Biomedical Discovery Informatics Using Knowledge Graphs

Vít Nováček, PhD

April 24th, 2019

Outline

- Institute / group overview
- Knowledge graphs
- Biomedical discovery informatics
  - Signalling prediction
  - Drug target prediction
  - Ultimate goal

Outline

- Institute / group overview
- Knowledge graphs
- Biomedical discovery informatics
  - Signalling prediction
  - Drug target prediction
  - Ultimate goal

DSI @ NUI Galway

- **Data Science Institute**
  - Formerly DERI (2003-2013), a **leading Semantic Web institute** directed by Stefan Decker (formerly of Stanford)
  - Founding member of **Insight**, a €75M+ **national research centre** for **data analytics**
  - Part of **National University of Ireland Galway** (https://www.nuigalway.ie/)
  - For details, see https://datascienceinstitute.ie/, https://insight-centre.org/
- **Research topics** covered
  - AI, Machine Learning, Linked Data, NLP/Text Mining, Recommender Systems, IoT, ...
- **Verticals** covered
  - Healthcare, Financial, Green IT, ...

Vít’s Group at DSI

- **Basic research** on **knowledge graphs** (KGs)
  - *Regularizing Knowledge Graph Embeddings via Equivalence and Inversion Axioms*. In ECML/PKDD, 2017 (https://doi.org/10.1007/978-3-319-71249-9_40)
- Straightforward **biomedical applications of KGs**
  - *Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models*. In Briefings in Bioinformatics, 2019 (https://doi.org/10.1093/bib/bbx099)
- **Link prediction** for **systems biology** and **drug discovery**
  - See the next slides
- **Clinical applications** of **KG embeddings** and **explainable AI**
  - See the next slides

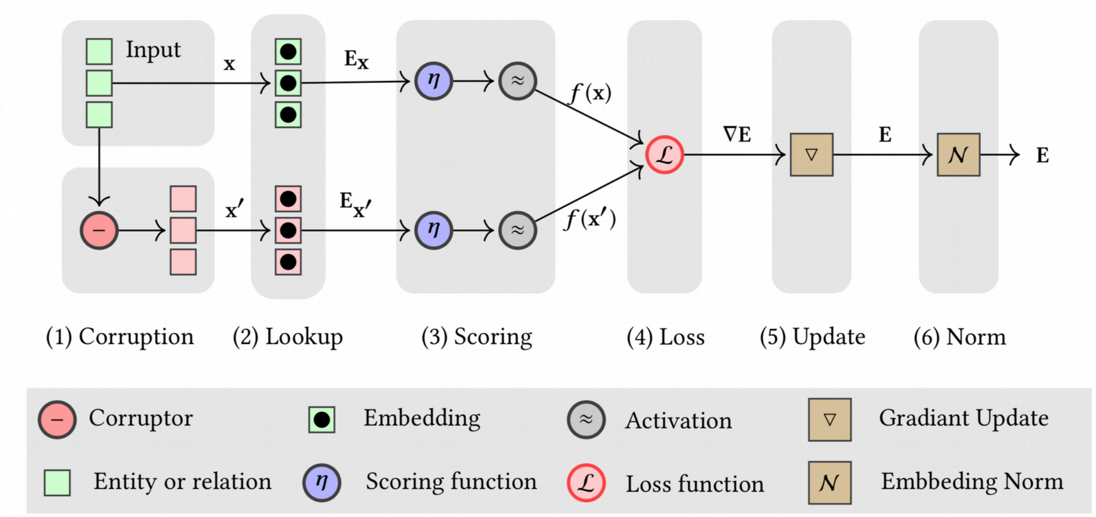
Outline

- Institute / group overview
- Knowledge graphs
- Biomedical discovery informatics
  - Signalling prediction
  - Drug target prediction
  - Ultimate goal

Knowledge Graphs

- A powerful way to organise descriptions of properties of **objects** and their **connections**
- The “**Semantic Web** done right”
  - **Lightweight** knowledge representation formalism
  - Suitable for many **domains** and **use cases**
  - Straighforward **automated population** and **knowledge integration**
  - Rather **complex inferences** possible
    - Link prediction and knowledge base completion
    - Relation extraction
    - Analogical reasoning
    - FOL / DL axioms can be incorporated to some extent
  - **Scalable algorithms** taking advantage of the most recent **AI developments**

