PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Computational systems biology in the 21st century: data sharing and crowd-sourced challenges

Dimitris Messinis

Postdoctoral Fellow PMI R&D

The research described in this presentation was sponsored by Philip Morris International.

About me

Science & technology

- 2018–present: **Systems Pharmacology Postdoctoral Fellow**, Philip Morris R&D
- 2014–2018: **PhD in Systems Bioengineering**, National Technical University of Athens (NTUA)
  - 2016: **Data Scientist**, U.S. Food & Drug Administration (FDA), Washington D.C.
  - 2017: **Health Alumnus**, European Institute of Innovation & Technology (EIT)
- 2011–2014: **Research Scientist**, Protavio Ltd (biotech startup)
- 2005–2011: **Diploma in Mechanical Engineering**, NTUA
  - 2009–2011: **Research Assistant**, Systems Bioengineering Group, NTUA

Social activism

- 2017: **Alumnus**, U.S. Department of State International Visitor Leadership Program (IVLP)
  Selected to attend 'Social & Economic entrepreneurship for Young Leaders' in recognition of founding the non-profit Mindspace
- 2017–2018: **Shaper**, Global Shapers, Athens and Geneva Hubs
  A global network of young people driving dialogue, action, and change, initiated by the World Economic Forum
- 2012–present: **Founder & CEO**, Mindspace
  A nonprofit aiming to make students entrepreneurial – bridges Balkan/U.S. ecosystems with educational trips

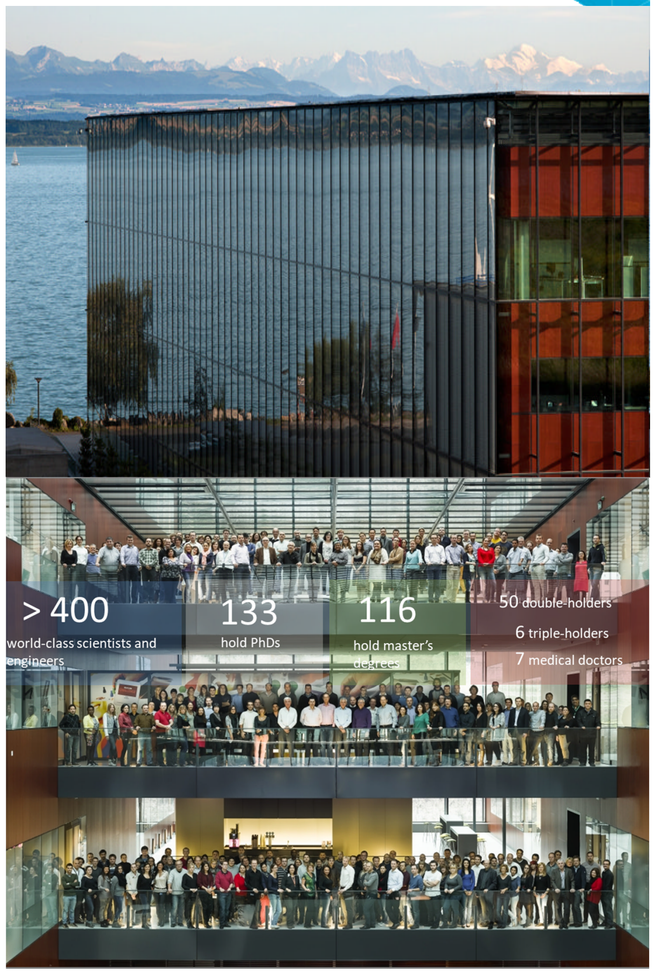
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

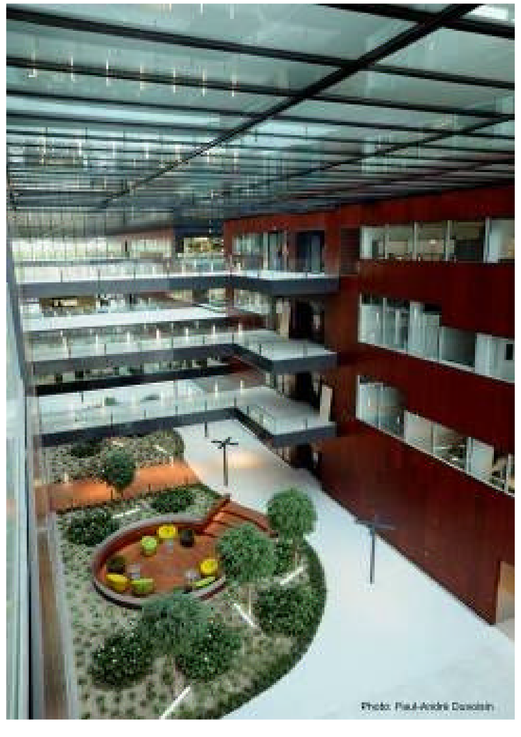
Background – PMI R&D

- Smoking causes serious diseases, such as cardiovascular disease, lung cancer, and chronic obstructive pulmonary disease.

- Philip Morris International is developing, assessing, and commercializing a number of Reduced-Risk Products* that have the potential to present less risk of harm compared with smoking cigarettes.

- Scientific determination of the reduced risk potential of these products includes comparison of the biological impact with that of a 3R4F reference cigarette on a mechanism-by-mechanism basis using robust experimental and computational methodologies.





* Reduced-Risk Products ("RRPs") is the term we use to refer to products that present, are likely to present, or have the potential to present less risk of harm to smokers who switch to these products versus continued smoking. We have a range of RRPs in various stages of development, scientific assessment, and commercialization. Because our products do not burn tobacco, they produce far lower quantities of harmful and potentially harmful compounds than found in cigarette smoke.

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv IMPROVER

A case on crowd-sourced challenges and research reproducibility

Problematic

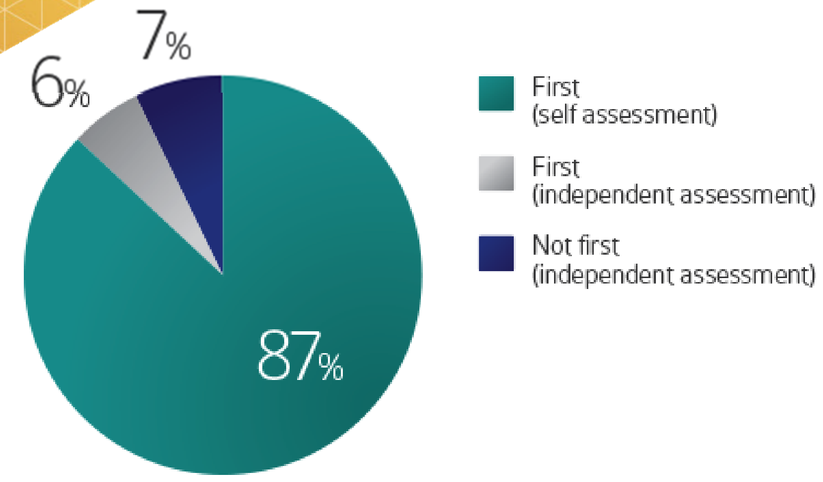
- We are experiencing a data deluge
  …but we lack the corresponding validation tools.

- A number of publications highlighted the fact that industry funding increases the likelihood that researchers will produce pro-industry conclusions and suppress the publication of negative findings (Babor & Miller, 2014).

- Bauchner and colleagues (Bauchner & Fontanarosa, 2013) describe possible ways to restore industry's credibility:
  - Give the data of industry-sponsored research to (re)analyze to academic scientists.
  - Preparation of the manuscript should primarily be the task of the academic partner, and the roles, responsibilities, contributions, and identities of all persons involved should be reported.
  - Data should be made publicly available.
  - Avoid direct-to-consumer advertising until post-marketing studies are completed.

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

The self-assessment trap



Can we all be better than average?

- Researchers wishing to publish their methods are usually required to compare their methods against others.
- Authors' method tends to be the best in an unreasonable majority of cases.
  - Selective reporting of performance: inadvertent or disingenuous
  - Choice of only one, best metric

Develop a robust methodology that verifies systems biology-based approaches

Raquel Norel et al., Mol Syst Biol. 2011 Oct 11;7:537. doi: 10.1038/msb.2011.70.

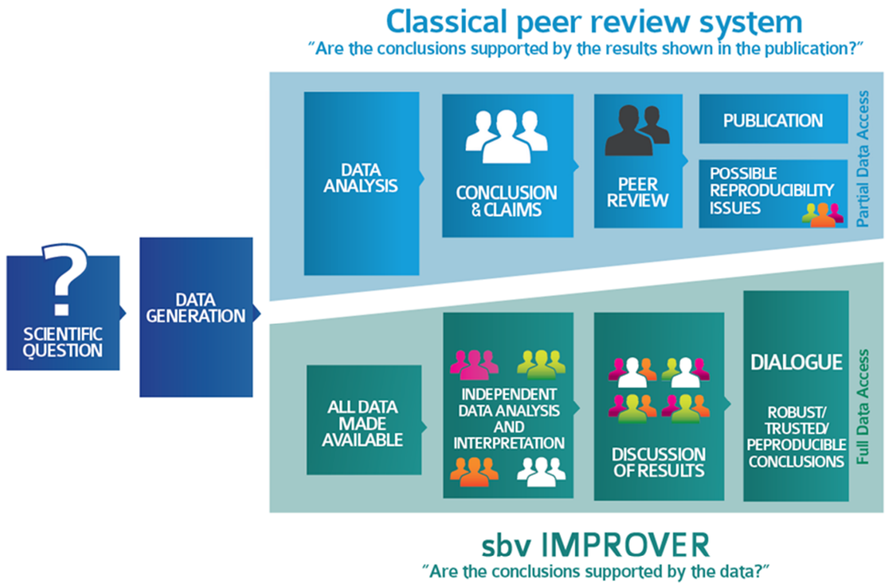
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

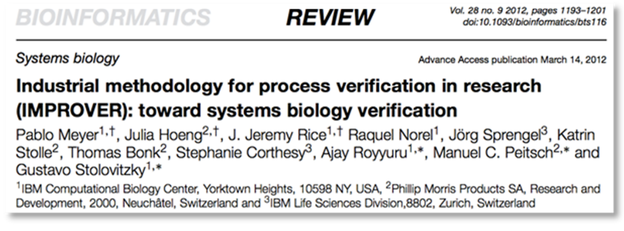
sbv IMPROVER

sbv IMPROVER stands for **S**ystems **B**iology **V**erification combined with **I**ndustrial **M**ethodology for **Pro**cess **Ve**rification in **R**esearch.

