From Data to Knowledge:
Informatics at Medicines Discovery Catapult

G. Holliday[1], D. James[1], K. Burusco-Goni[1,2], A. Ioannidou[1], M. Warren[1], C. Southan[1,3], S. Rehman[1,2],
H. Barjat[1], A. Pallo[1], N. Etherington[1], R. Jimenez[1], M. Hodgkiss[1], J. P. Overington[‡], and I. Dunlop[1]
Medicines Discovery Catapult. Alderley Park, Macclesfield, Cheshire, SK10 4ZF, UK.

md.catapult.org.uk

Medicines Discovery is hard. With high failure rates, often late in the discovery pipeline and ever-increasing costs it is critical that we look for new ways to innovate in this field. Here at Medicines Discovery Catapult, we seek to do just that using a data driven, patient centric approach. We have a highly inter-disciplinary and collaborative team with expertise ranging from protein informatics to systems biology; imaging to genomics, and cheminformatics to data science and software engineering; we are here to help you innovate and are on the lookout for those "someone really needs to" challenges!

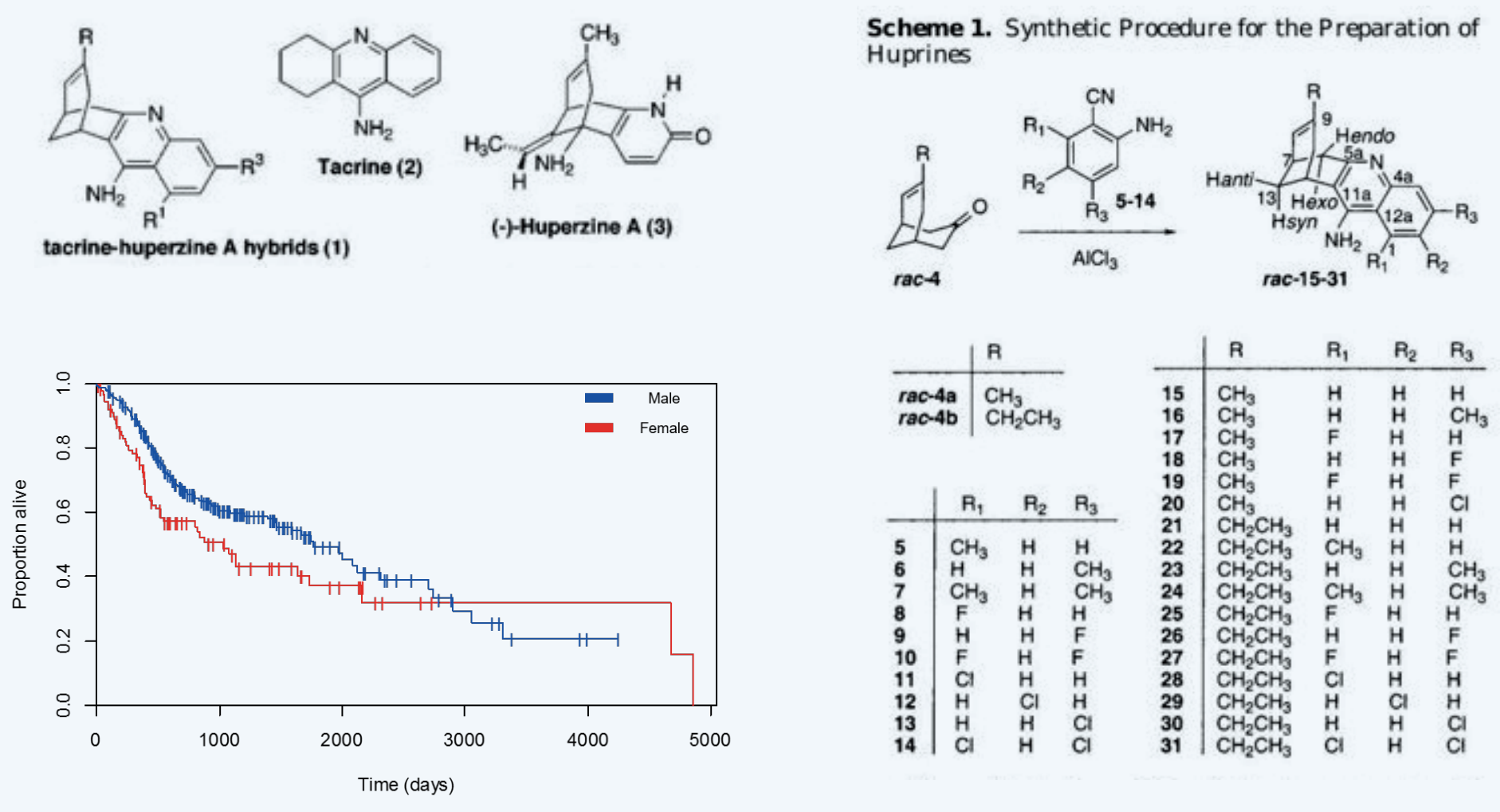
Data are being generated almost faster than their impact can be understood but are absolutely critical to our understanding and ability to treat patients. From high throughput screening to highly specialised wet-lab experiments, data has a central place in our ecosystem.