Knowledge Graph Embeddings



- **Supervised machine learning** problem
- Falls under **statistical relational learning**
  - Effectively, fitting a **multivariate** probability density function to the **positive** and **negative** “links” (i.e. *subject-predicate-object* triples) in the knowledge graph
  - **Negatives** typically generated as **corruptions** of positives
    - Fixing *subject-predicate*, generating random *objects*, or the other way around

Outline

- Institute / group overview
- Knowledge graphs
- Biomedical discovery informatics
  - Signalling prediction
  - Drug target prediction
  - Ultimate goal

Opportunities for AI / KGs in Life Sciences and Healthcare

- **AI** has not seen much direct application in **healthcare** (with few rather experimental exceptions like the expert systems of old)
- The tides may be changing, though
- The **deep learning hype** is largely responsible
  - Image analysis for super-human diagnostics (***DSI active in the domain***)
  - Large-scale analysis of patterns in experimental omics data (***DSI active in the domain***)
  - Prediction of depression based on social network data analysis (***DSI active in the domain***)
- But it's not only about that
  - Biomedicine comes with **wealth of curated, highly expressive network data** that are barely ever processed
  - **EHRs largely untapped** due to lack of text mining solutions integrated into reliable predictive models
  - **Knowledge graph techniques** can be the next big thing here (***DSI has some pieces of world-first technology here***)

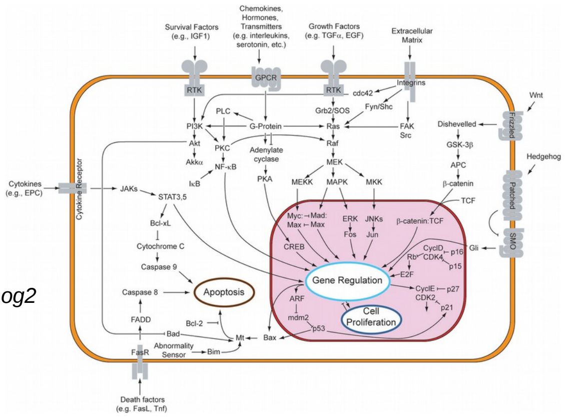
- The **biggest challenges** at the moment
  - **Explainable AI** much needed (***DSI active in the domain***)
  - The field would tremendously benefit from much **more communication** between biologists, clinicians, pharma experts and computer scientists to **inform novel models** that inherently address the challenges of current biomedicine (***DSI paving the way here with some recent research***)
  - Deep learning may not be the best for **clinical decision support** (related to the above points) - the biomedical field may need to trigger a **paradigm shift in the AI** itself
  - New **healthcare policies** are required to use AI in a **safe**, **ethical** and **privacy-preserving** manner

Outline

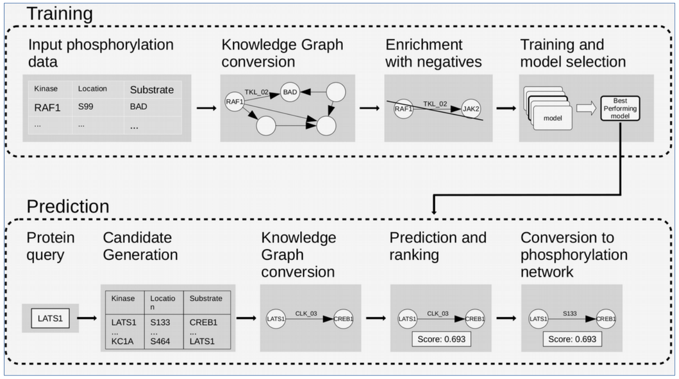
- Institute / group overview
- Knowledge graphs
- Biomedical discovery informatics
  - Signalling prediction
  - Drug target prediction
  - Ultimate goal

Signalling Prediction – Outline of the Problem and Solution

- **Problem**
  - Many **diseases** are associated with **dysregulated cellular signalling** (e.g. cancer or neurodegenerative disorders)
  - **Making sense of signalling** is a hard, expensive and time-consuming **biological problem**
  - **Computational predictions** accelerate **research** aiming at **evidence-based therapies**
  - Current systems struggle with **low accuracy** and **limited proteome coverage**, though
- **Solution**
  - Representing **phosphorylation signalling** data as **knowledge graphs** (KGs)
  - Training **statistical relational models** on the KGs to
    - Be able to make predictions on **any protein** in the input data
    - **Increase prediction accuracy** by taking the **latent features** of **signalling networks** into account



The cell signalling image was created by Boghog2 at English Wikipedia - Transferred from en.wikipedia to Commons., Public Domain, https://commons.wikimedia.org/w/index.php?curid=4851717