This approach aims to provide a measure of quality control of industrial research and development by verifying the methods used.

The sbv IMPROVER project is a collaborative effort led and funded by PMI Research and Development.





BIOINFORMATICS **REVIEW** Vol. 28 no. 9 2012, pages 1193–1201 doi:10.1093/bioinformatics/bts116

Systems biology Advance Access publication March 14, 2012

Industrial methodology for process verification in research (IMPROVER): toward systems biology verification

Pablo Meyer[1,†], Julia Hoeng[2,†], J. Jeremy Rice[1,†] Raquel Norel[1], Jörg Sprengel[3], Katrin Stolle[2], Thomas Bonk[2], Stephanie Corthesy[3], Ajay Royyuru[1,*], Manuel C. Peitsch[2,*] and Gustavo Stolovitzky[1,*]

[1]IBM Computational Biology Center, Yorktown Heights, 10598 NY, USA, [2]Phillip Morris Products SA, Research and Development, 2000, Neuchâtel, Switzerland and [3]IBM Life Sciences Division,8802, Zurich, Switzerland

Bioinformatics 2012 28(9):1193-1201

_computational BIOLOGY COMMENTARY

Verification of systems biology research in the age of collaborative competition

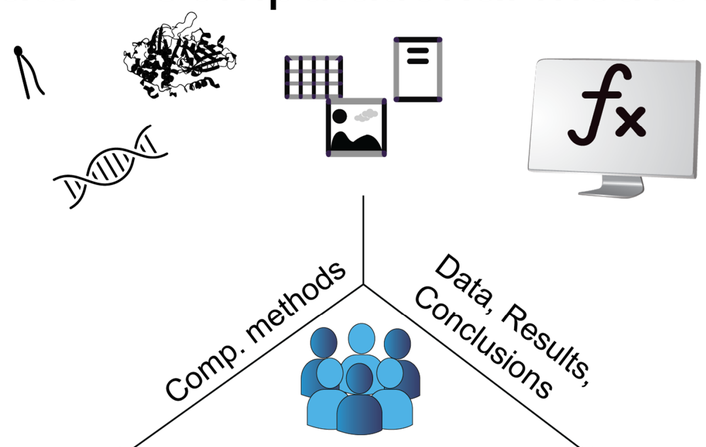
Pablo Meyer[1], Leonidas G Alexopoulos[2], Thomas Bonk[3], Andrea Califano[4], Carolyn R Cho[5], Alberto de la Fuente[6], David de Graaf[7], Alexander J Hartemink[8], Julia Hoeng[3], Nikolai V Ivanov[3], Heinz Koeppl[9], Rune Linding[10], Daniel Marbach[11], Raquel Norel[1], Manuel C Peitsch[3], J Jeremy Rice[1], Ajay Royyuru[1], Frank Schacherer[12], Joerg Sprengel[13], Katrin Stolle[3], Dennis Vitkup[4] & Gustavo Stolovitzky[1]

Nature Biotechnology 2011 Sep 8;29(9):811-5

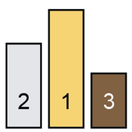
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv IMPROVER

Data + Computational methods



Computational Challenges



Benchmarking

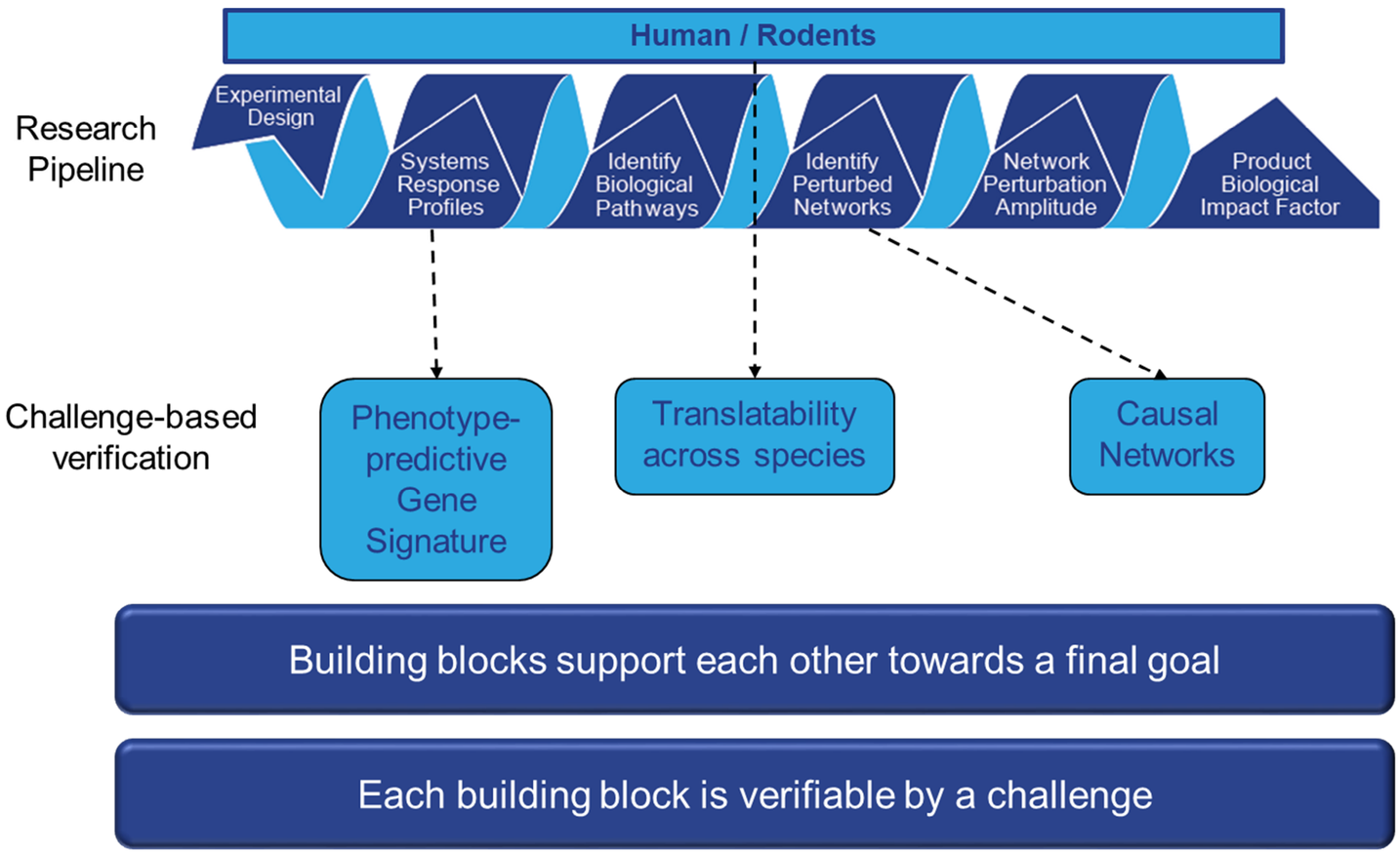
Verification

Independent Review by Panels of experts

Boue et al. Toxicological assessment of Tobacco Heating System 2.2: findings from an independent peer review. *Under review in Research Evaluation*

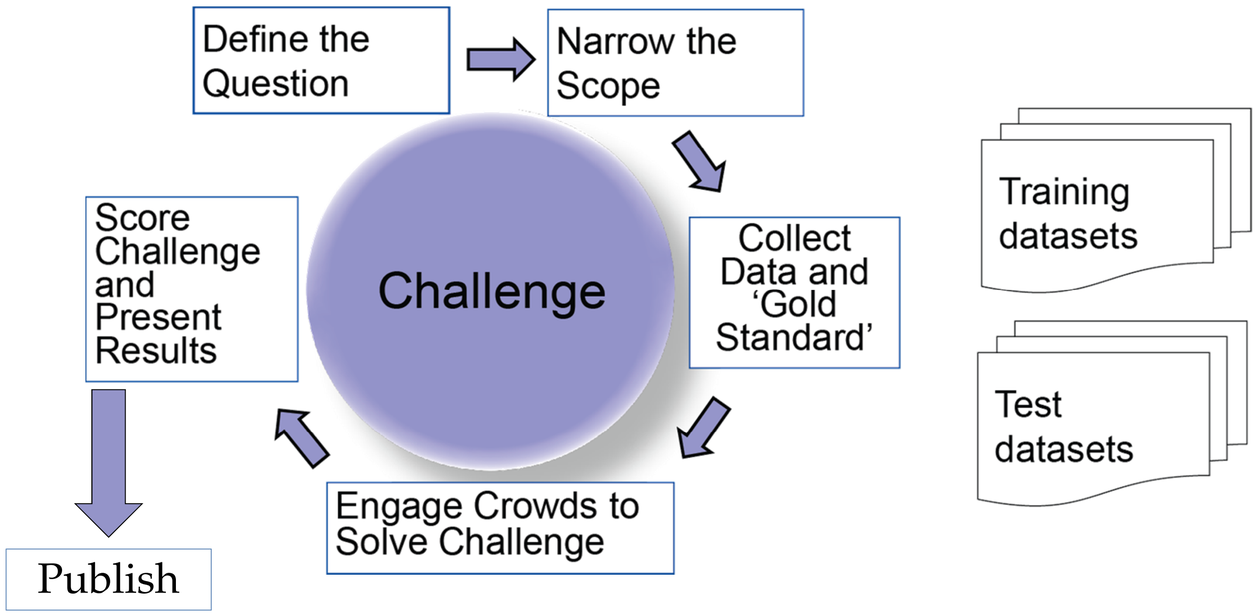
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Complex industrial research pipeline divided into verifiable building blocks



PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

How to develop a Challenge?