We have utilised Natural Language Processing (NLP) techniques combined with text mining to identify new elements for knowledge discovery and developed novel models with our SME partners. Knowledge extraction from clinical trials data combined with NLP has led to successful collaborations with patient-led charities in the area of drug repurposing and production of target product profile. We use a combination of chem- and bioinformatics techniques to assess the druggability of targets and perform high-throughput screening to help partners drive forward their target and small molecule assets. Combining our state-of-the-art platforms for multiplex tissue imaging and medical imaging technology, informatics is embedded into our biomedical research infrastructure to advance the understanding of biological samples to a range of clinical subjects. Additionally, workflows encompassing automated identification and extraction of relevant data from graphs is applied to several projects with great success, as have systems biology and knowledge graph-based approaches.

Here we highlight several of our successes and ask: how can we help you?

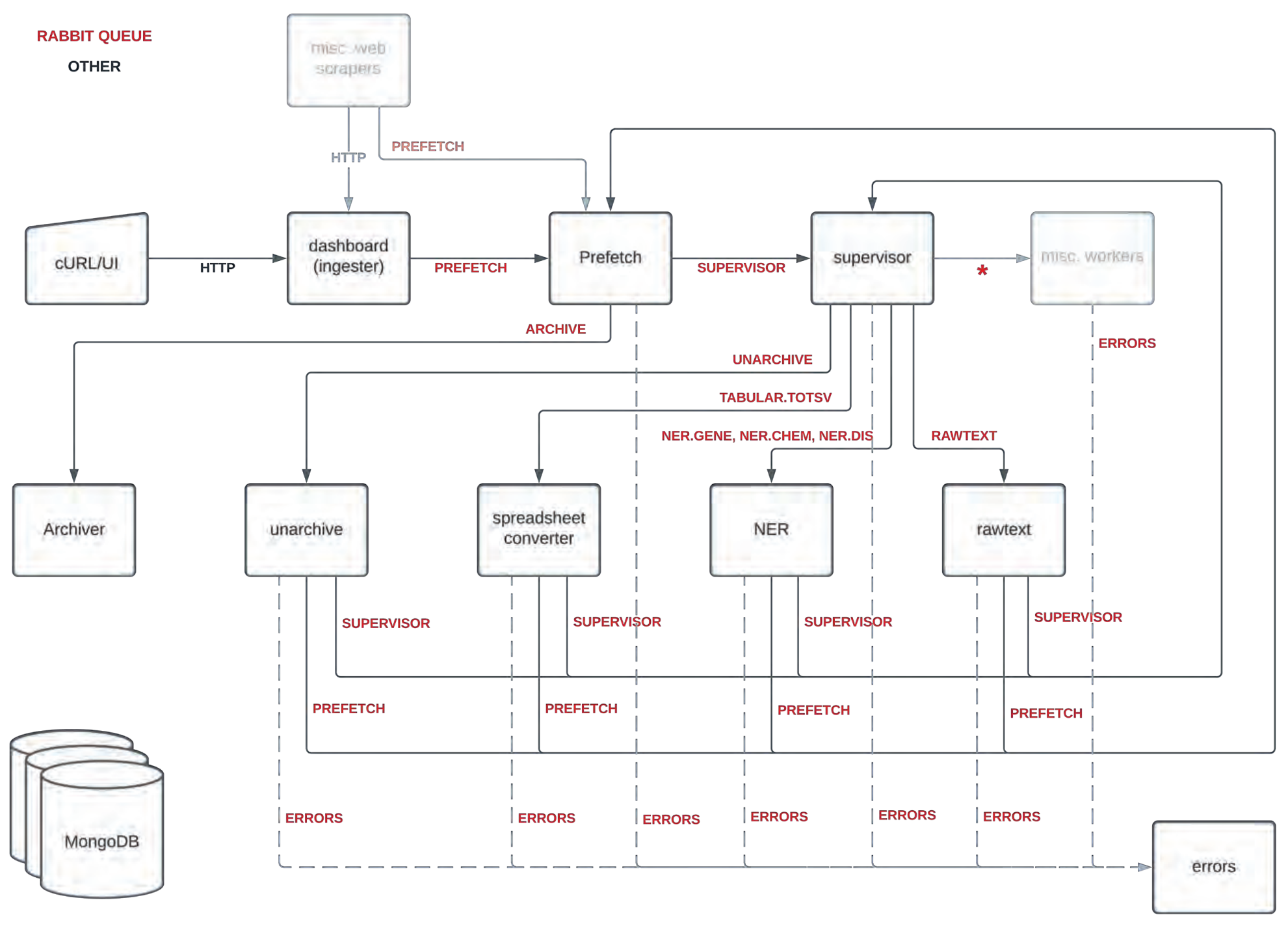
Extracting and Classifying Scientific Figures

Text mining biomedical literature is an established practice within drug discovery pipelines. MDC have identified additional value by extracting, classifying and indexing the figures from drug discovery relevant scientific papers. Our Machine Learning tools find and extract figures from scientific journals, providing the user with useful information quickly and efficiently. Replace hours of manually searching through literature with automated ML pipelines. Along with our partner Chief.AI we developed AI models to classify figures, and tables. Our data pipelines can then extract data from these including SMILES, text and CSV.


tacrine-huperzine A hybrids (1)
Tacrine (2)
(-)-Huperzine A (3)
Scheme 1. Synthetic Procedure for the Preparation of Huprines
rac-4
rac-15-31
Proportion alive
Time (days)
Male
Female

