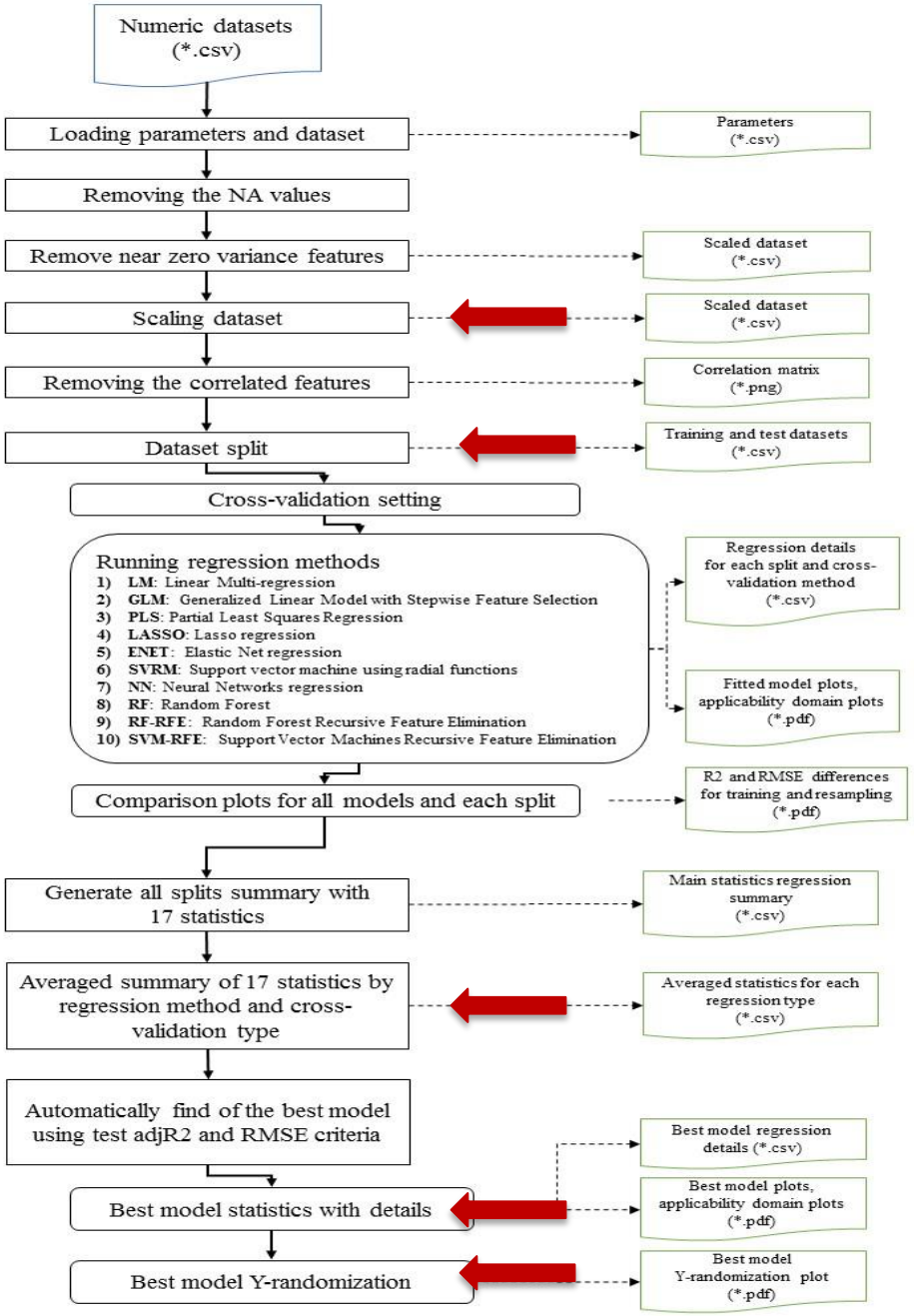
- Multivariate linear regression
- Lasso/ ridge regression
- Elastic Net
- Hierarchical clustering
- Bi-clustering
- ID3 decision tree
- Partial Least Squares
- PLS with VIP selection
- Radial basis function neural networks
- Support vector machines
- RRegrs

RRegrs: an R package for computer-aided model selection

RRegrs: Develop a tool to explore the space of linear and non-linear QSAR prediction models (Tsiliki et al. 2015).

- An **easy-to-use framework** for model selection offering extensive capabilities for model comparisons
- Apply several simple and complex **regression methods**
- Other features include: data set **splitting**, **cross-validation** methods, specific **regression parameters** and **best model criteria** which affect the accuracy and efficiency of the produced predictive models
- Produce **standardized summary** and comparison outputs (text format, graphs)
- An easy-to-use tool accessible to all users irrespectively of their statistical background
- A free programming package which can be continuously improved by the users and adopted to their needs
- https://github.com/enanomapper/RRegrs

RRegrs workflow



eNanoMapper

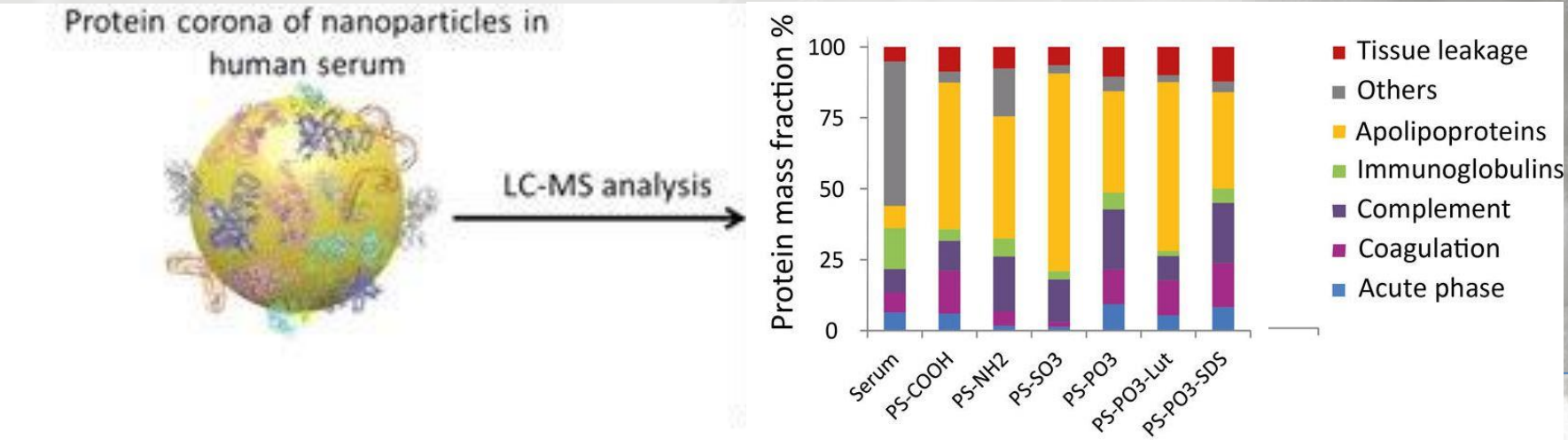
Protein corona data

NPs protein corona

- When NPs are exposed to a biological medium, different biomolecules (proteins, lipids and glycans) will compete to interact with the NP surface to form a layer called **'protein corona'**
- Important factors for the protein corona composition are the **physicochemical properties** of NPs, such as particle size, shape, and charge, and the characteristics of NPs biological environment
- The protein corona **modifies NP's physicochemical properties**, thus affecting biological responses such as cellular uptake, kinetics, signaling, accumulation, transport and toxicity
- Understanding nanoparticle-proteins interactions is a crucial issue in the development of targeted nanomaterial delivery:
  - Besides unravelling the composition of the NP protein coronas, distinct proteins could control NP's uptake into specific cell types

NPs protein corona

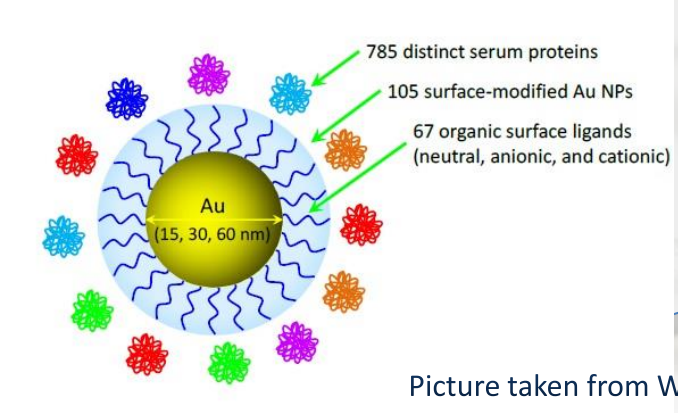
- The protein corona establishes the **biological identity** of the NP given the proteins absorbed onto its surface when that comes into contact with a biofluid
- It has been found that the interaction with cell membranes and the mechanisms of **cellular uptake** is controlled by the absorbed proteins [6]