Double-blind performance assessment

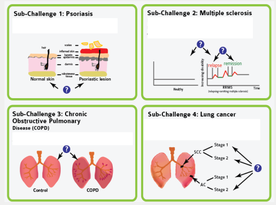
- Predefined scoring strategy approved by a Scoring Review Panel of external experts
- Scoring metrics released after the challenge closure
- Scoring of anonymized participants' submissions
- Final team ranking reviewed and approved by the Scoring Review Panel

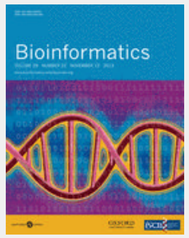
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

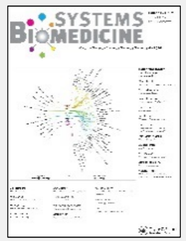
Previous sbv IMPROVER Challenges

Diagnostic Signature Challenge – 2012

To identify gene signatures for diagnostic classification in four disease areas

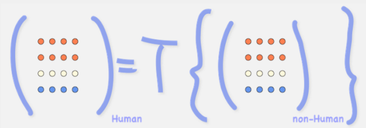






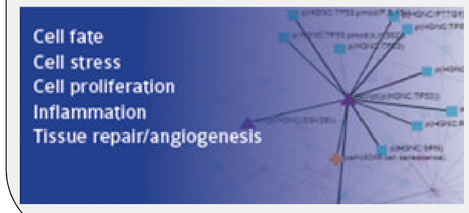
Species Translation Challenge – 2013

To identify and quantify a function of translatability of biological perturbations across human and rodent species



Network Verification Challenge – 2014-2015

To review biological network models that are suitable for drug discovery, toxicological, and mechanistic research in respiratory and cardiovascular diseases

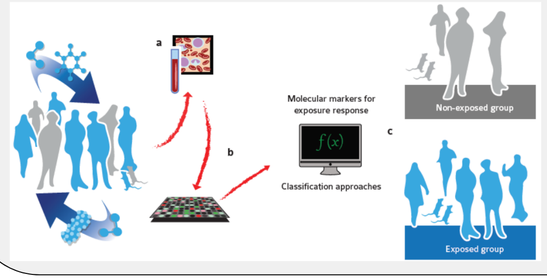


F1000Research
Open for Science

Pacific Symposium on Biocomputing 2015

Systems Toxicology Challenge – 2015-2016

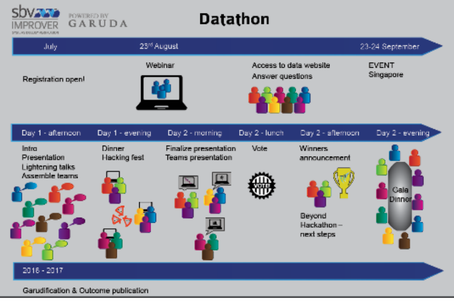
To identify robust blood-based gene signatures as predictors for smoking and cessation status



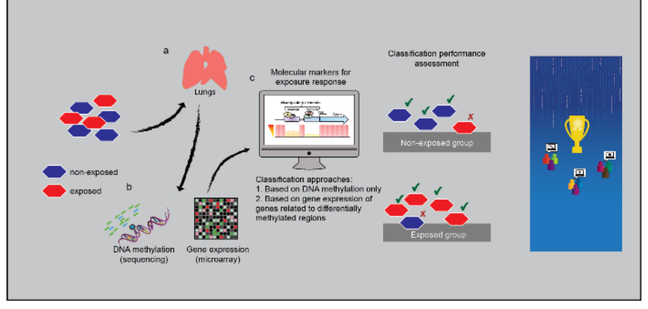
STAY TUNED!

Datathons & mini-computational challenges

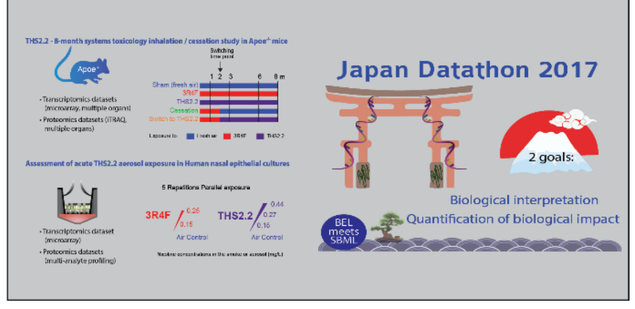
sbv IMPROVER Datathon - Singapore (Sept 2016)



sbv IMPROVER Epigenomics Challenge - Israel (Feb-May 2017)

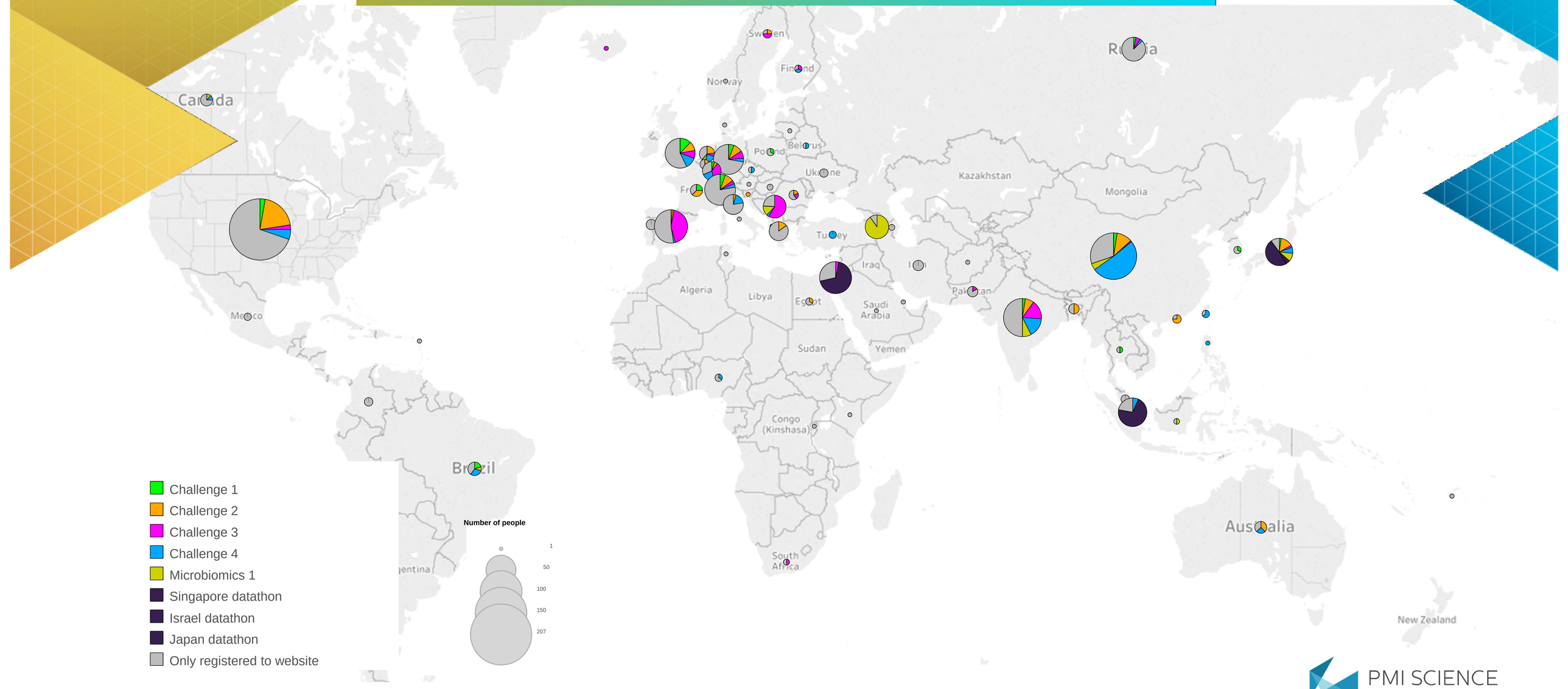


sbv IMPROVER Datathon - Japan (Oct 2017)



PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Participation map



PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv Symposia


Boston, 2012


Athens, 2013


Montreux, 2014


Orlando, 2016


Neuchâtel, 2018


Tel Aviv, 2017


Barcelona, 2015

Singapore, 2016

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Resulting publications

1. Ansari, S. *et al. On crowd-verification of biological networks. Bioinformatics and biology insights* **7 (2013)**
2. Belcastro, V. *et al. The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status. Computational Toxicology, doi:https://doi.org/10.1016/j.comtox.2017.07.004 (**2017**).*
3. Bilal, E. *et al. A crowd-sourcing approach for the construction of species-specific cell signaling networks. Bioinformatics* **31, 484-491, doi:10.1093/bioinformatics/btu659 (2015).**
4. Binder, J. *et al. in Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 270-281.*
5. Boue, S. *et al. Enhancement of COPD biological networks using a web-based collaboration interface. F1000Research* **4 (2015).**
6. Hoeng, J., Peitsch, M. C., Meyer, P. & Jurisica, I. Where are we at regarding species translation? A review of the sbv IMPROVER challenge. *Bioinformatics* **31, 451-452, doi:10.1093/bioinformatics/btv065 (2015).**
7. Meyer, P. *et al. Verification of systems biology research in the age of collaborative competition. Nature biotechnology* **29, 811-815, doi:10.1038/nbt.1968 (2011).**
8. Meyer, P. *et al. Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. Bioinformatics* **28, 1193-1201, doi:10.1093/bioinformatics/bts116 (2012).**
9. Poussin, C. *et al. Crowd-Sourced Verification of Computational Methods and Data in Systems Toxicology: A Case Study with a Heat-Not-Burn Candidate Modified Risk Tobacco Product. Chemical research in toxicology* **30, 934-945, doi:10.1021/acs.chemrestox.6b00345 (2017).**
10. Poussin, C. *et al. The species translation challenge-a systems biology perspective on human and rat bronchial epithelial cells. Scientific data* **1, 140009, doi:10.1038/sdata.2014.9 (2014).**
11. Rhrissorrakrai, K. *et al. Understanding the limits of animal models as predictors of human biology: lessons learned from the sbv IMPROVER Species Translation Challenge. Bioinformatics* **31, 471-483, doi:10.1093/bioinformatics/btu611 (2015).**
12. sbv IMPROVER project team *et al. On Crowd-verification of Biological Networks. Bioinformatics and biology insights* **7, 307-325, doi:10.4137/BBI.S12932 (2013).**
13. sbv IMPROVER project team *et al. Reputation-based collaborative network biology. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 270-281 (2015).*
14. sbv IMPROVER project team *et al. Enhancement of COPD biological networks using a web-based collaboration interface. F1000Research* **4, 32, doi:10.12688/f1000research.5984.2 (2015).**
15. sbv IMPROVER project team *et al. Community-Reviewed Biological Network Models for Toxicology and Drug Discovery Applications. Gene regulation and systems biology* **10, 51-66, doi:10.4137/GRSB.S39076 (2016).**
16. Tarca, A. L. *et al. Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. Bioinformatics* **29, 2892-2899, doi:10.1093/bioinformatics/btt492 (2013).**