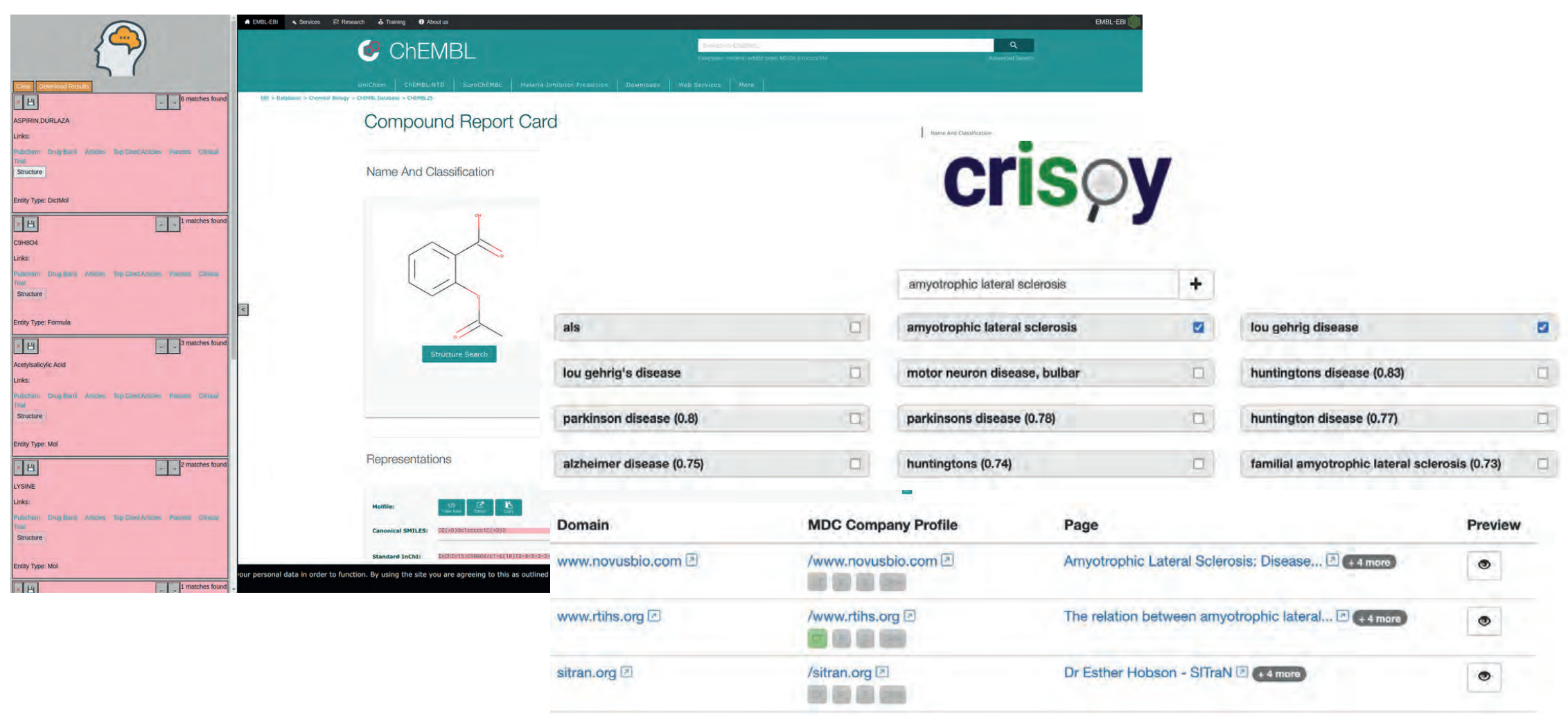
Deep ADMET

Working with our partners Optibrium and Inteligens we developed machine learning algorithms and pipelines to find ADMET data within unstructured biomedical literature that isn't available through existing open databases e.g. ChEMBL. Additional ADMET data can then be used to train improved QSAR models to identify relevant chemical structures for a particular target. Optibrium, in collaboration with Inteligens have developed Cerella™, a state-of-the-art QSAR platform leveraging the power of deep learning.