Picture taken from Ritz et al. [6]

Protein corona fingerprinting

- Protein corona data in [1,2] consist of
  - **84** gold and 12 silver NPs
    - 84 anionic/cationic, 21 neutral surface ligands
  - **785** distinct serum proteins were identified by LC-MS/MS
    - 129 proteins suitable for relative quantification → **'fingerprint'** to characterize the protein corona
    - 76 proteins selected based on iterative PLSR with VIP>=0.6 (model training)
- Cell association using A549 human lung epithelial carcinoma cells was quantified to model biological interaction
- Net cell association (**cellular interaction**) was chosen as the response variable (Y) and the **relative abundancies** of LC-MS/MS data were set as the explanatory variables (X) of the model



Picture taken from Walkey et al. [2]

Protein corona fingerprinting

- Scope: predict NP's toxicity solely by proteomics data and by mean of fully validated QSAR models
- Net cell association (**cellular interaction**) was chosen as the response variable (Y) and the **relative abundancies** of LC-MS/MS data were set as the explanatory variables (X) of the model

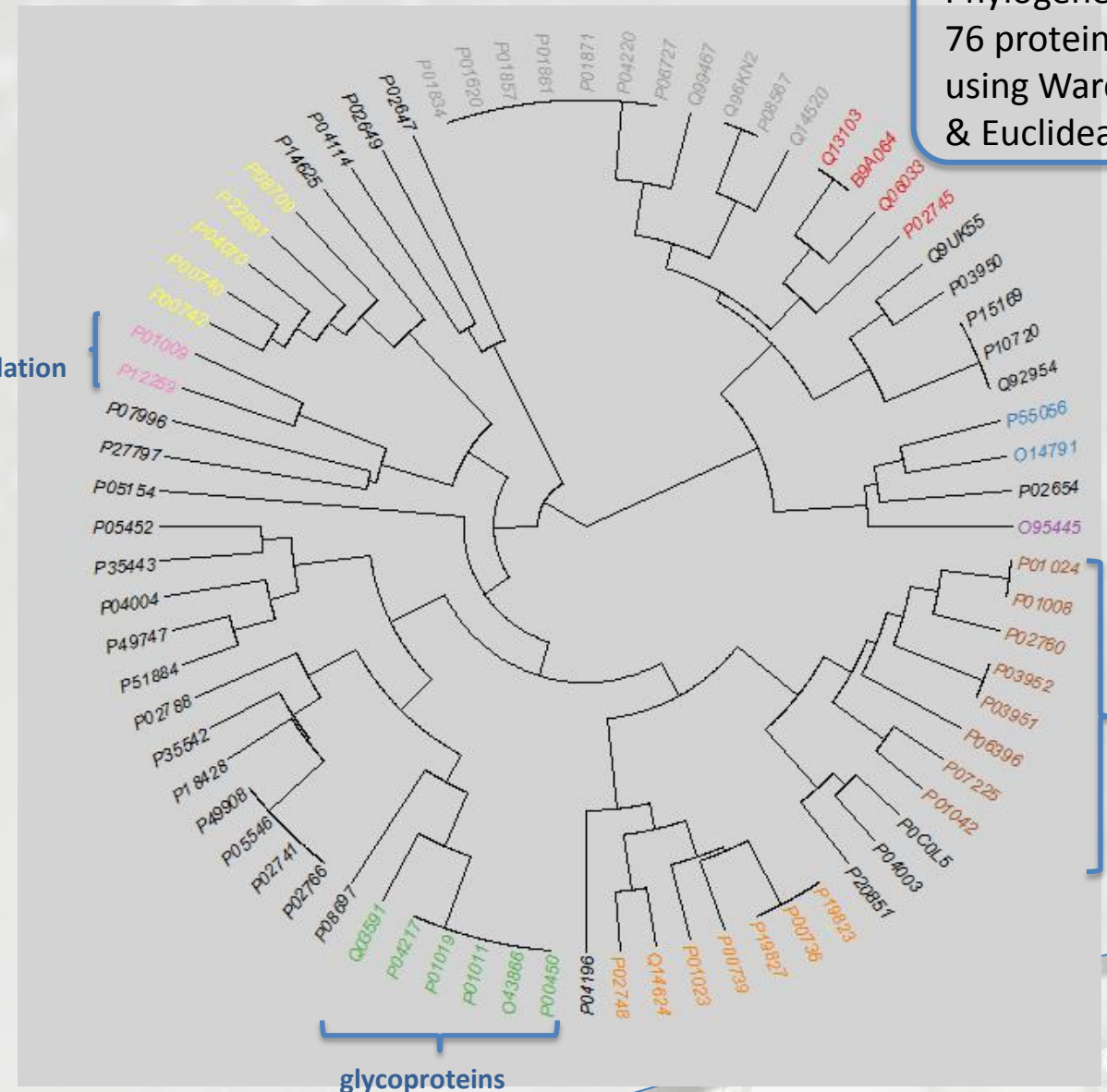
Walkey et al.[1] analysis:

- Proteins are sorted by the Variable Importance to the projection (**VIP**) i.e. the importance of a particular protein to the prediction task
- For the PLSR-VIP selected proteins R^2_{LOO}=0.81, R^2_{4CV}=0.61
- Models that use the **top 48**, **32**, **16**, **6** serum proteins are 99%, 95%, 83% or 74% as accurate as the full model

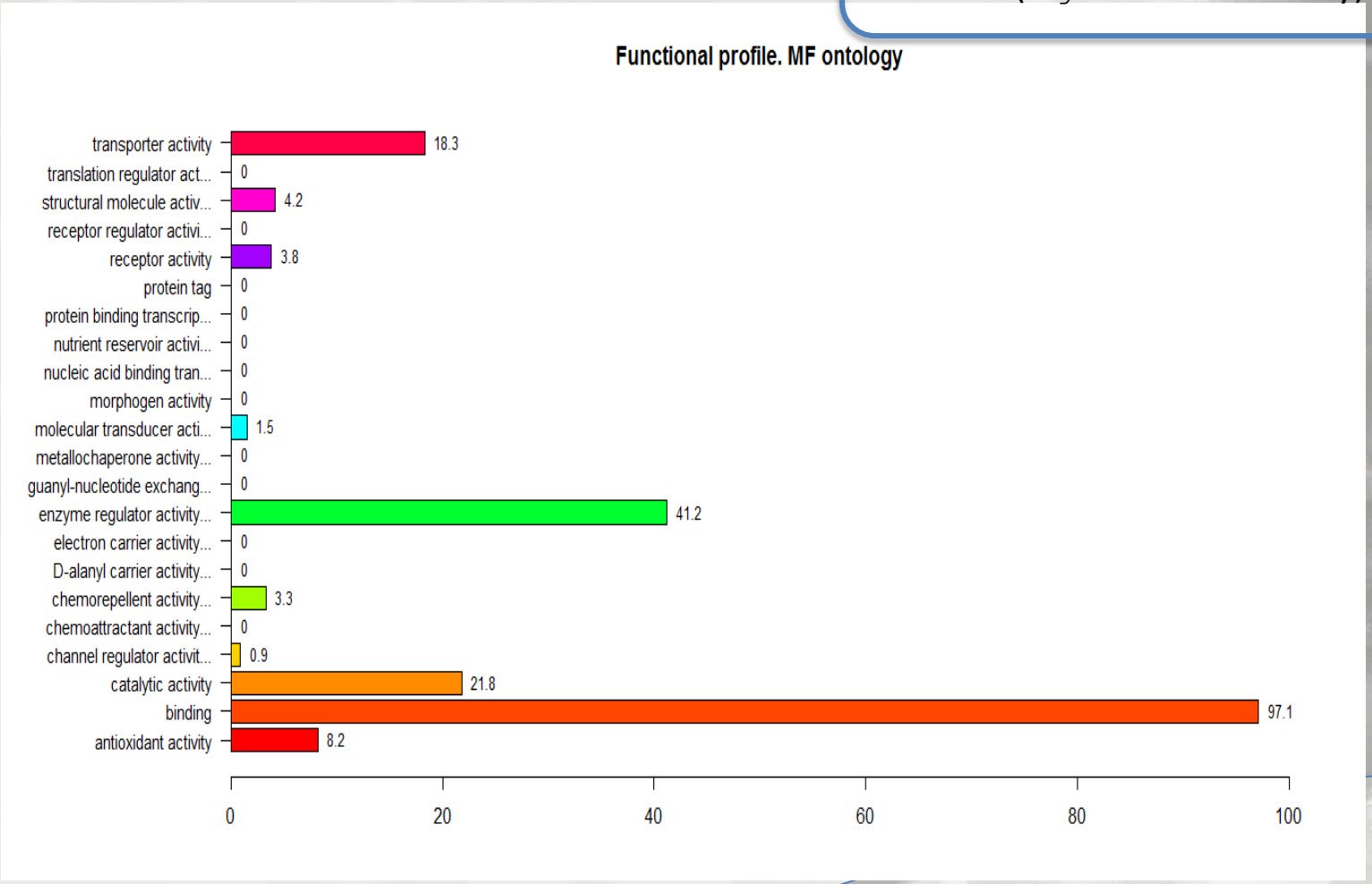
Liu et al.[2] analysis:

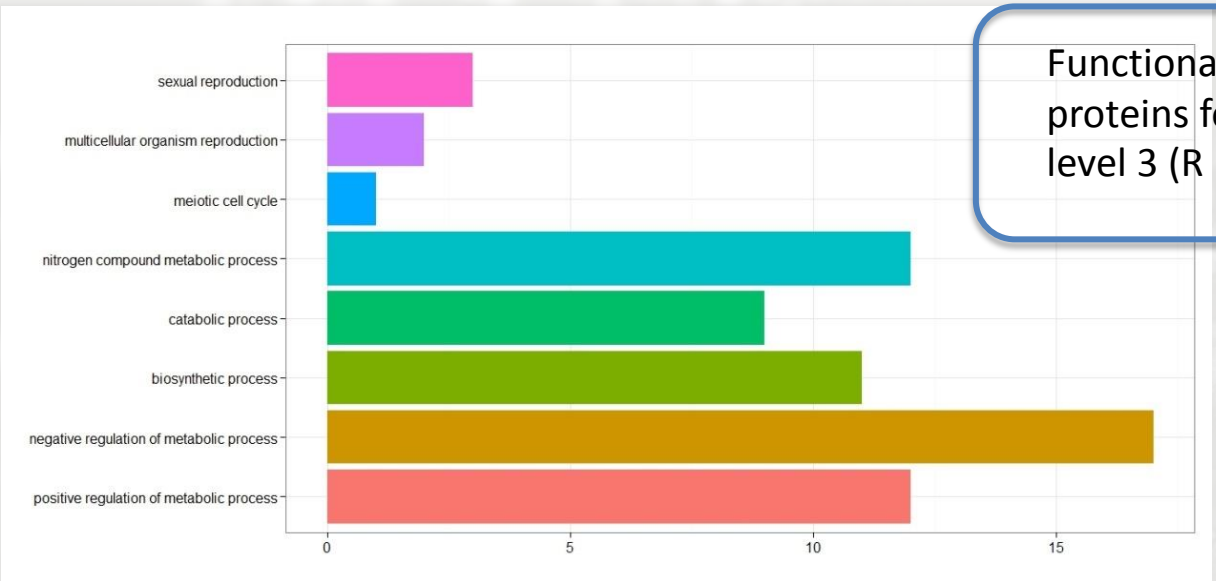
- Linear and Support vector regression models are employed together with sequential forward floating selection
- Models consider protein 'descriptors' and physichochemical properties
- R^2_{4CV}=0.862 (SVR- 6 serum proteins & zeta potential), R^2_{4CV}=0.843 (LM- 11 serum proteins)

Phylogenetic-like dendrogram for 76 proteins (hierachical clustering using Ward agglomeration method & Euclidean distance)

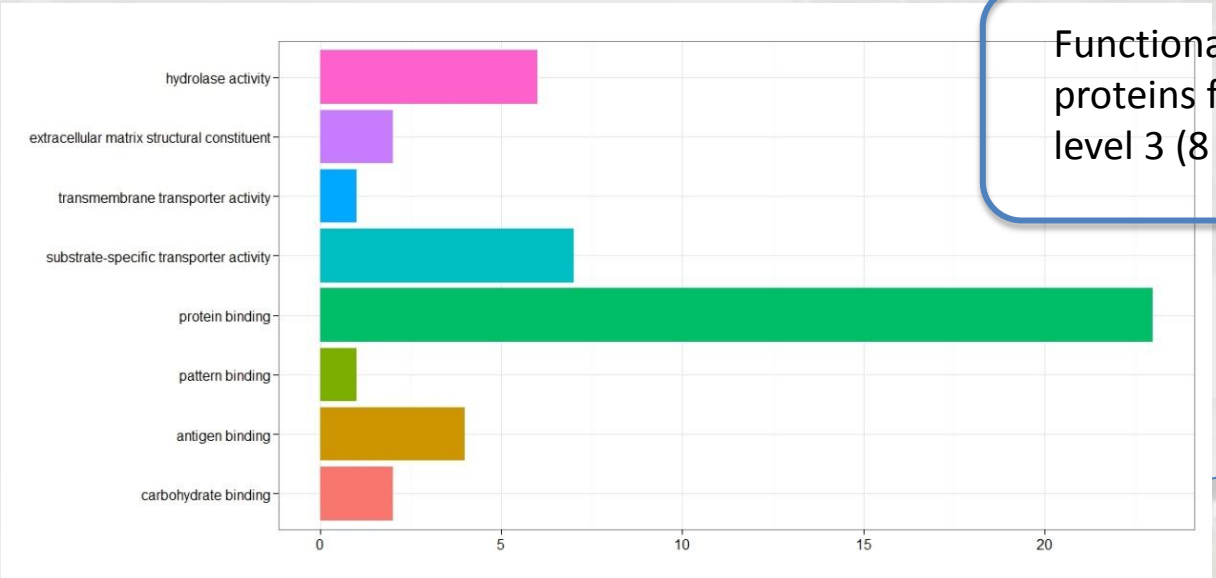


Functional profiles of the 76 proteins for MF GO ontology built at level 2 (R `goProfiles` library)