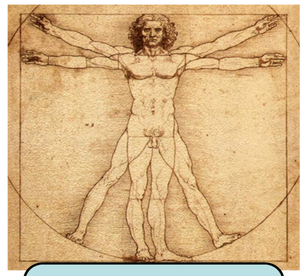
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

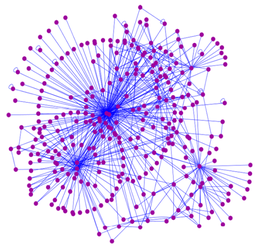
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Case study I

Species Translation Challenge

Background and goal

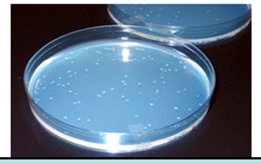


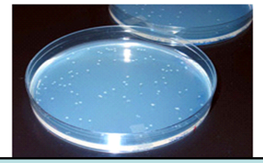


Rat

Human

Predict human impact and then validates with human data





Rat cellular model

Human cellular model

Concept of
*« **Translatability** »*

Goal: Verify that a mapping exists and allows the translation of biological effects of stimulus-induced perturbations in one species given information about the same perturbations in another species.

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

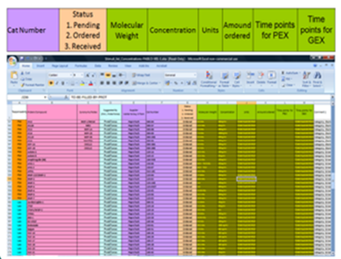
Overall experimental workflow

Step 1

Culture Cells Human/Rat

Step 2

In-silico Screening

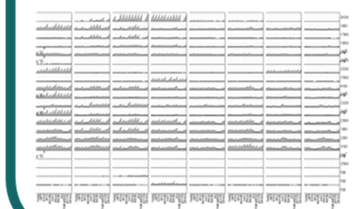


Step 3

Validation of RNA & Protein assays

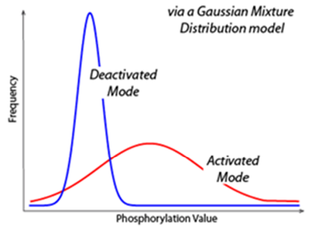
Step 4

Experimental Screen of Compounds



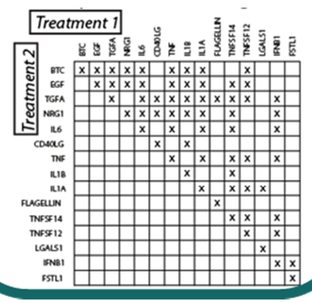
Step 5

Compound Selection



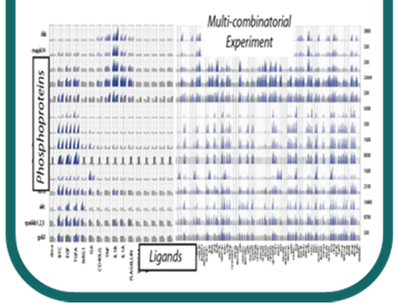
Step 6

Experimental Design



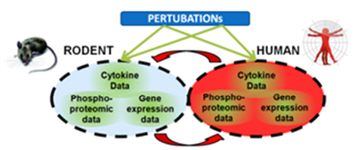
Step 7

Data Acquisition



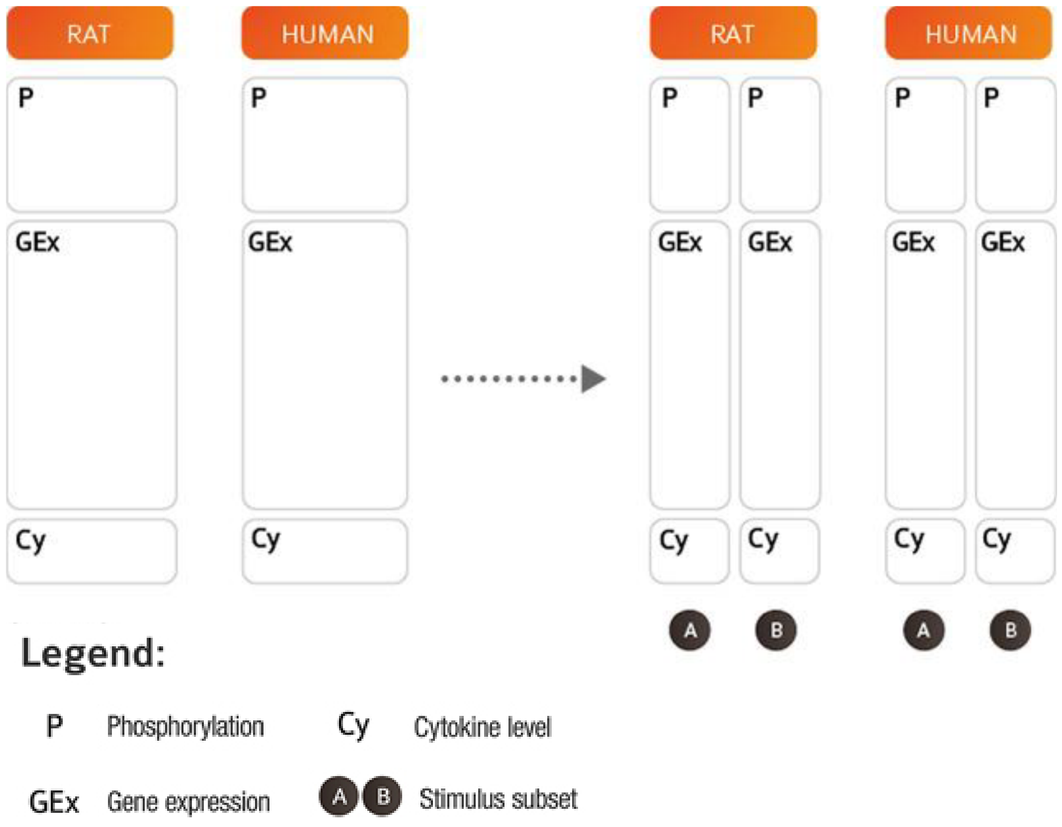
Step 8

Data Analysis Species Comparison



PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Data compendium



2 species: human and rat

52 stimuli

Phospho-proteomics data (~10,000 data points)
~16 proteins
2 time points: 5 min and 25 min
3 biological replicates

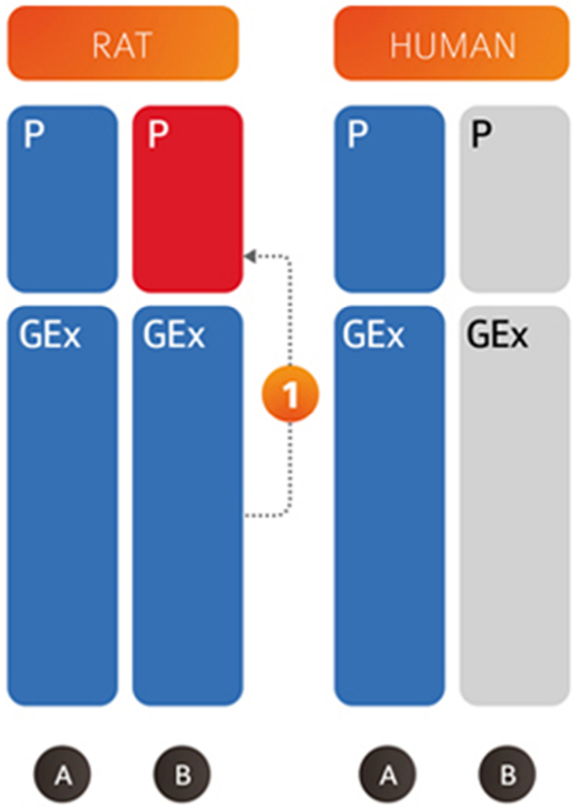
Gene expression data (> 300 Cel files)
~20,000 (human) and ~19,000 (rat) genes
1 time point: 6 h
3 biological replicates

Cytokine level data (~7,000 data points)
~22 proteins
1 time point: 24 h
3 biological replicates

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Sub-Challenge 1

Intra-species protein phosphorylation prediction



- Predict the protein phosphorylation status for each stimulus in Subset B of rat, from the corresponding gene expression information.
- Question:
  - Is gene expression data sufficiently informative to infer the phosphorylation status through a backward inference process?

Legend:

P	Phosphorylation		Not provided
GEx	Gene expression		Provided data
A B	Stimulus subset		Predicted data

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Sub-Challenge 2

Inter-species protein phosphorylation prediction

- Predict the protein phosphorylation status for each stimulus in subset B in human from the protein phosphorylation status for the same stimulus in subset B in rat.