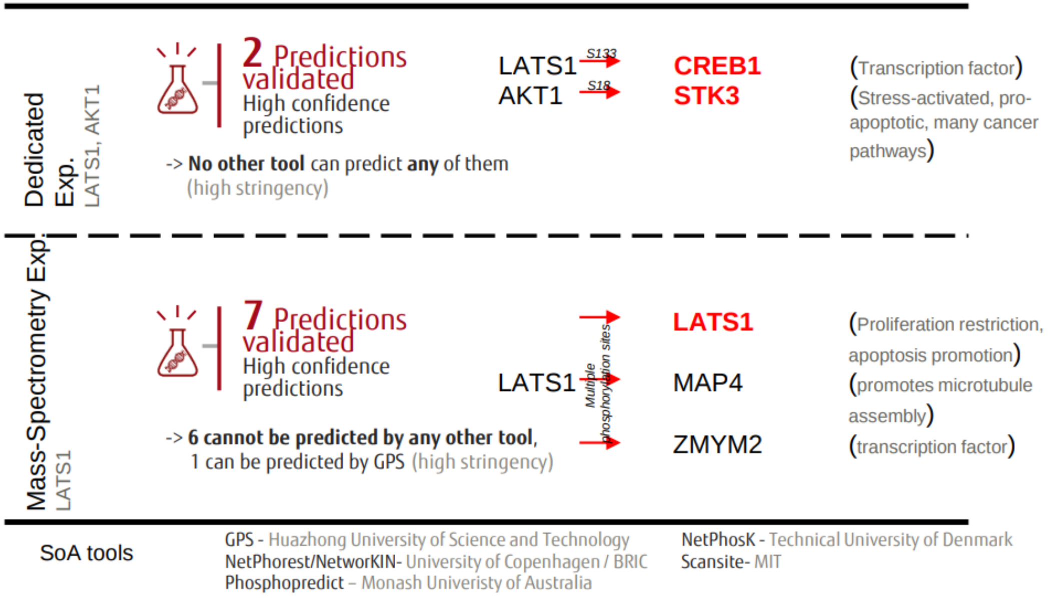

Training
Input phosphorylation data
Knowledge Graph conversion
Enrichment with negatives
Training and model selection
TKL_02
RAF1
BAD
JAK2
model
Best Performing model
Prediction
Protein query
Candidate Generation
Knowledge Graph conversion
Prediction and ranking
Conversion to phosphorylation network
LATS1
CLK_03
CREB1
S133
Score: 0.693

Kinase	Location	Substrate
RAF1	S99	BAD
...	...	...

Kinase	Locatio n	Substrate
LATS1 ... KC1A	S133 ... S464	CREB1 ... LATS1

Signalling Prediction – Result Summary

– AU-PR: 0.906 (+- 0.02), AU-ROC: 0.958 (+- 0.006)



Dedicated Exp.
LATS1, AKT1

2 Predictions validated
High confidence predictions

LATS1 $\xrightarrow{S133}$ **CREB1** (Transcription factor)

AKT1 $\xrightarrow{S18}$ **STK3** (Stress-activated, pro-apoptotic, many cancer pathways)

-> **No other tool** can predict **any** of them (high stringency)

Mass-Spectrometry Exp.
LATS1

7 Predictions validated
High confidence predictions

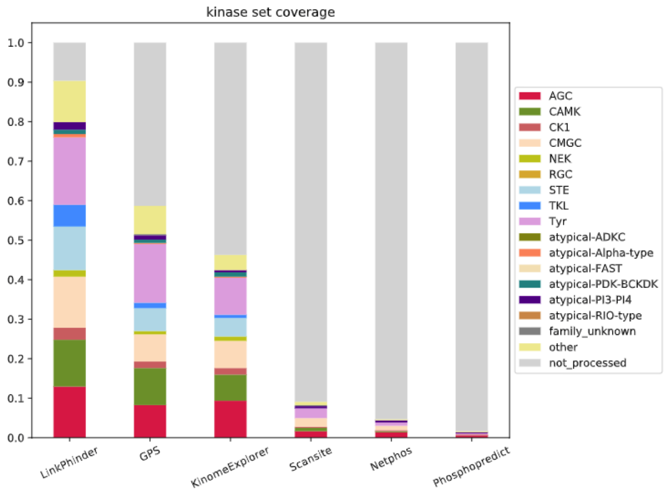
LATS1 → Multiple phosphorylation sites

- **LATS1** (Proliferation restriction, apoptosis promotion)
- MAP4 (promotes microtubule assembly)
- ZMYM2 (transcription factor)