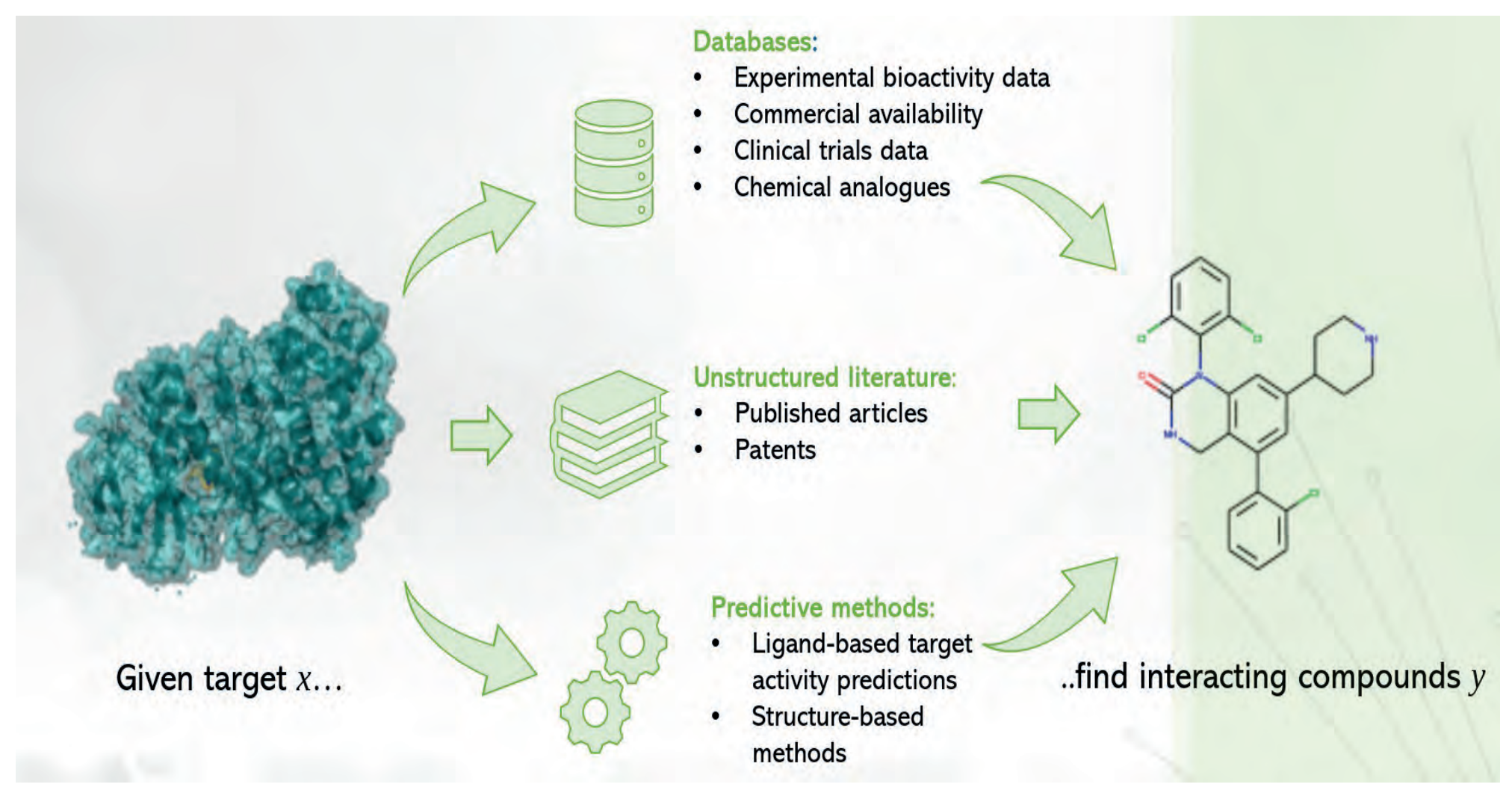
Finding Collaborators and Enriching Biomedical Web Pages

MDC have developed a domain specific search engine which facilitates easy searching of UK drug discovery organisations by integrating website, patents, clinical trials and grant data. We have also developed a browser extension which can find, annotate, cross reference and download compound, gene/protein and disease references within web pages and PDFs.



