

Data challenges for deep learning in drug discovery



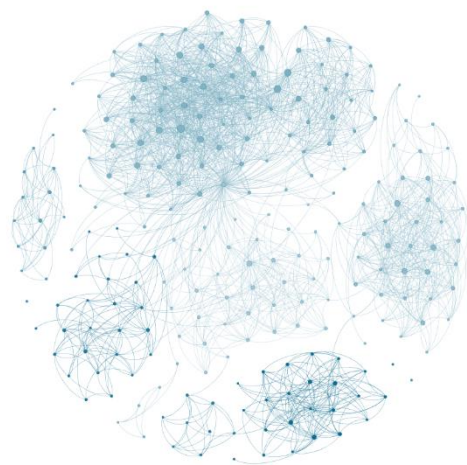
Intellegens

Tom Whitehead
10th July 2019

Deep learning data



Exhaustive



Precise



Strictly defined

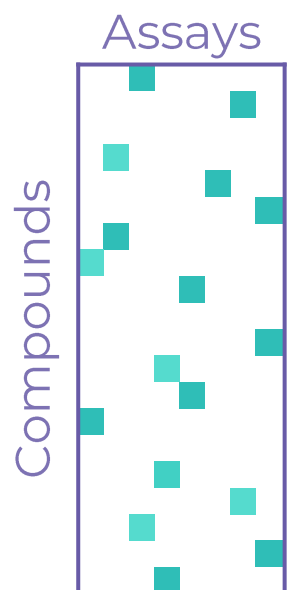


CAT

Life science data



Sparse



Noisy

$$\text{pIC}_{50} = 6.5 \pm 0.5$$

Weakly defined



Intellegens



Deep learning for **sparse** and **noisy**
experimental data

Combine all data sources to **extend horizons**

Focus on most **confident** predictions to
increase utility



Data problems

Sparsity

Noise

Weak definitions

Size

Confidentiality

Data ~~problems~~ opportunities



Sparsity

Knowledge transfer

Noise

**Robust confidence
estimates**

Weak definitions

Flexibility

Size

Cloud computing

Confidentiality

**Anonymisation,
federated learning**



Case study

Big Pharma dataset

Sparsity – $\ll 1\%$ compound/assay pairs

Noise – included cell-based assays

Size – 700,000 compounds, 2,500 assays

Confidentiality – no structures, no assay labels



Case study

Big Pharma dataset

Sparsity – $\ll 1\%$ compound/assay pairs

Used Alchemite™ deep learning

Noise – included cell-based assays

Focussed on most confident predictions

Size – 700,000 compounds, 2,500 assays

Used AWS to scale rapidly

Confidentiality – no structures, no assay labels

Compound descriptors, generic method

Conclusions



Life science data provides challenges, but they are surmountable

Potential for deep learning to add significant value to drug discovery projects

tom@intellegens.ai