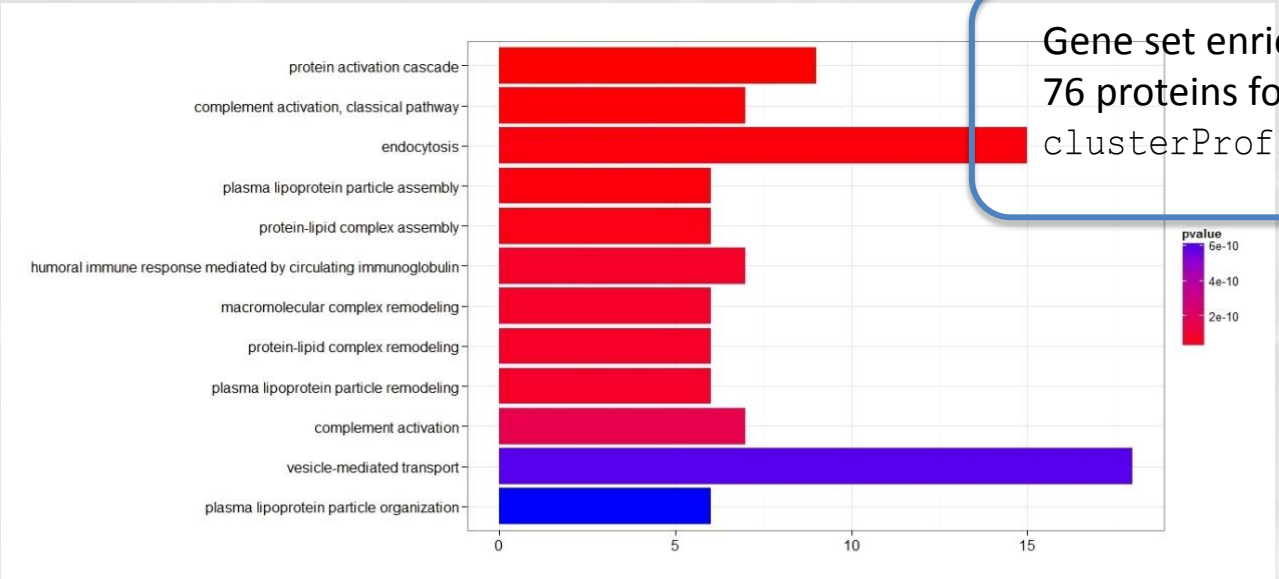




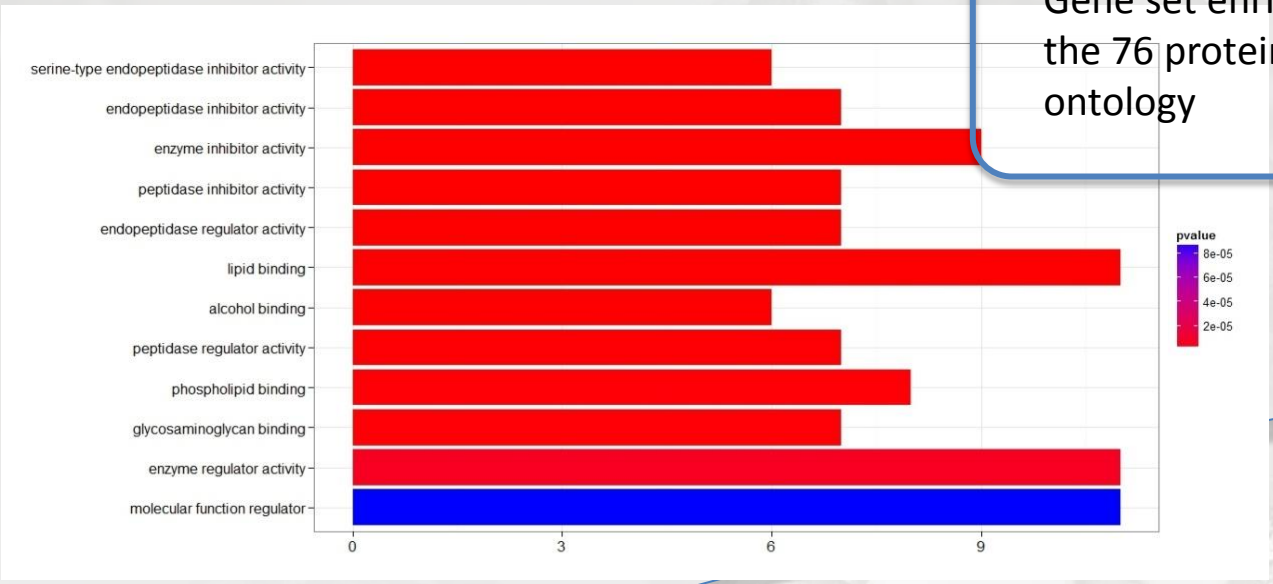
Functional profiles of the 76 proteins for BP GO ontology at GO level 3 (R `clusterProfiler` library)



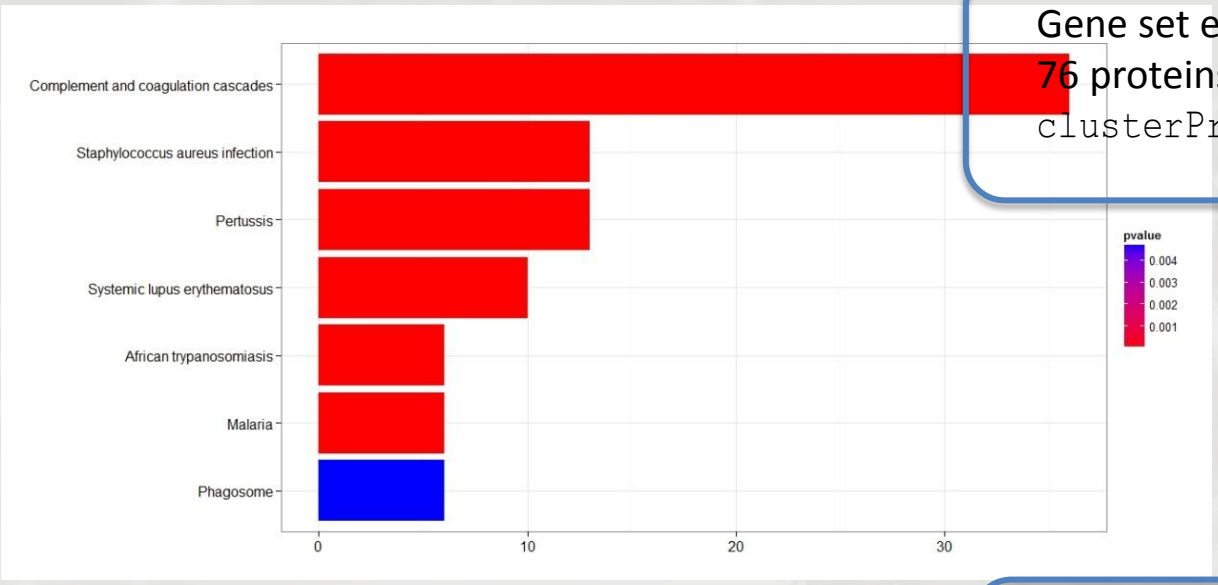
Functional profiles of the 76 proteins for MF GO ontology at GO level 3 (8 first categories)



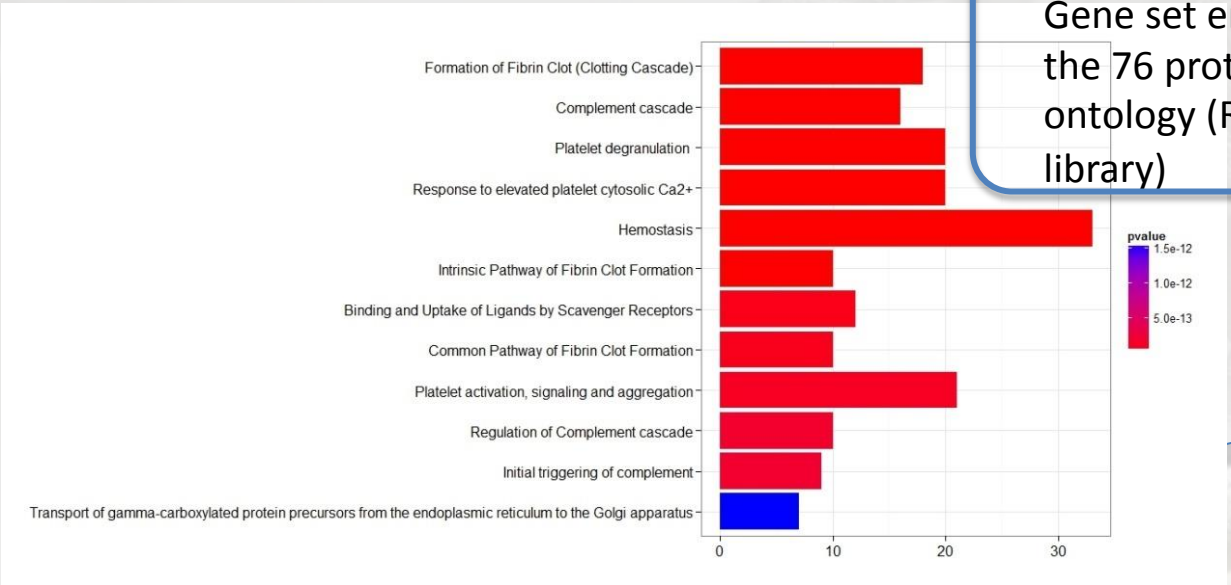
Gene set enrichment analysis of the 76 proteins for BP GO ontology (R `clusterProfiler` library)



Gene set enrichment analysis of the 76 proteins for MF GO ontology



Gene set enrichment analysis of the 76 proteins for KEGG ontology (R `clusterProfiler` library)



Gene set enrichment analysis of the 76 proteins for Reactome ontology (R `clusterProfiler` library)