- Question:
  - Are gene expression and phosphorylation data in one species sufficiently informative to infer the phosphorylation status in another species?

Legend:

P	Phosphorylation
GEx	Gene expression
A B	Stimulus subset

- Not provided
- Provided data
- Predicted data
- Provided after 1 July

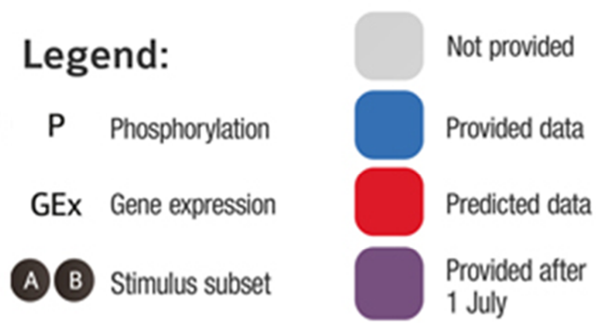
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Sub-Challenge 3

Inter-species pathway perturbation prediction

- Predict the gene sets representative of pathways/biological processes that are the most to least enriched among differentially expressed genes with respect to control for each stimulus in Subset B in human based on the corresponding data in rat.

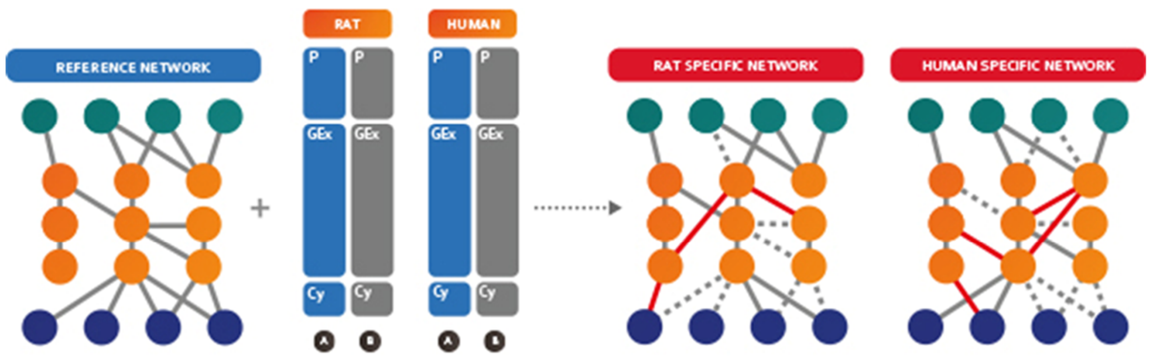
- Question:
  - Can the perturbation of pathways be predicted in human from equivalent information in rat?



PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Sub-Challenge 4

Species-specific network inference



Legend:

P	Phosphorylation		Not to be used		Edges
GEx	Gene expression		Provided data		Removed edges
Cy	Cytokine level		Inferred network		Added edges
		A B	Stimulus subset		

- The goal is to infer human and rat networks given phosphoprotein, gene expression and cytokine data and a reference map provided as prior knowledge. Participants will use network inference to add or remove edges from the reference map to produce specific rat and human networks.
- Question:
  - Can biological networks be built by leveraging diverse ‘omics’ data to assess the commonalities and differences between the species?

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Species Translation Challenge outcome

To learn more about the outcome of the Species Translation Challenge, the following articles have been published in the Bioinformatics Journal:

- **Understanding the limits of animal models as predictors of human biology: lessons learned from the sbv IMPROVER Species Translation Challenge**
  (Bioinformatics Overarching Paper, 17 September 2014)
- **Inter-Species Pathway Perturbation Prediction via Data Driven Detection of Functional Homology**
  (Bioinformatics, 4 August 2014)
- **Predicting protein phosphorylation from gene expression: Top methods from the IMPROVER Species Translation Challenge**
  (Bioinformatics, 23 July 2014)
- **Inter-species prediction of protein phosphorylation in the sbv IMPROVER Species Translation Challenge**
  (Bioinformatics, 3 July 2014)

To learn more about the data set used during the Species Translation Challenge, the following article has been published in Scientific Data:

- **The species translation challenge - A systems biology perspective on human and rat bronchial epithelial cells.**
  (Scientific Data, 10 June 2014)

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Case study II

Systems Biology Approach for Compounds Mode of Action Discovery

Introduction

- Compounds derived from plants, such as acetylsalicylic acid, morphine, digitoxin, and quinine, have shown a plethora of beneficial therapeutic anti-inflammatory, analgesic, cardiotonic, and anti-malaria effects, respectively, and are the most successful source of potential drug leads in pharmaceutical research [1].

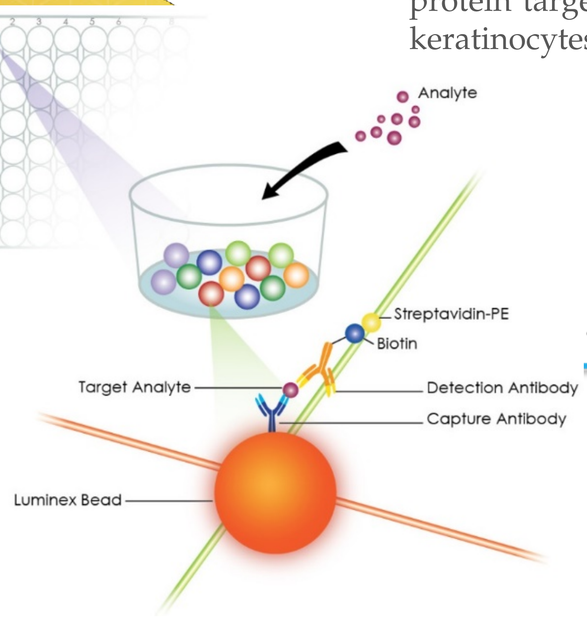
- However, the elucidation of their mechanism(s) of action can be challenging, because these compounds can bind multiple protein targets with unrelated structures [2]. A systems biology approach that integrates numerous cellular data layer components may help to link downstream effect(s) with the cascade(s) of signaling molecular changes in response to a natural compound.

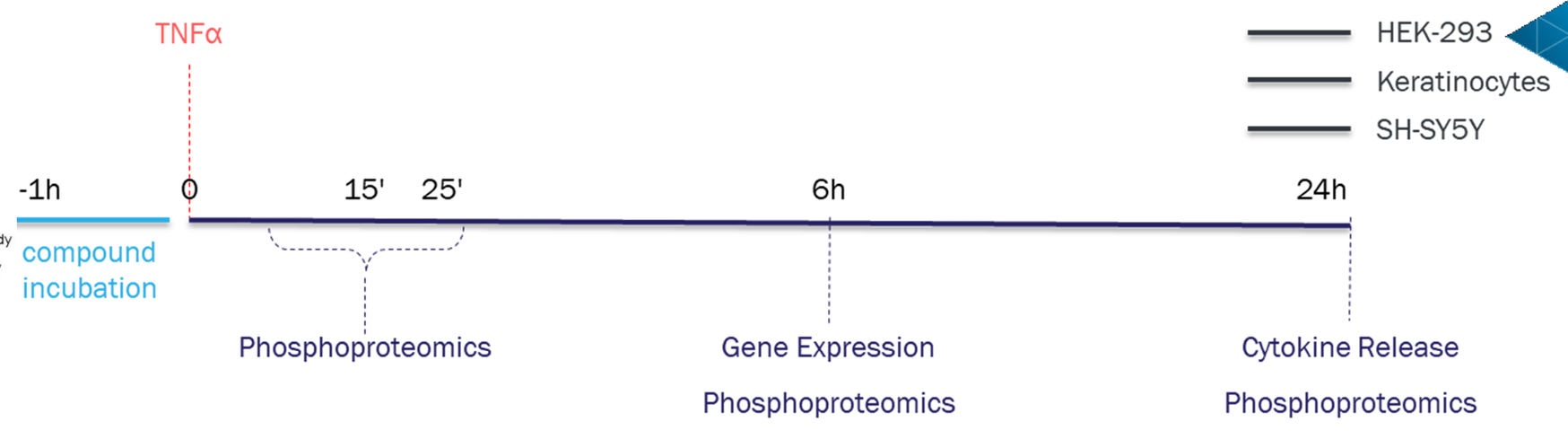
1. Dias, D.A., Urban, S. & Roessner, U. Metabolites 2, 303-336 (2012).
2. Breinbauer, R., Vetter, I.R. & Waldmann, H. Angew Chem Int Ed Engl 41, 2879-2890 (2002).

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Generation of multiomics datasets

- To investigate the mode of action of a natural compound (e.g., anti-oxidant, anti-inflammation) on pre-activated cells with an inflammatory stimulus, a large dataset including 18 phosphoproteomic targets, transcriptomics, and 35 soluble protein targets (e.g., cytokines and chemokines) have been generated in 3 cell systems (HEK-293 cells, primary human keratinocytes, and SH-SY5Y cells) and 4 exposure time points (15 min, 25 min, 6 h, 24 h).