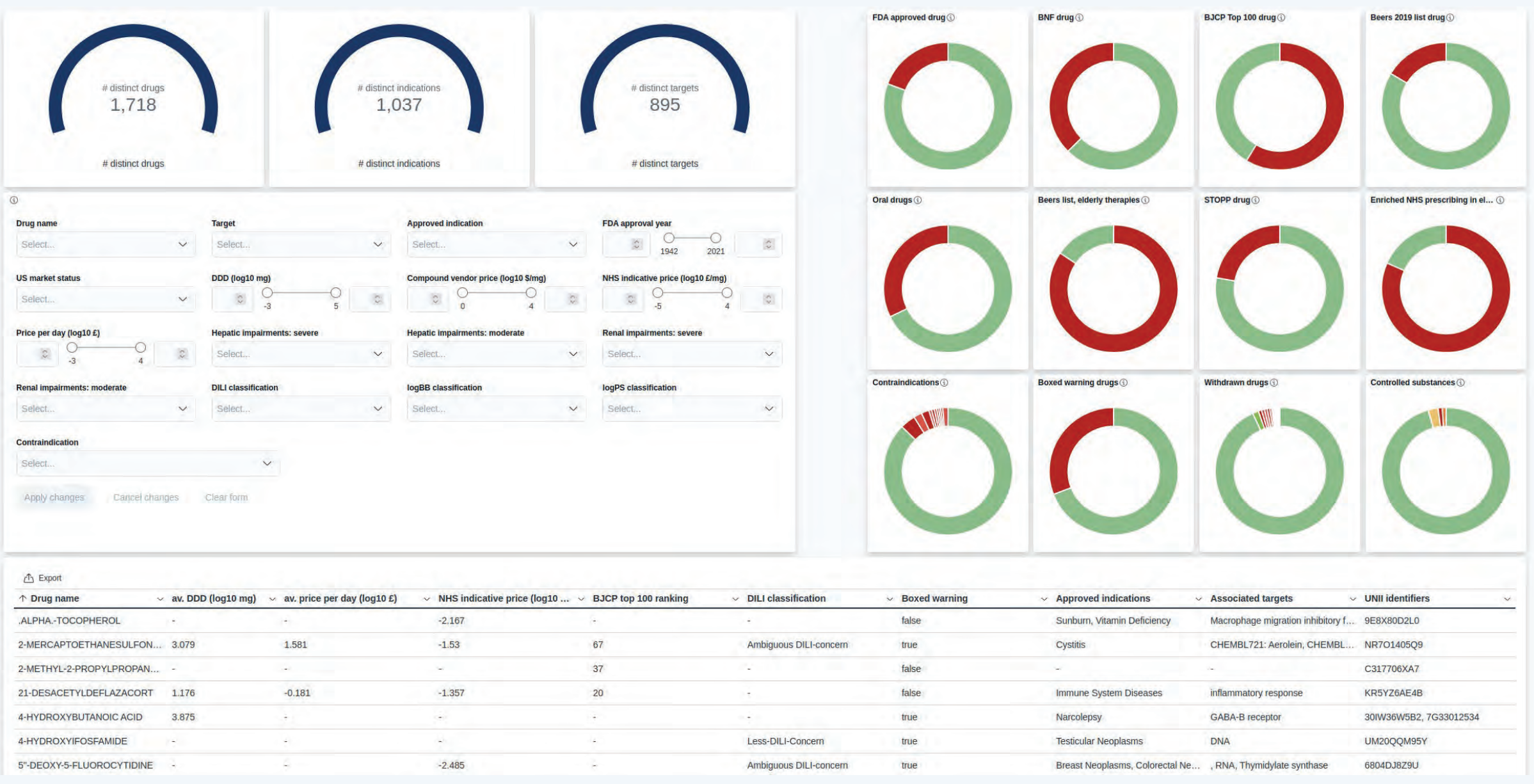
Deep Chemotyping

Our enhanced docking/reverse-docking modular tools can help determine the optimal biological target for a given compound and vice versa using an integrated protocol of Docking, Molecular Dynamics and 3D Deep Neural Networks techniques.