Biological Process	GO	KEGG	Reactome	GO: topGO	GO:GOSim	toxicity
Complement activation	GO:0072376 GO:0006958 GO:0006956	hsa04610	140877;109582 140837;140875 76002	GO:0042060 GO:0030193 GO:0050817 GO:0061041 GO:1900046	GO:0030195 GO:0010543	√
Inflammation	GO:0002455;GO:0006959 GO:0016064;GO:0019724 GO:0006952;GO:0019724 GO:0002526	hsa04610	166658 977606 166663	GO:0030449 GO:0006956 GO:0006958	GO:0072376 GO:0006958 GO:0006956	√
Lipid transport	GO:0034377;GO:0065005 GO:0034368;GO:0003469 GO:0071827;GO:0071825 GO:0033344;GO:0097006 GO:0042157;GO:0030301 GO:0010876	hsa04145		GO:0006629;GO:0008610 GO:0006869;GO:0019915 GO:0016042;GO:0097006	GO:0034377;GO:0065005 GO:0034368;GO:0071827 GO:0033344;GO:0010873 GO:0043691;GO:0042632 GO:0033700;GO:0001523	√
Coagulation	GO:1903034 GO:0030193 GO:0042060 GO:0007596		114608 76005 2173782	GO:0006898;GO:0018200 GO:0018214;GO:0006909	GO:0006897 GO:0006898	√
Cell association	GO:0006897 GO:0006898	hsa05322	166786	GO:0002520;GO:0030097 GO:0048534;GO:0045087 GO:0002697;GO:0002703 GO:0002706;GO:0002712 GO:0002819;GO:0002920	GO:0002455;GO:0006952 GO:0016064;GO:0006959 GO:0019724;GO:0006953 GO:1903027;GO:0004507	√
Metabolic processes	GO:0051248;GO:0070613 GO:0010955		159763;159740 159782;159854 163841	GO:0017187;GO:0018200 GO:0018214	GO:0010951	
Infectious diseases		hsa05133 hsa05143 hsa05144				

GO descriptors

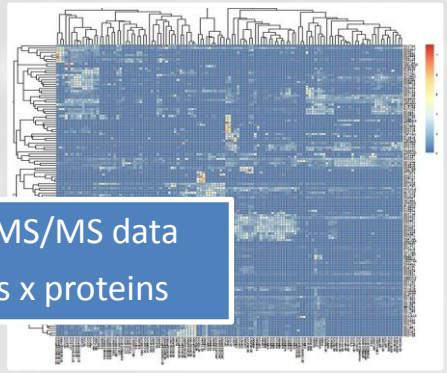
GO descriptors

Scope: Create new descriptors that would summarize proteomics data and yet build statistically significant predictive models

- Based on pathway analysis: **groups of proteins** corresponding to particular pathways could produce new biologically interpretable descriptors
- Systematic approach of **integrating various types of data** (-omics) and be able to simultaneously assess the biological meaning of the analysis outcome. For example there are studies that incorporate genomic knowledge such as pathways or protein-protein interaction networks to increase their power in predicting biologically relevant information [6] [7]

GO descriptors: Integrate **Gene Ontology information** with **proteomics data** (relative abundancies) to produce new descriptors

- Select GO category (Biological Process, Cellular Component, Molecular Function)
- Identify the important GO ids (hypergeometric test p-values)
- Apply clustering algorithm to produce protein clusters
- Summarize proteomics data based on the clusters produced
- **Report results** for protein corona data using RRegrs
- **Biological validation**: Gene set enrichment analysis - Ingenuity Pathway Analysis



LC-MS/MS data

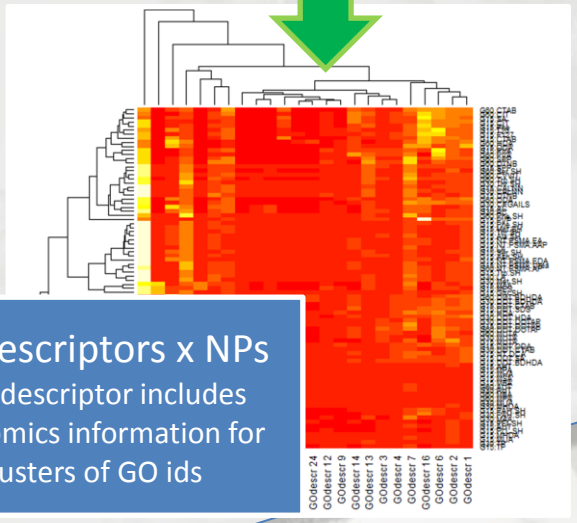
NPs x proteins

Get protein ids
Find the statistically significant GO ids

GO membership matrix

GO ids x proteins

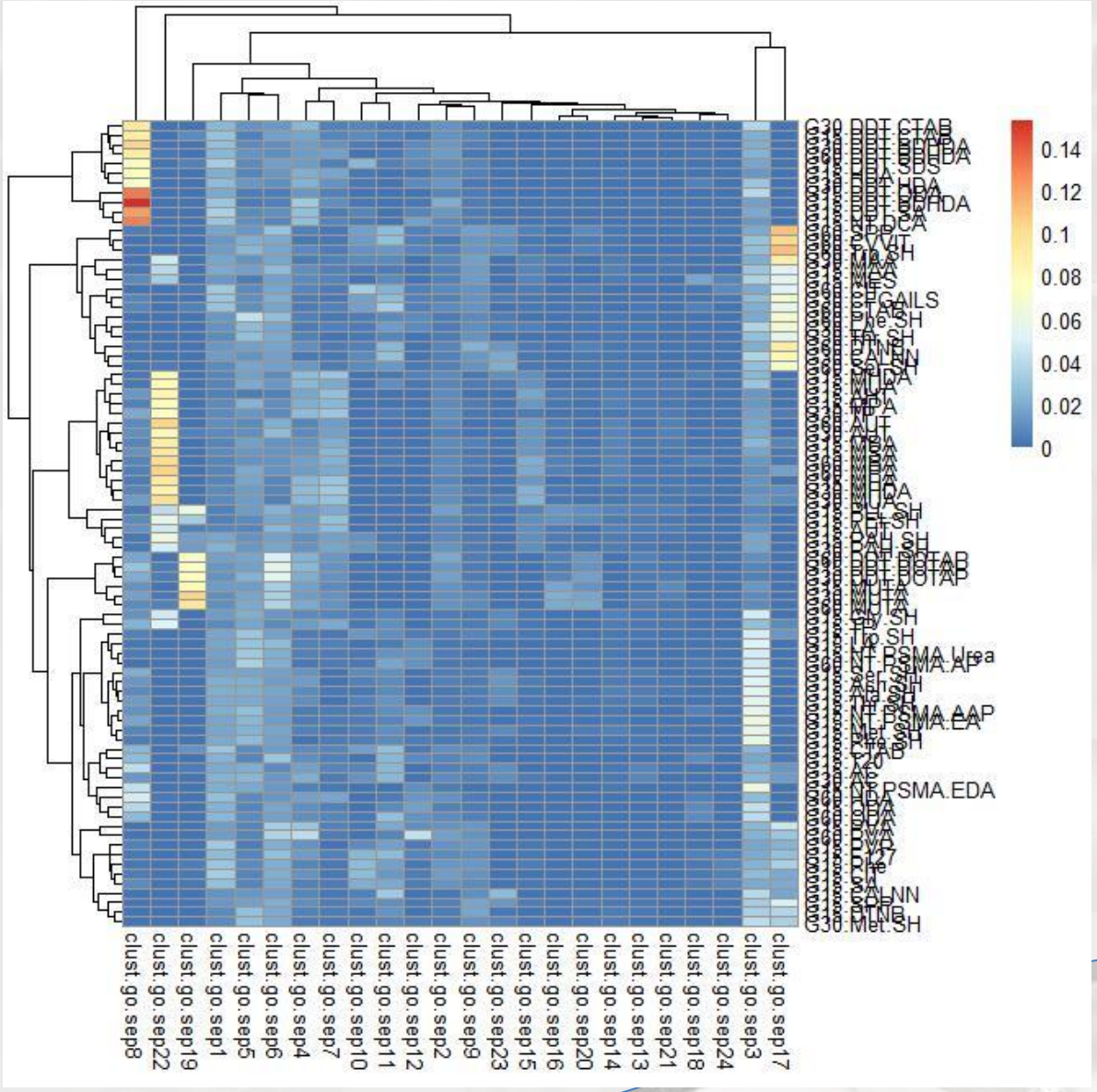
Find clusters of proteins
from GO matrix and
apply to summarize LC-MS/MS data