	Transcriptomics	Phosphoproteomics/secreted proteins
Measurement	Affymetrix GeneChip® Human Genome U133 Plus 2.0 Array (~20k genes)	Luminex (bead-based antibody multiplexed assay)
Normalization	Frozen robust multiarray analysis[1]	-
Differential Expression	limma R package [2]	stats R package (Author: R Core Team and contributors worldwide)

Phosphoprotein targets	
AKT S473	MEK1 S218/S222
cJun S63	mTOR S2448
CREB S133	NFkB S536
ERK1 T202/Y204	p38MAPK T180/Y182
GSK3AB S21/S9	p53 S15
Hsp27 S78/S82	p70S6K T389
IKBa S32/S36	p90RSK S380
JNK T183/Y185	PRAS40 T246
MARCKS S170	STAT3 Y705

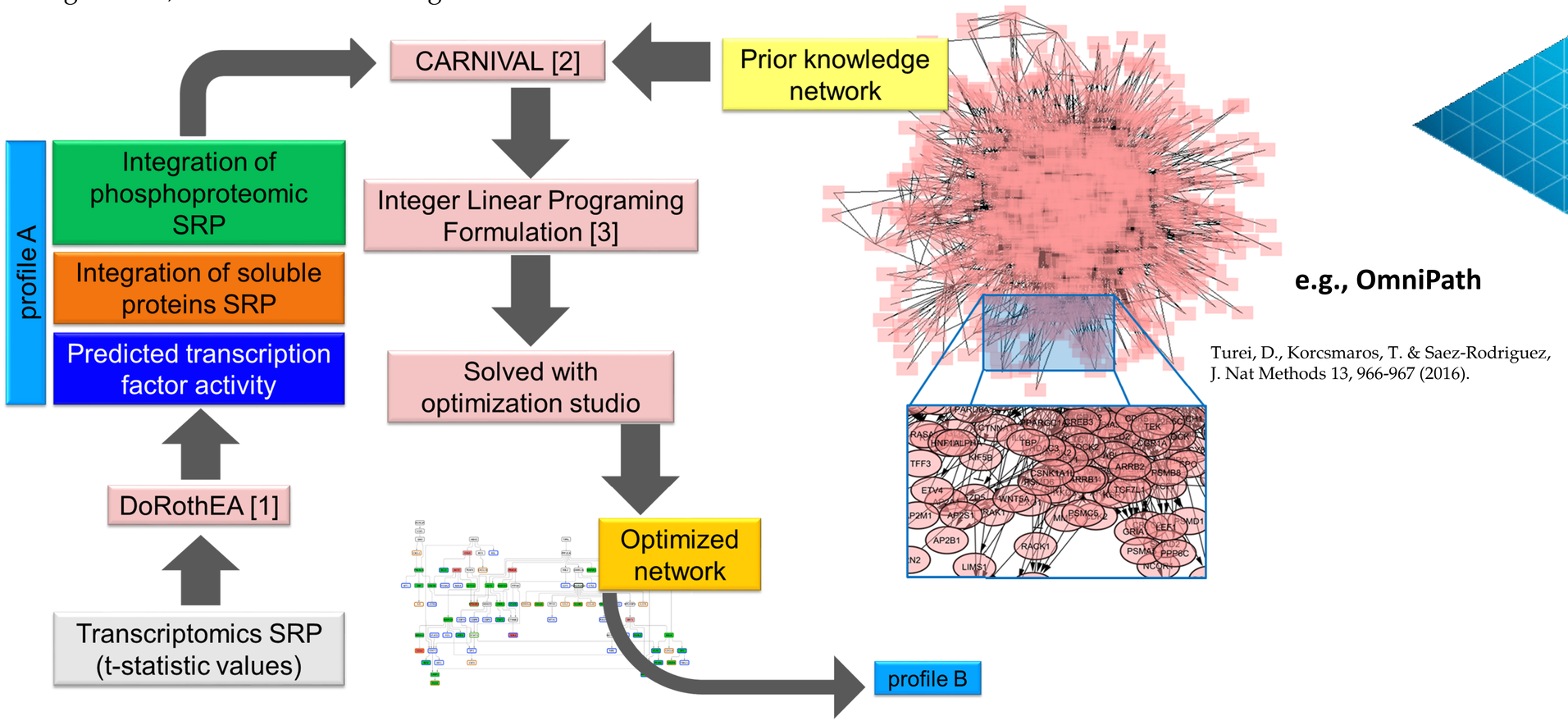
Cytokine/Chemokine targets				
CCL11	CSF3	IL10	IL1RA	IL9
CCL2	CXL10	IL12A	IL2	TGFA
CCL22	EGF	IL12B	IL3	TNFB
CCL3	FGF2	IL13	IL4	VEGFA
CCL4	FLT3L	IL15	IL5	X3CL1
CCL7	GROA	IL17	IL6	
CD40L	IFNA2	IL1A	IL7	
CSF2	IFNG	IL1B	IL8	

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

1 McCall MN et al. 2010 Biostatistics. Apr;11(2):242-53. doi: 10.1093/biostatistics/kxp059. Epub 2010 Jan 22.
2 Ritchie ME et al. 2015 Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.

Computational network inference

- These various data types can be integrated in an optimized inferred network representative of the mode of action using a computational approach robustly combining backward reasoning, integer linear programming (ILP) optimizer algorithms, and curated knowledge networks.



e.g., tumor necrosis factor alpha (TNFα) vs. Medium control systems response profile (SRP)

CARNIVAL tool website: https://saezlab.github.io/CARNIVAL/

DoRothEA tool website: https://dorothea.opentargets.io/#/

1. Garcia-Alonso, L. et al. Cancer Res 78, 769-780 (2018).
2. Liu, A. et al. bioRxiv, 541888 (2019).
3. Mitsos, A. et al. PLoS Comput Biol 5, e1000591 (2009).

ILP

ILP formulation

Objective function: $min \sum_{j,k} \left| x_j^k - x_j^{k,m} \right|$ j=1,..,$n_{species}$ and k =1,..,$n_{experiments}$

Secondary objective function: $min \sum_i y_i$ i=1,..,$n_{reactions}$

CONSTRAINTS

$$z_i^k \le y_i, \quad i = 1,\ldots,n_r, \quad k = 1,\ldots,n_e$$

$$z_i^k \le x_j^k, \quad i = 1,\ldots,n_r, \quad k = 1,\ldots,n_e, \quad j \in \mathsf{R}_i$$

$$z_i^k \le 1 - x_j^k, \quad i = 1,\ldots,n_r, \quad k = 1,\ldots,n_e, \quad j \in \mathsf{I}_i$$

$$z_i^k \ge y_i + \sum_{j \in \mathsf{R}_i} \left(x_j^k - 1\right) - \sum_{j \in \mathsf{I}_i} \left(x_j^k\right), \quad i = 1,\ldots,n_r, \quad k = 1,\ldots,n_e$$

$$x_j^k \ge z_i^k, \quad i = 1,\ldots,n_r, \quad k = 1,\ldots,n_e, \quad j \in \mathsf{P}_i$$

$$x_j^k \le \sum_{i=1,\ldots,n_r : j \in \mathsf{P}_i} z_i^k, \quad j = 1,\ldots,n_s, \quad k = 1,\ldots,n_e$$

