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/](https://www.nuigalway.ie/))
  - For details, see [https://datascienceinstitute.ie/](https://datascienceinstitute.ie/), [https://insight-centre.org/](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**
  - Straightforward **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
Signalling Prediction – Result Summary

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

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

<table>
<thead>
<tr>
<th></th>
<th>ROC (%)</th>
<th>PR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DB</td>
<td>99.97</td>
<td>64.60</td>
</tr>
<tr>
<td>E</td>
<td>99.96</td>
<td>95.91</td>
</tr>
<tr>
<td>IC</td>
<td>99.97</td>
<td>93.91</td>
</tr>
<tr>
<td>GPCR</td>
<td>98.96</td>
<td>80.77</td>
</tr>
<tr>
<td>NR</td>
<td>97.93</td>
<td>84.81</td>
</tr>
</tbody>
</table>

- Joint work with **University of Bristol**
- Submission in **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

Diagnosis:
- ER+ Breast Cancer
- Resistant to Tamoxifen

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

Medical Decision
- Stopping Tamoxifen
- Starting off-the-shelf chemotherapy

Outcome:
- Bad side effects
- Sub-optimal treatment

Overexpression of AKT1 and SGK1 lead to uncontrolled cell growth (= cancer)
Patient data, structured and unstructured data

**Diagnosis:**
- ER+ Breast Cancer
- Resistant to Tamoxifen

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

---

**KlnCom technology**
(Discoveries + experimental data)

**Unprecedented results**
(Commercial and societal impact)

---

**Novel Targeted Treatment Proposed**
- The resistance of Patient’s X cancer to Tamoxifen can be overcome by combining it with approved drug Copanlisib 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