GO descriptors x NPs

Each descriptor includes proteomics information for clusters of GO ids

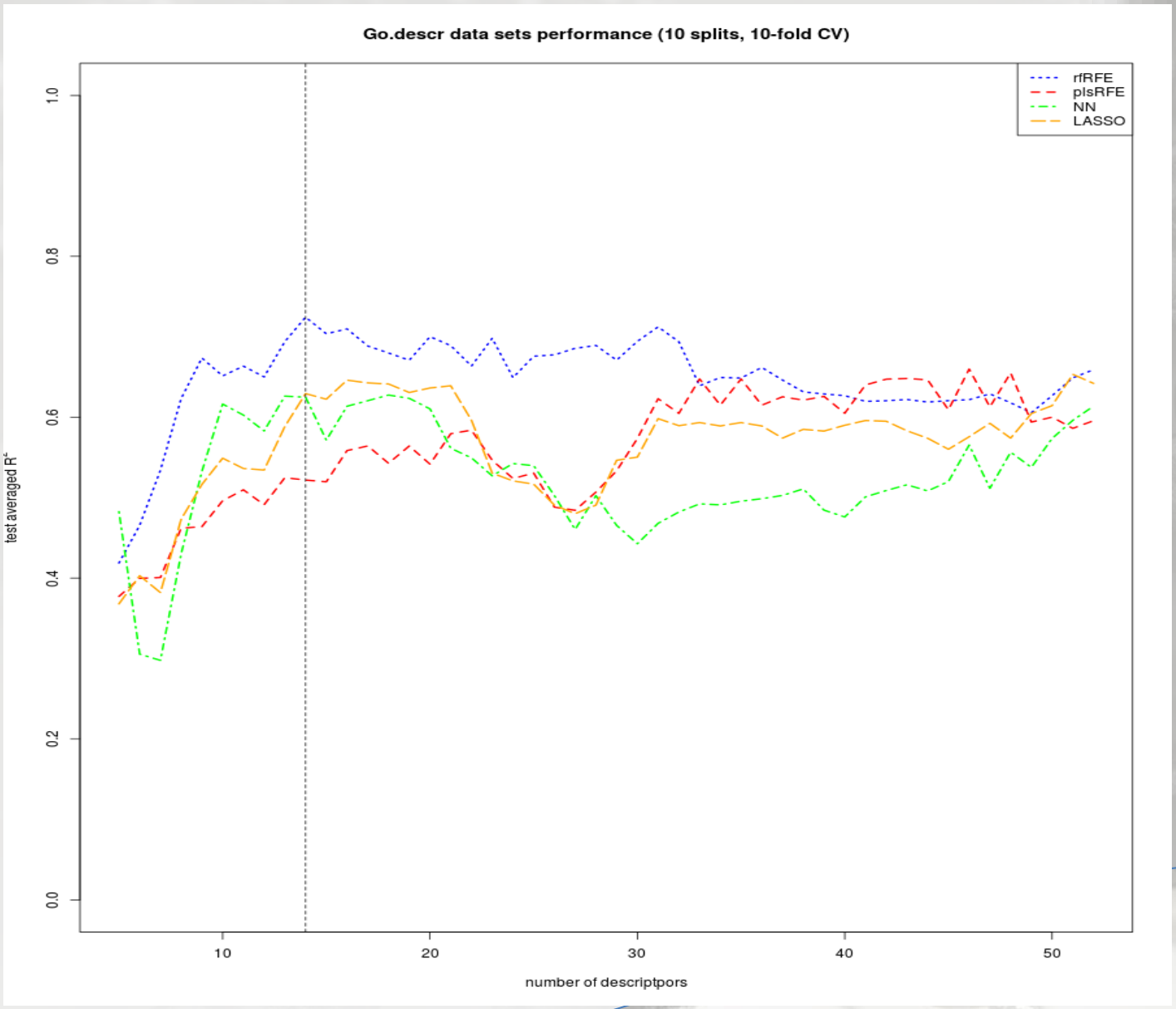
eNanoMapper


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
clust.go.sep8
clust.go.sep22
clust.go.sep19
clust.go.sep1
clust.go.sep5
clust.go.sep6
clust.go.sep4
clust.go.sep7
clust.go.sep10
clust.go.sep11
clust.go.sep12
clust.go.sep2
clust.go.sep9
clust.go.sep23
clust.go.sep15
clust.go.sep16
clust.go.sep20
clust.go.sep14
clust.go.sep13
clust.go.sep21
clust.go.sep18
clust.go.sep24
clust.go.sep3
clust.go.sep17

GO descriptors: application to protein corona data

Results:

- Apply RRegrs framework to GO descriptors data
- Best GO.descr data set: the best set of GO descriptors found to summarize the 76 proteins set consists of **14 GO descriptors**
- Our best model is reported for **RF methodology** (embedded with feature selection, RFE) **R^2_{Test}(RF)=0.73**, which is only marginally lower than the best results reported here for the protein corona data set, i.e. R^2_{Test}(SVMRADIAL)=0.74. Its is worth noticing that for the PLS model which is the one presented in the original publication, R^2_{Test}(PLS)=0.73 (R^2_{4CV}=**0.61** in [1])
- R^2_{Test}: random split, averaged R^2 test values


Go.descr data sets performance (10 splits, 10-fold CV)
rfRFE
plsRFE
NN
LASSO
test averaged R²
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
number of descriptpors

GO descriptors: fully validated results using RRegrs

	RegrMeth	Split.No	CVtype	14 Go.descr Averaged R2 (in test)	76 prot.cor Averaged R2 (in test)
1	glmnet	10	LOOCV	0.62134	0.68539
2	glmnet	10	repeatedcv	0.616039	0.674395
3	svmRFE	10	repeatedcv	0.572766	0.664595
4	rf	10	repeatedcv	0.721779	0.656561
5	svmRadial	10	LOOCV	0.676144	0.738791
6	svmRadial	10	repeatedcv	0.689997	0.74111
7	rbfDDA	10	repeatedcv	0.048119	0.114047
8	lasso.RMSE	10	repeatedcv	0.629373	0.638921
9	pls	10	LOOCV	0.537477	0.723371
10	pls	10	repeatedcv	0.537477	0.729684
11	glmStepAIC	10	LOOCV	0.60609	0.057425
12	glmStepAIC	10	repeatedcv	0.60609	0.057425
13	lm	10	LOOCV	0.620863	0.057425
14	lm	10	repeatedcv	0.620863	0.057425
15	rfRFE	10	repeatedcv	0.728865	0.651195
16	pls.WSel	10	LOOCV	0.502733	0.678559
17	pls.WSel	10	repeatedcv	0.500781	0.678559
18	nnet	10	LOOCV	0.622386	0.610303
19	nnet	10	repeatedcv	0.612911	0.617143

RRegrs call with default parameters normalization option:

- 10 random splits
- 100 Y-randomization runs

GO descriptors: RF stepwise model

Based on the above results, we focus on the RF methodology and report the GO descriptors importance as well as their performance in terms of averaged R^2 values in the test set.

- perform **10-fold repeated CV**
- new variables are added one at a time, starting by the first three most important

Number of GO descriptors	Descending order of descriptors' importance	Averaged R^2 values (RF in test)
1	10	
2	11	
3	6	0.518204
4	12	0.526117
5	9	0.627913
6	5	0.670172
7	4	0.699355
8	7	0.704904
9	8	0.691104
10	3	0.71467
11	2	0.723231
12	1	0.718921
13	14	0.719899
14	13	0.728865

RF & embedded variable importance feature (permuting the Out-Of-Bag data per tree optimizing mean decrease in accuracy/ mean decrease in MSE)

GO descriptors: application to protein corona data

We have placed GO descriptors in five groups G1-G5 based on their performance, R^2_{Test}(RF), by including protein sets with decreasing predictive power

Group	G1	G2	G3	G4	G5
Description	*go.10*	*go.11*	*go.10*	*go.10*	*go.10*
			go.11	*go.11*	*go.11*
			go.6	*go.6*	*go.6*
				go.12	*go.12*
				go.9	*go.9*
				go.5	*go.5*
				go.4	*go.4*
				go.7	*go.7*
				go.8	*go.8*
				go.3	*go.3*
				go.2	*go.2*
					go.1
					go.14
					go.13
Size GO descr	5	1	17	50	76
Size PLS sets	6	16	32	48	76