-> **6 cannot be predicted by any other tool,**
1 can be predicted by GPS (high stringency)

SoA tools

GPS - Huazhong University of Science and Technology
NetPhorest/NetworKIN- University of Copenhagen / BRIC
Phosphopredict – Monash Univeristy of Australia
NetPhosK - Technical University of Denmark
Scansite- MIT



Collaboration with

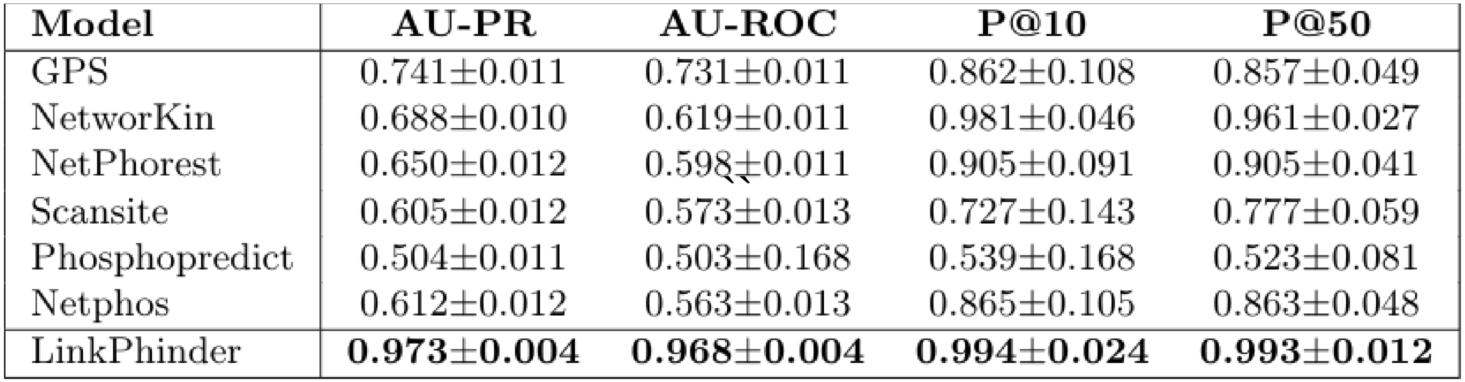
SYSTEMS BIOLOGY IRELAND

Prof. Walter Kolch

#3 World leader in Systems Medicine
#10 World leader in Precision Medicine

Submission in

nature biotechnology



Model	AU-PR	AU-ROC	P@10	P@50
GPS	0.741±0.011	0.731±0.011	0.862±0.108	0.857±0.049
NetworKin	0.688±0.010	0.619±0.011	0.981±0.046	0.961±0.027
NetPhorest	0.650±0.012	0.598±0.011	0.905±0.091	0.905±0.041
Scansite	0.605±0.012	0.573±0.013	0.727±0.143	0.777±0.059
Phosphopredict	0.504±0.011	0.503±0.168	0.539±0.168	0.523±0.081
Netphos	0.612±0.012	0.563±0.013	0.865±0.105	0.863±0.048
LinkPhinder	**0.973±0.004**	**0.968±0.004**	**0.994±0.024**	**0.993±0.012**

UI available at

linkphinder.insight-centre.org

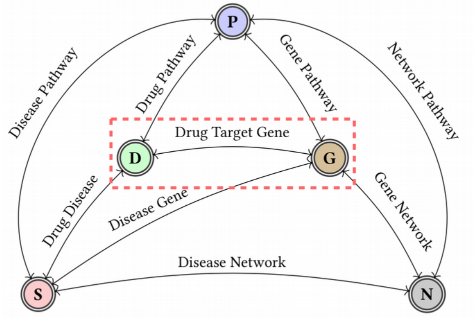
Nature Biotechnology logo adapted from https://www.nature.com/nbt/.