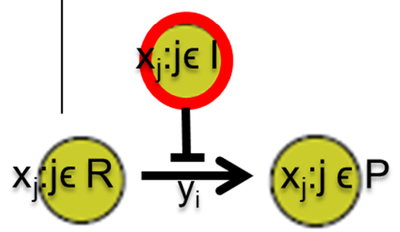
Solved with standard solver: ILOG CPLEX

VARIABLES

y_i Reaction i possible

z_i^k Reaction i occurs in experiment k

x_j^k Species j formed in experiment k

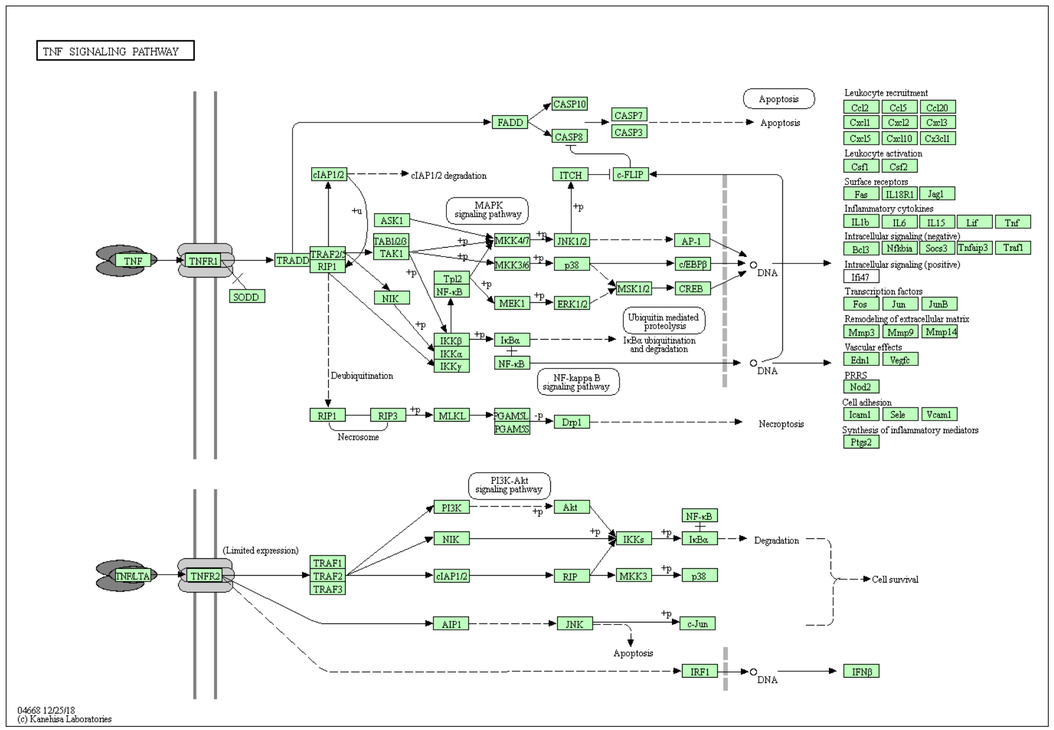


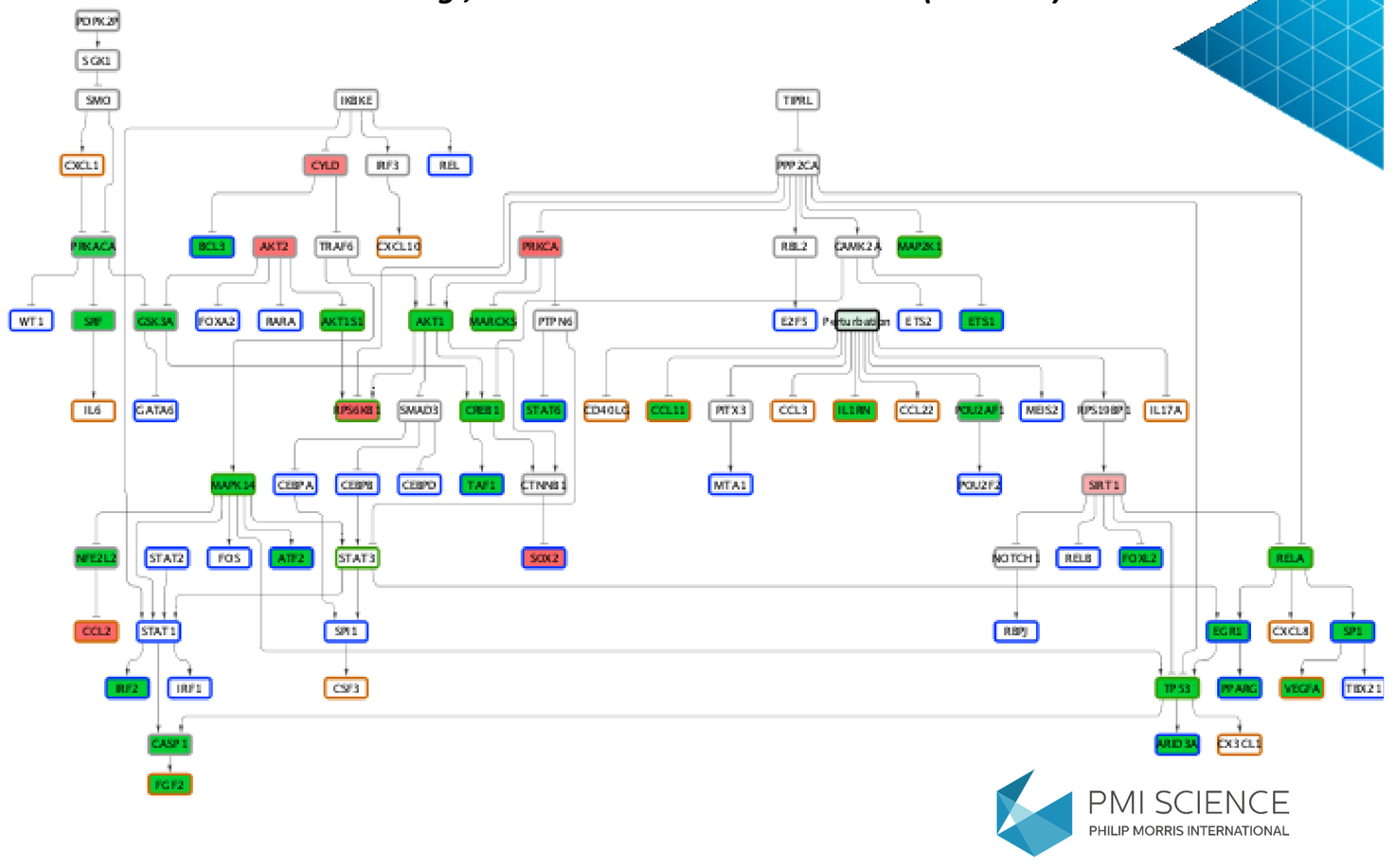
Mitsos, A. et al. PLoS Comput Biol 5, e1000591 (2009).

Inferred network

As a case study to show the relevance of the approach, we have applied this computational pipeline on transcriptomic, phosphoproteomic, and soluble protein data from SH-SY5Y cells treated with TNFα (TNFα vs Medium control SRP), for which the mechanism of action is known.

e.g., TNFα vs Medium Control SRP (SH-SY5Y)





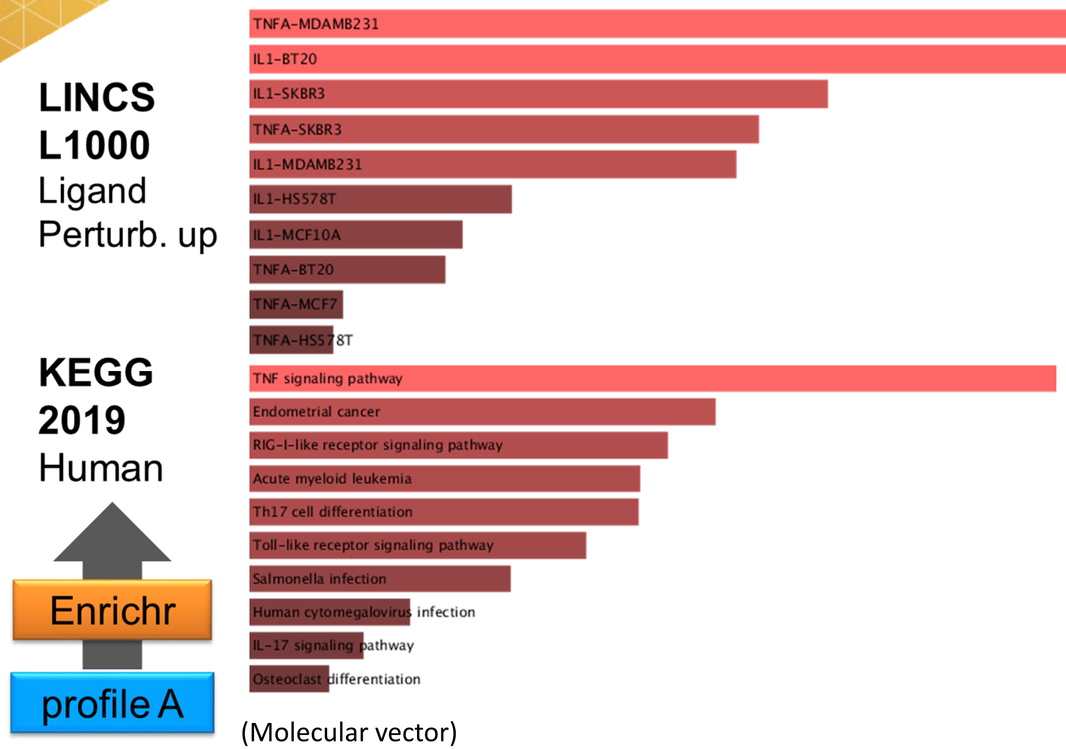
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

KEGG pathways, TNF signaling pathway

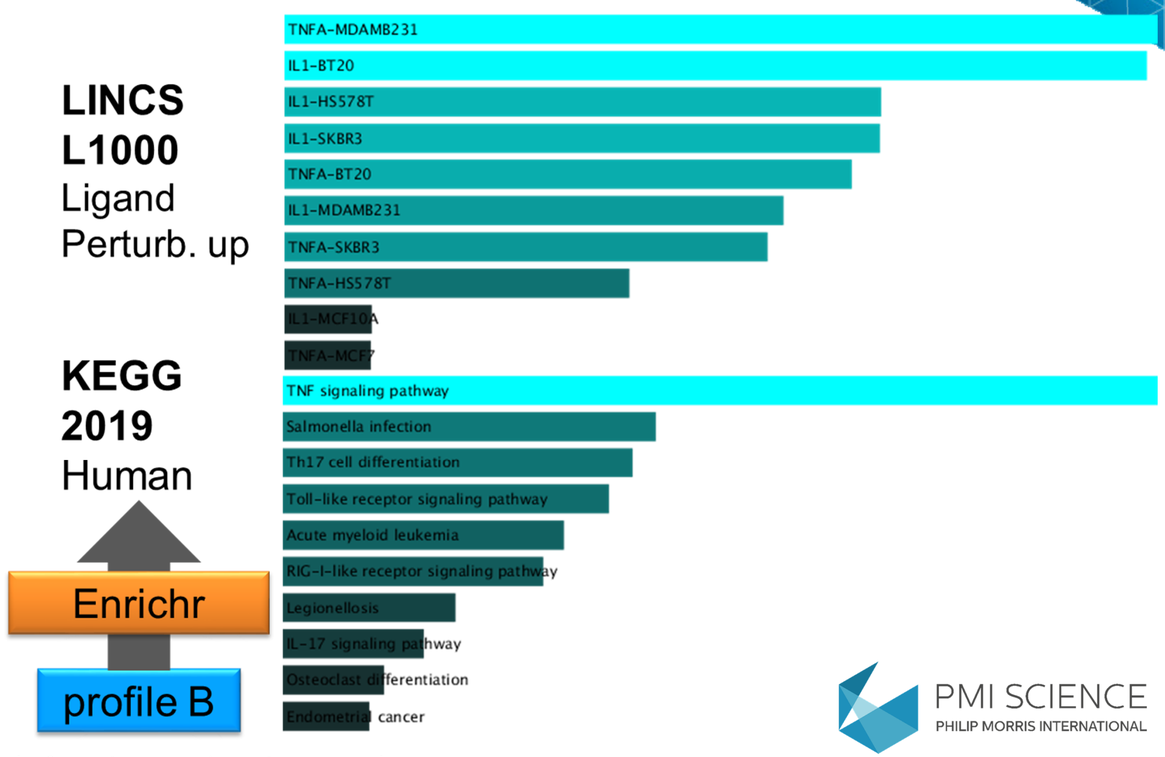
Mode of action investigation

The enrichment analysis conducted with Enrichr [1] and using KEGG and LINCS databases as sources of *a priori* pathway- and ligand-based gene/molecule sets highlights the fact that the molecular profile A and profile B, corresponding to nodes of the optimized network, are enriched in a set of molecules known to be activated by TNFα or interleukin 1β.

e.g., TNFα vs Medium Control SRP (SH-SY5Y)



(Molecular vector)



(Inferred network nodes vector)

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

1. Chen, E.Y. et al. BMC Bioinformatics 14, 128 (2013).

Conclusions

- A large dataset, including 18 phosphoproteomic targets, transcriptomics, and 35 soluble protein targets (e.g., cytokines and chemokines), have been generated in 3 cell systems (HEK-293 cells, primary human keratinocytes, and SH-SY5Y cells) and 4 exposure time points (15 min, 25 min, 6 h, 24 h).