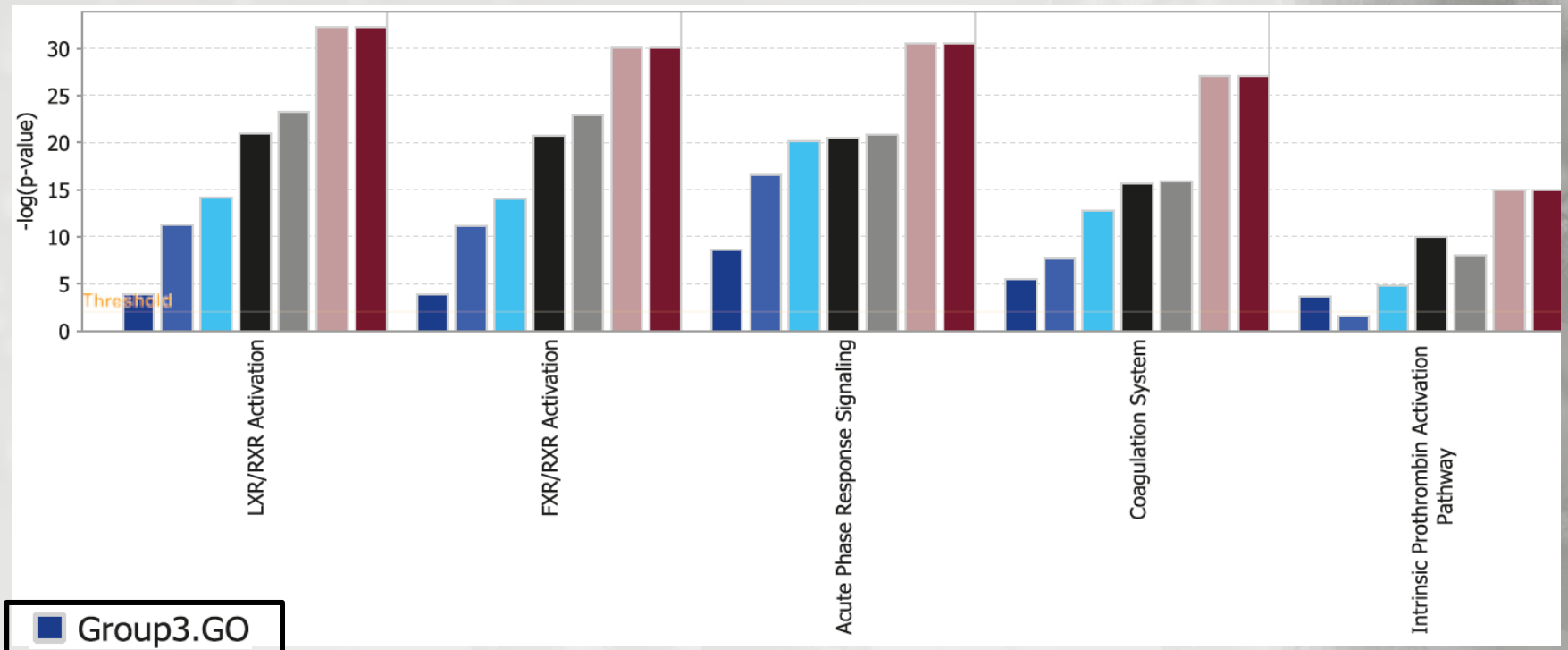
P01009
P04003
P02766
P43866
P12259
P03760
P49908
P15169
P35542
P08709
Q03591
P01011
P07225
P00740
P08567

eNanoMapper

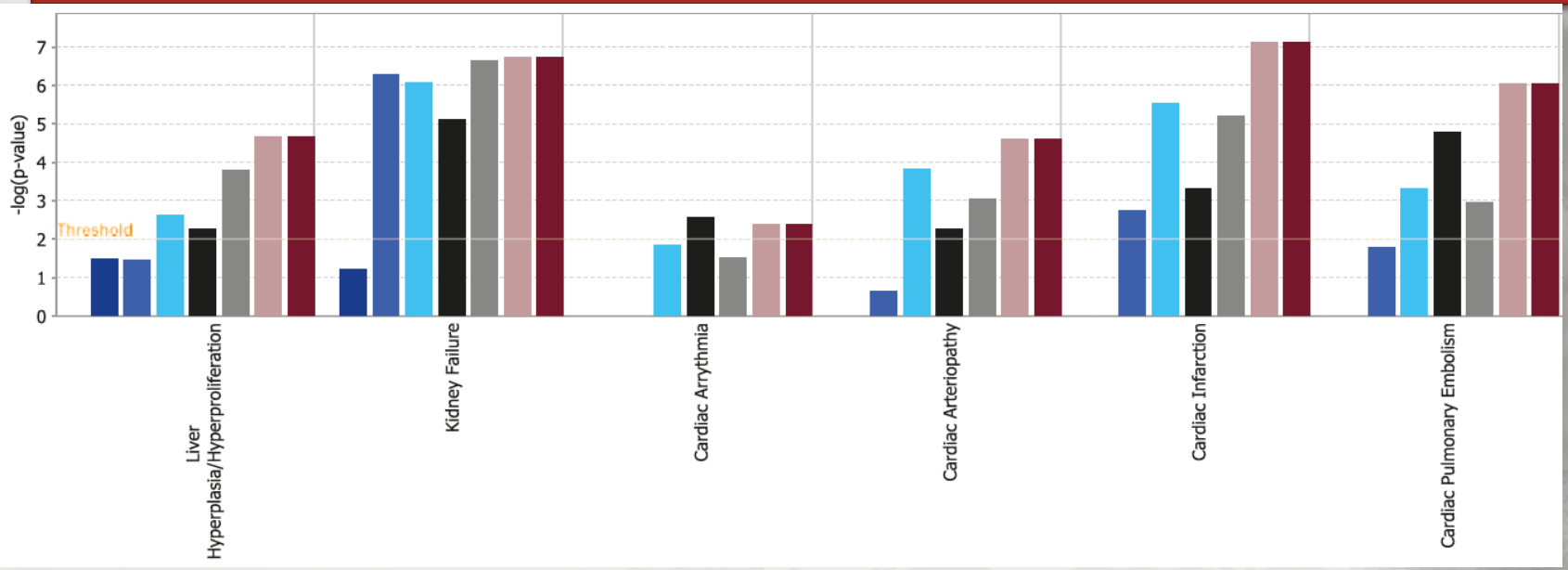
IPA comparative analysis

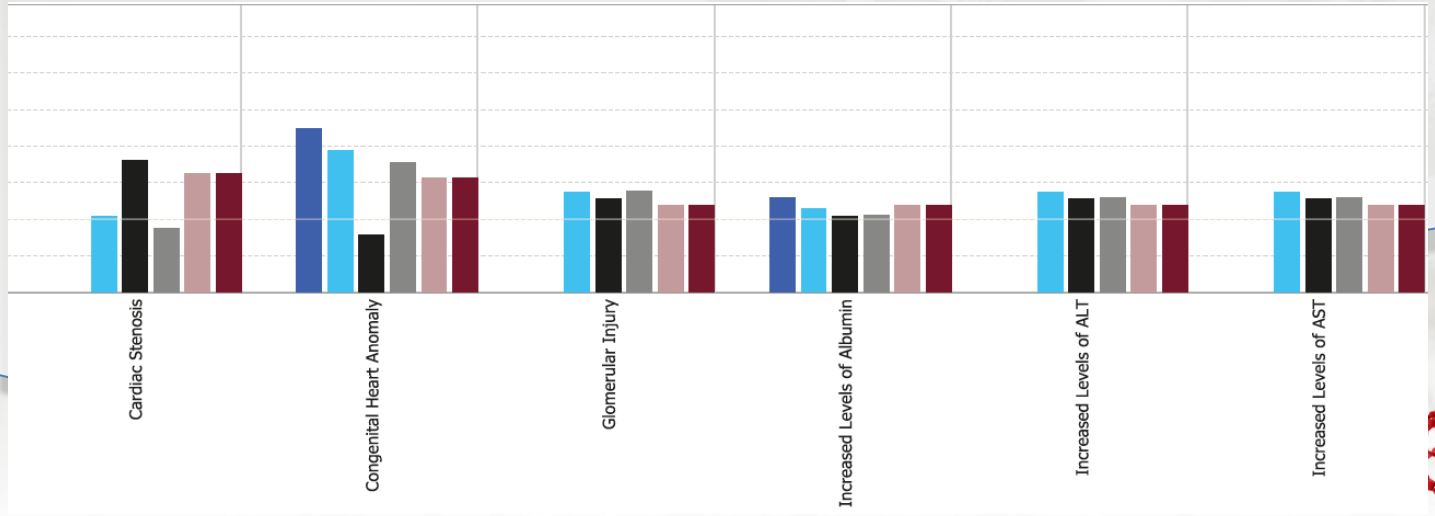
- Predictive proteins identified by GO-based models are arranged into 5 groups (as in table before)
- For comparison the Partial Least Squares (PLS)-based modelling from Walkey et al.[1] is repeated using methods as presented in the original publication and results are also arranged into 5 groups using VIP score. The distribution of the protein sets is:
  - G1: 6, G2: 16, G3: 32, G4:48, G5: 76
- Comparative IPA is carried out between GO and PLS groups with at least 10 proteins in the analysis results

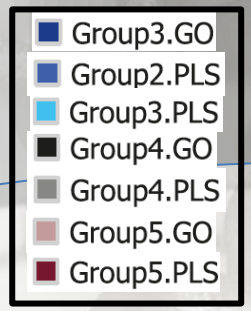
IPA comparative analysis: Canonical Pathways