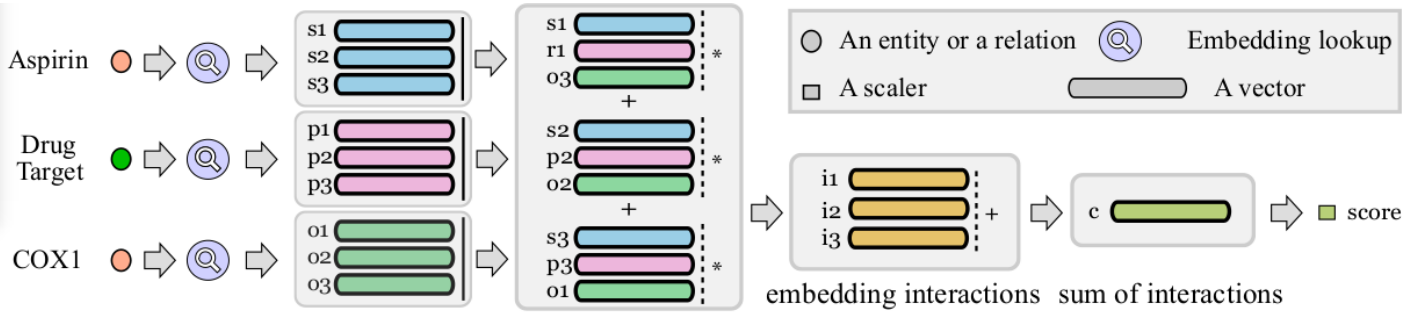
Outline

- Institute / group overview
- Knowledge graphs
- Biomedical discovery informatics
  - Signalling prediction
  - Drug target prediction
  - Ultimate goal

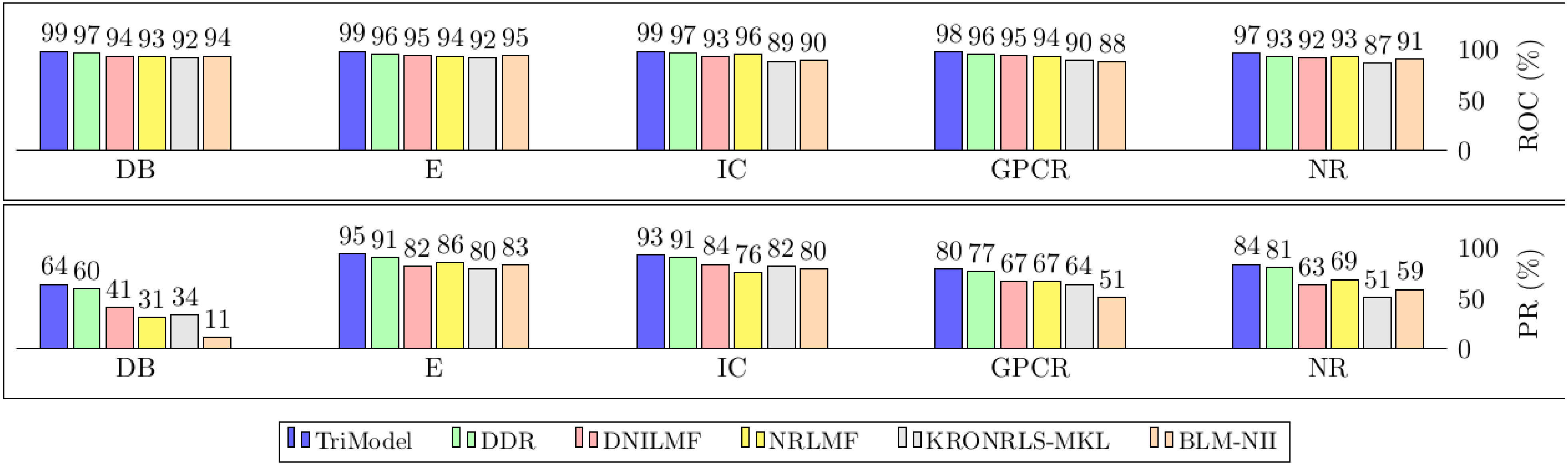
Drug Target Prediction – Outline of the Problem and Solution

- **Problem**
  - Drugs work by **interacting** with **proteins** in the human body
    - Interactions with the “right” proteins lead to **therapeutic effects**
    - Interactions with unwanted proteins may lead to **adverse (side) effects**
  - The human body has over **20k genes** that produce around **100k proteins**
  - **Hard to screen** for drug interactions at this **sheer scale**
  - **Computational predictions** can give new insights into therapeutic activities and adverse effects of both **de novo** and **approved** compounds
  - Current techniques do not fully utilise all **available knowledge**
- **Solution**
  - Integrate relevant **curated information** in **knowledge graphs**
  - Train a custom-made **knowledge graph embedding** model to make predictions





Drug Target Prediction – Result Summary



- Joint work with **University of Bristol**
- Submission in OXFORD ACADEMIC Bioinformatics
- UI available at http://drugtargets.insight-centre.org/

OUP Bioinformatics logo adapted from https://academic.oup.com/bioinformatics.