- These various data types can be integrated in an optimized inferred network representative of the mode of action using a computational approach robustly combining backward reasoning, ILP optimizer algorithms, and curated knowledge networks.

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Next Challenge

The Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge

Background

- Inflammatory bowel diseases (IBD) constitute a spectrum of chronic inflammatory disorders that recurrently affect the gastrointestinal tract of millions of patients worldwide (1, 2).

- Ulcerative colitis (UC) and Crohn's disease (CD) are the two main clinically defined manifestations of IBD, each with distinctive clinical and pathological features (1).

- Endoscopy constitutes the gold standard for the diagnosis and monitoring of IBD. The diagnosis is usually confirmed by biopsies on colonoscopy and complemented with the measurement of clinical molecular biomarkers. However, their low sensitivity and high variability characteristics limit clinical efficacy (1).

- Thus, there is a need to identify novel molecular biomarkers that could be assessed with less-invasive methods.

- The link between pathogenesis of IBD and the intestinal microbiota has been established. Evidence points out that microbiome disequilibrium (dysbiosis) may cause an inappropriate immune response that results in alteration of the intestinal epithelium barrier integrity (3).

- In this new challenge, we will investigate the diagnostic potential of metagenomics data to discriminate patients with IBD from non-IBD subjects. Furthermore, within the IBD category, it will attempt to separate UC and CD subjects.

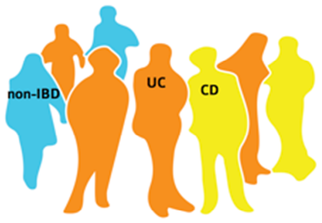
(1) Titz et al., Int J Mol Sci, 19, 2018
(2) Ng et al., The Lancet, 390, 2017
(3) Maloy et al., Nature 474, 2011

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

The Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge (MEDIC)

CHALLENGE DATA

Shotgun metagenomics sequencing data from faecal samples



POPULATIONS
- non-IBD
- IBD (UC and CD)

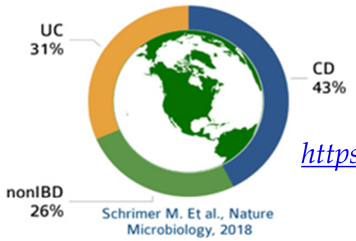
Sub-challenge 1

Sub-challenge 2

RAW FASTQ

Process raw data with your own pipeline

Training dataset



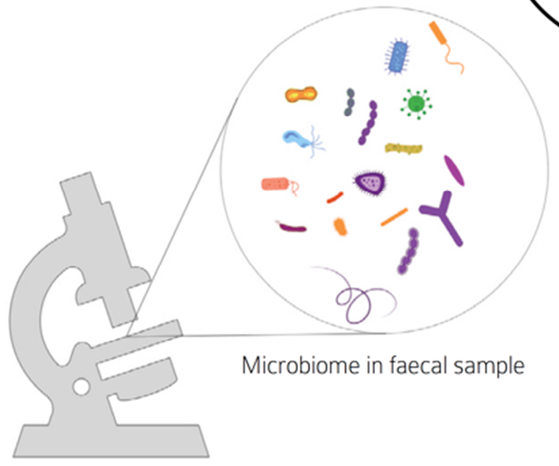
Schrimer M. Et al., Nature Microbiology, 2018

https://ibdmdb.org/

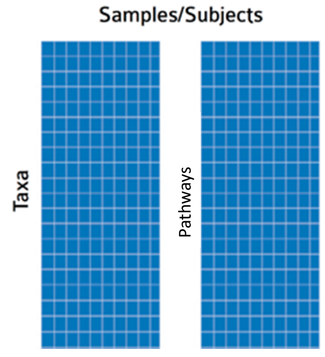
Free to use other publicly available datasets or your private datasets

He Q. et al., GigaScience 2017

http://gigadb.org/



TAXONOMY AND PATHWAY ABUNDANCES



MACHINE LEARNING

Build classifiers for 2-class problems:
- IBD vs non-IBD
- UC vs non-IBD
- CD vs non-IBD
- UC vs CD

+

Test dataset

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv IMPROVER MEDIC – scientific questions

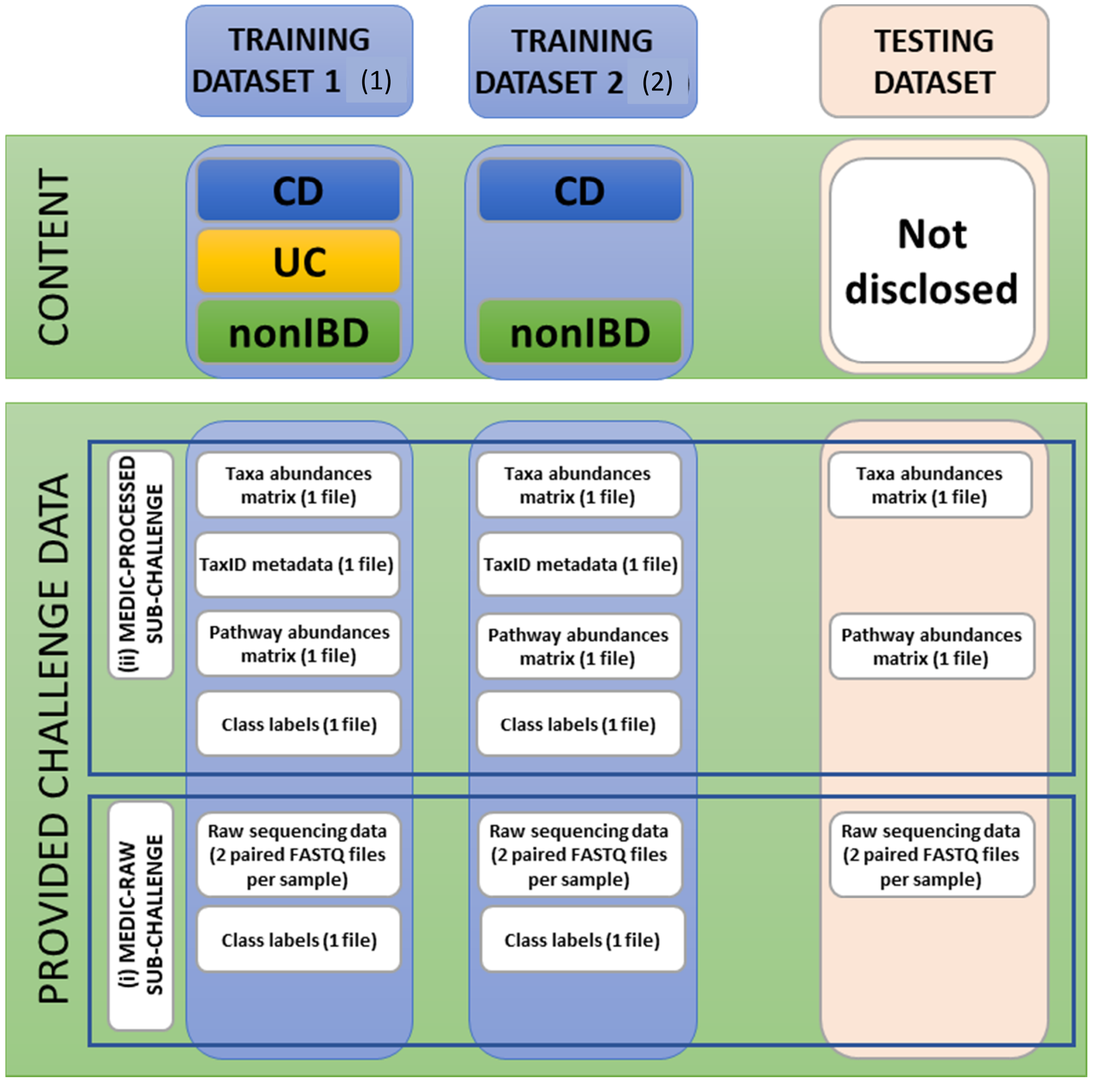
- Which predictive computational approaches are the most accurate across the following four 2-class problems ?
  - IBD vs. non-IBD
  - UC vs. non-IBD
  - CD vs. non-IBD
  - UC vs. CD

- What do the most discriminative metagenomic features tell us?
  - Are they based on taxonomy, functions/pathways, and/or other types (e.g., k-mers)?

  - Are they distinct between UC vs. non-IBD and CD vs. non-IBD, or do they show commonalities?

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv IMPROVER MEDIC – datasets



(1) Schirmer et al., Nat Microbiol 3, 2018
(2) He et al., GigaScience 6, 2017

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv IMPROVER MEDIC – timelines

www.sbvimprover.com/challenge-5

June 2019 — October 2019 — November 2019

Challenge opens	Submission of predictions	Challenge closes	Scoring	Team ranking	Winners announcement	Drafting publication

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv IMPROVER MEDIC – scoring

Double-blind performance assessment

- Predefined scoring strategy approved by a Scoring Review Panel of external experts
- Scoring metrics released after the challenge closure
- Scoring of anonymized participants' submissions
- Final team ranking reviewed and approved by the Scoring Review Panel

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

sbv IMPROVER MEDIC – why participate?

SHOWCASE	Show your data science skills.
COMPETE	6 × 2.000 $ prizes *
NETWORK	Grow your professional network.
BENCHMARK	Receive an independent assesment of your methods.
PUBLISH	Co-author scientific article(s) describing the outcome of the challenge.
COLLABORATE	Have fun working with others.

Be part of the Journey!

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

***Disclaimer:** Please see the conditions for winning prizes and the rules of participation in the MEDIC at www.sbvimprover.com*

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

www.sbvimprover.com

Questions? Contact us:
sbvimprover.RD@pmi.com

Thank you!

Contact: Dimitris.Messinis@contracted.pmi.com

The sbv IMPROVER project, the websites and the Symposia are part of a collaborative project designed to enable scientists to learn about and contribute to the development of a new crowd sourcing method for verification of scientific data and results. The project is led and funded by Philip Morris International.
For more information on the focus of Philip Morris International's research, please visit www.pmiscience.com