IPA comparative analysis: ToxFunctions

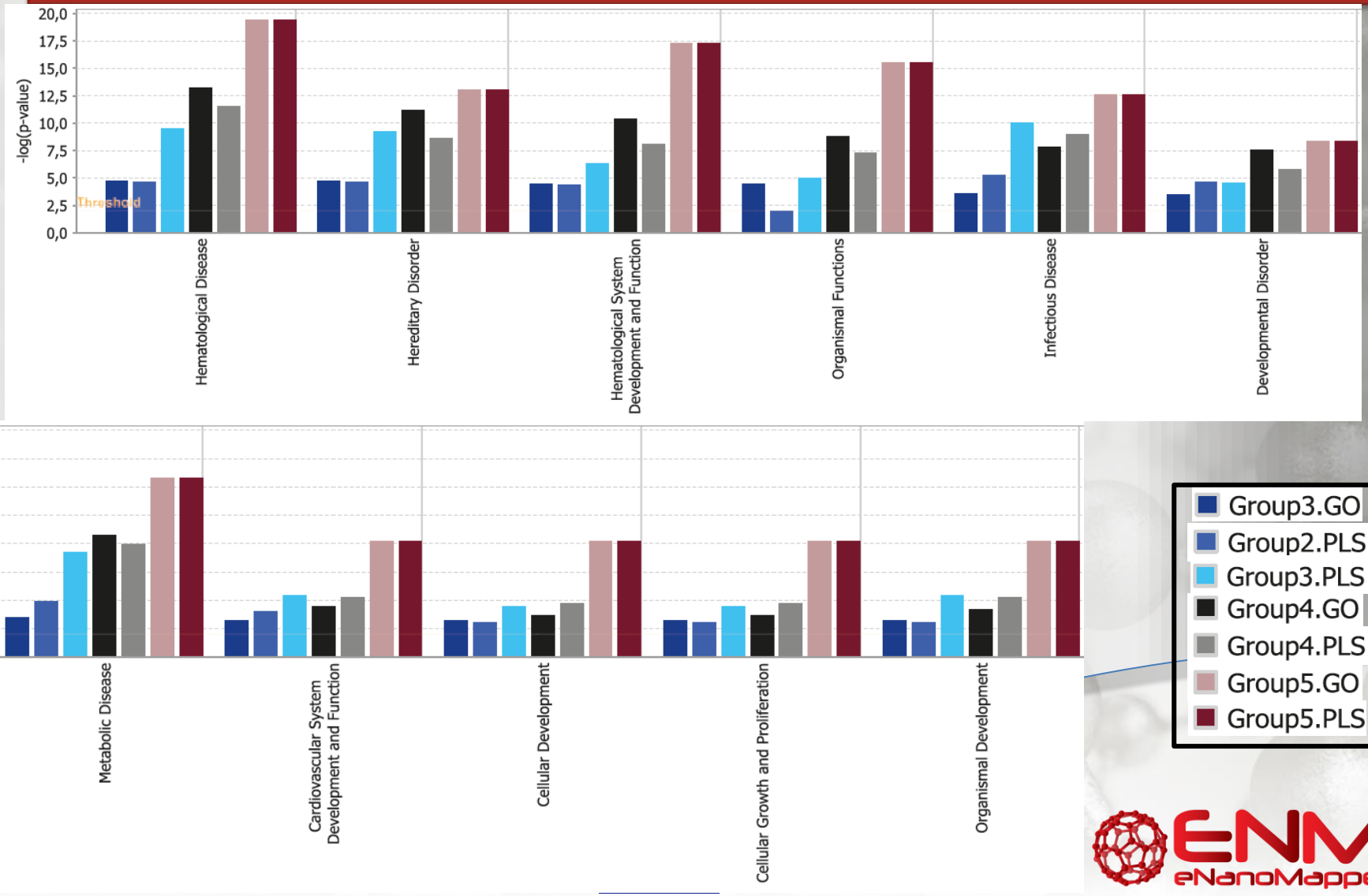






eNM eNanoMapper

IPA comparative analysis: Disease BioFunctions



eNanoMapper

IPA comparative analysis: summary of results

- GO-derived and PLS modelling results give similar biological output results when the number of genes is above 10
- The proteins classify primarily to **extracellular** or **secreted proteins** (60/76). Processes implicated include **lipid metabolism**, complement activation and **blood coagulation**, however, also with association to hematological, hepatic, **renal and cardiac toxicity**, including hematological and cardiac related diseases.
- The main Canonical Pathways and ToxLists results are seen already in the first group. These include: LXR/RXR Activation, **Acute Phase Response Signaling**, FXR/RXR Activation, Coagulation System, Complement System, Atherosclerosis Signaling
- ToxLists and Canonical Pathways produced mechanistically informative results with expected correlations between enrichment and number of genes
- Specific IPA ToxFunctions were implicated, including **liver hyperplasia/hyperproliferation** with Group1 list
- Overall many disease processes typical of repeated dose toxicity were implied including: Cardiovascular Disease, Hematological Disease, Hematological System Development and Function, Metabolic Disease and Infectious Disease that are also active with the Top6 protein list

Conclusions

Integration analysis findings

- Protein corona data predicts NPs' toxicity with high accuracy
- Mechanistic modelling – pathway analysis helped us to get an insight of toxicity mechanisms
  - Statistically significant pathways found by **R** (**GO**, **KEGG**, **REACTOME** databases) and **IPA** software,
    - **Acute Phase Response Signaling** [8][9]
  - GO analysis produced similar (and complementary) results to the original publication (results were validated with IPA also)
  - Those established correlations were used for producing GO descriptors
- **Why integrating GO information?**
  - Biological information would assist **biological interpretation**
    - Identify set of proteins of potential relevance to toxicity
    - Some predefined gene sets are over-represented & thus play a role in disease etiology
    - Available biological information is used to supplement the disease gene hierarchy provided by proteomics data
  - Improve **power** & **reproducibility** for QSAR models
  - Compact data set (reasonably **small number** of descriptors, 129 → 14)
- Currently working towards increasing prediction accuracy
  - improving clustering performance,
  - stochastic search for number of clusters- i.e. descriptors
- Development of web-services compatible with APIs at the eNM infrastructure

eNM eNanoMapper

Thanks to Pekka Kohonen, Penny Nymark, Alexndra Roussi

References

1. Walkey CD, Olsenb JB, Songa F, Liug R, Guob H, et al. (2014) Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold Nanoparticles. *ACS Nano*, *8* (3); 2439–2455, DOI: 10.1021/nn406018q.
2. Liu R, Jiang W, Walkey CD, Chan WC, Cohen Y (2015) Prediction of nanoparticles-cell association based on corona proteins and physicochemical descriptors. *Nanoscale, 7(21): 9664-75*
3. Mistry, M., and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics, 9:327.
4. Frohlich, H., Speer, N., and Zell, A. (2006). Kernel based functional gene grouping. In Proc. Int. Joint Conf. Neural Networks, pages 6886 -6891.
5. Ritz S, Schöttler S, Kotman N, Baier G, Musyanovych A, et al. (2015) Protein Corona of Nanoparticles: Distinct Proteins Regulate the Cellular Uptake *Biomacromolecules* 2015 *16* (4), 1311-1321, DOI: 10.1021/acs.biomac.5b00108
6. Balbin OA, Prensner JR, Sahu A, Yocum A, Shankar S, et al (2013) Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat. Commun. 4*, 2617 DOI: 10.1038/ncomms3617.
7. Yang X, Regan K, Huang Y, Zhang Q, Li J, et al. (2012) Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Comput. Biol.* 8, DOI: 10.1371/journal.pcbi.1002350.
8. Saber AT, Lamson JS, Jacobsen NR, Ravn-Haren G, Hougaard KS, et al. (2013) Particle-Induced Pulmonary Acute Phase Response Correlates with Neutrophil Influx Linking Inhaled Particles and Cardiovascular Risk. *PLoS ONE* 8(7): e69020. doi: 10.1371/journal.pone.0069020
9. Higashisaka K, Yoshioka Y, Yamashita K, Morishita Y, Fujimura M, et al. (2011) Acute phase proteins as biomarkers for predicting the exposure and toxicity of nanomaterials.. *Biomaterials* 32, no. 1 (2011): 3-9.

eNanoMapper Associate Partner Program:
https://drive.google.com/file/d/0B79AKP6tuR7SOU54UjBDcmV1SEE/view?usp=sharing
associate-partner@lists.enanomapper.net