Outline

- Institute / group overview
- Knowledge graphs
- Biomedical discovery informatics
  - Signalling prediction
  - Drug target prediction
  - Ultimate goal

Ultimate Goals – Problem

Patient data + **Current technology** = **Limited results**

(Experimental data + expert knowledge)

(Limited understanding of complex interactions)

PATIENT DATA

Patient X

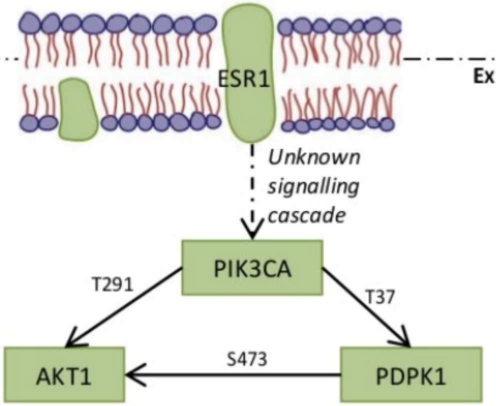
Diagnosis:

- ER+ Breast Cancer
- Resistant to Tamoxifen

Gene expression data:

- Has mutation of two genes:
  - SGK1
  - PIK3CA

EXPERIMENTS IN ER+ BREAST CANCER CELL LINES



EXPERT KNOWLEDGE

Overexpression of AKT1 and SGK1 lead to uncontrolled cell growth (= cancer)

MEDICAL DECISION

- Stopping Tamoxifen
- Starting off-the-shelf chemotherapy

Outcome:

- Bad side effects
- Sub-optimal treatment

Ultimate Goals – Solution

Patient data, structured and unstructured data +

KInCom technology

(Discoveries + experimental data)

= Unprecedented results

(Commercial and societal impact)

Patient Data

Patient X

Diagnosis:
- ER+ Breast Cancer
- Resistant to Tamoxifen

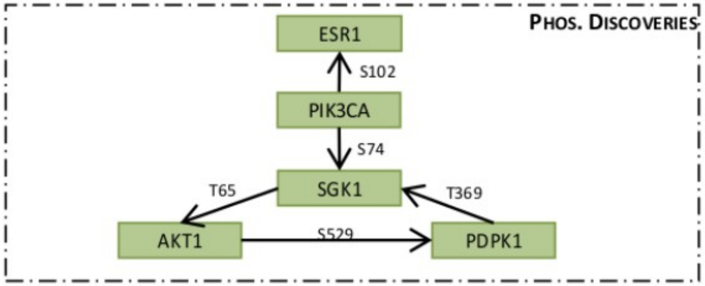
Gene expression data:
- Has mutation of two genes:
  - SGK1
  - PIK3CA

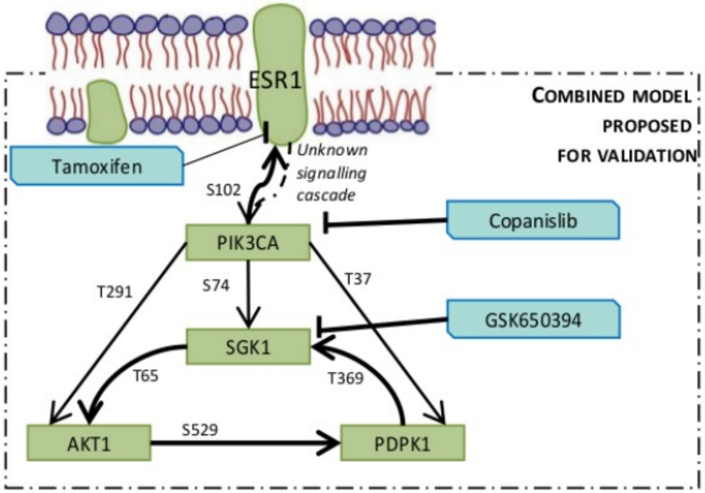
- Proteins
- Genes
- Drugs
- Mutations

Structured Data

Biomedical Scientific Literature

Unstructured Data





Novel Targeted Treatment proposed

- The resistance of Patient's X cancer to *Tamoxifen* can be overcome by combining it with approved drug *Copanislib* and experimental candidate pro-drug *GSK650394* (currently only tested for colon cancer)

Explainable AI:
- This is due to the specific pathways active in Patient's X tissue due to the mutations in two genes and data known from generic breast cancer tissues