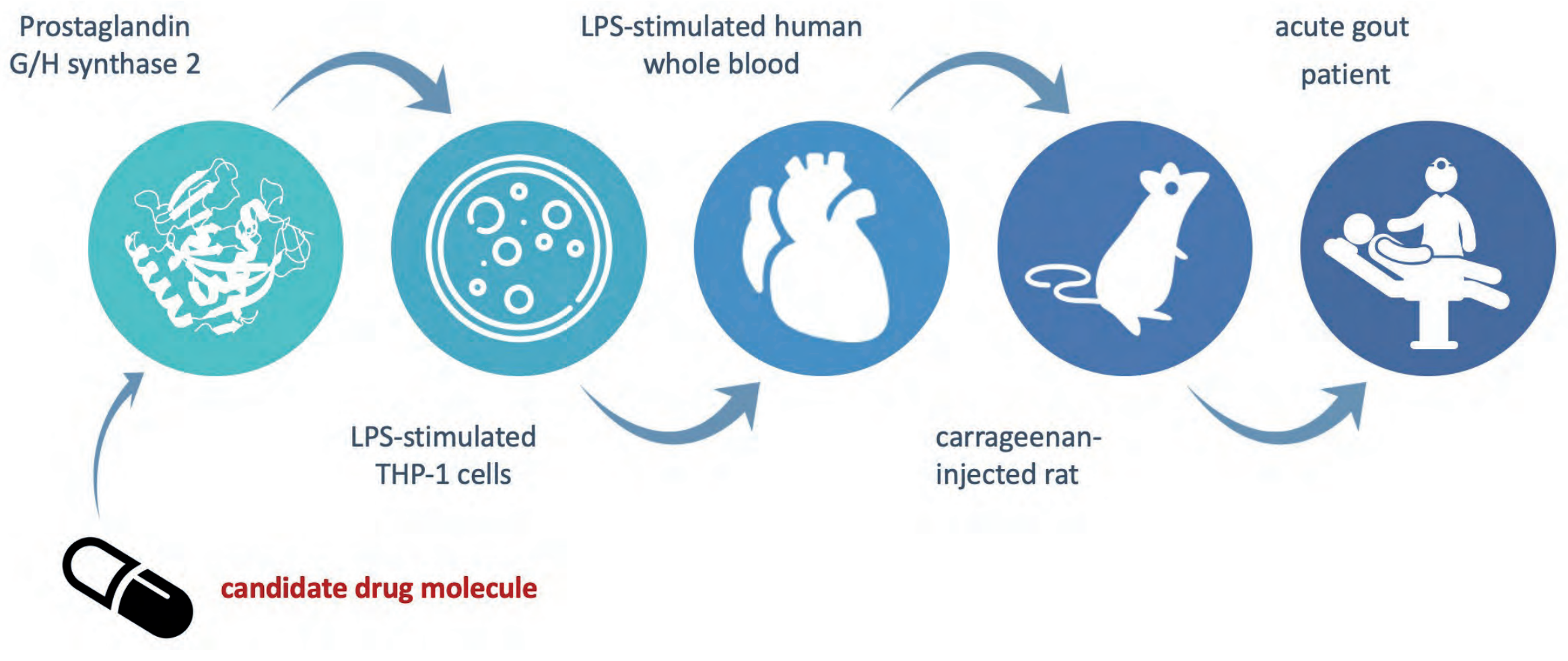
Drug repurposing

Working with our partners at UK-SPINE we used our tools and data pipelines to assemble a list of drugs which could be filtered based on efficacy, tolerability and safety in an elderly population with multi morbidities.



AssayNET

MDC are developing a machine learning based solution to identify, extract and classify assays from biomedical literature. The solution joins up clinical trial end points with preclinical assays to explore and optimise which assays should be used when developing a drug for a particular disease.


Prostaglandin G/H synthase 2
LPS-stimulated human whole blood
acute gout patient
LPS-stimulated THP-1 cells
carrageenan-injected rat
candidate drug molecule

Acknowledgements

We thank our funders UKRI Innovate UK. We also thank:

Andrew Pannifer[‡], Charles Bury[‡], Adam Poulston[‡], David Allen[‡], Mark Laing[‡], Matthew Cockayne[‡], Andrey Gagunashvili[1], Brad Rollings[1], Ashley Evans[1], Michael Sweeton[1], Alex Whitehorn[1] for their valuable contributions to the above suite of tools. None of this would have been possible without the guidance and support of JPO. Bon voyage.

1. MDC
2. University of Manchester
3. University of Edinburgh
‡ Ex MDC employee