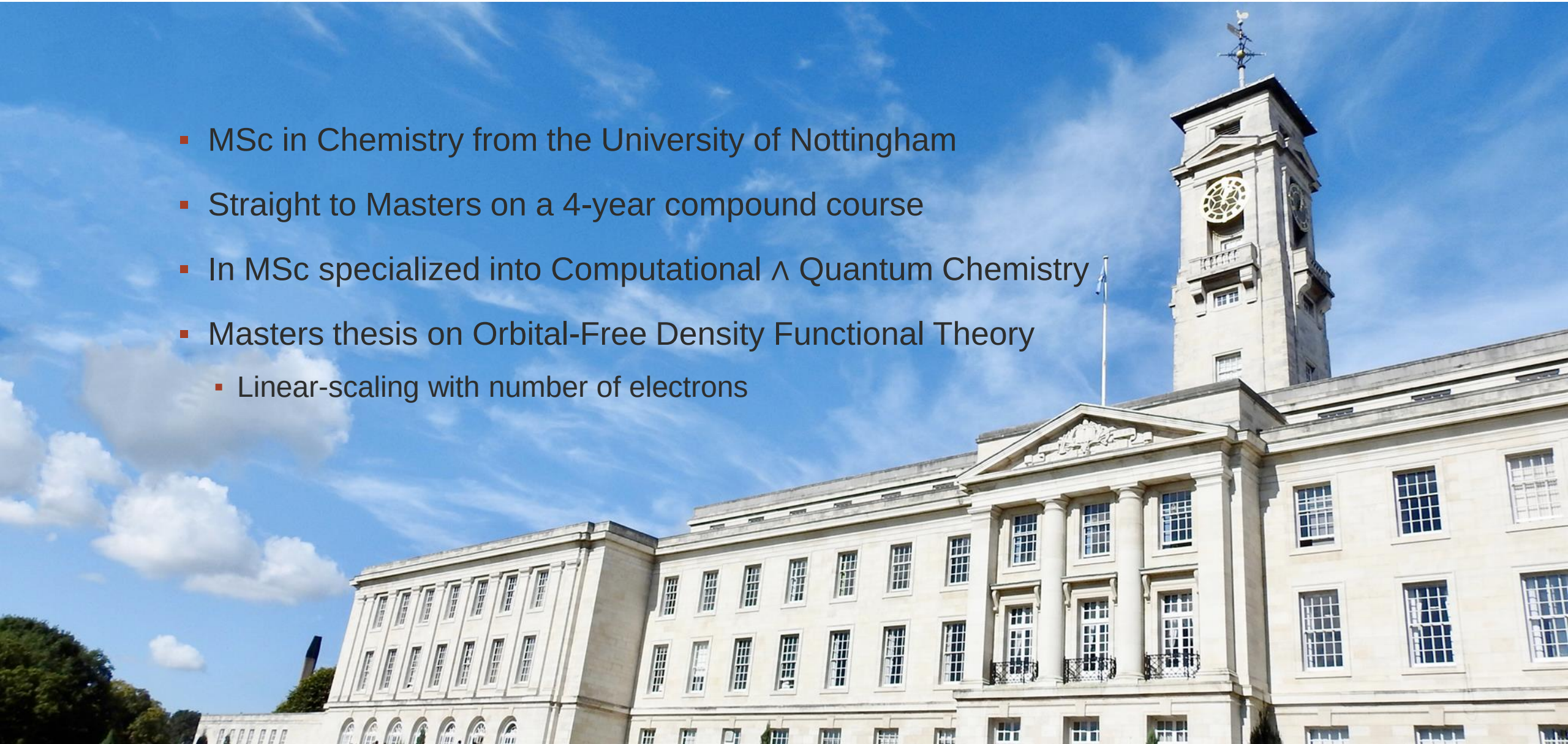
BIGCHEM Winter School Presentation

Michael Withnall – ESR10 Secure Sharing of Information

Personal Background

- Come from West Haddon
  - Small village in Northamptonshire, England
- 23 years old

Academic Background



- MSc in Chemistry from the University of Nottingham
- Straight to Masters on a 4-year compound course
- In MSc specialized into Computational ∧ Quantum Chemistry
- Masters thesis on Orbital-Free Density Functional Theory
  - Linear-scaling with number of electrons

BIGCHEM Project ESR10

Secure sharing of information using ensemble of machine learning methods and surrogate data

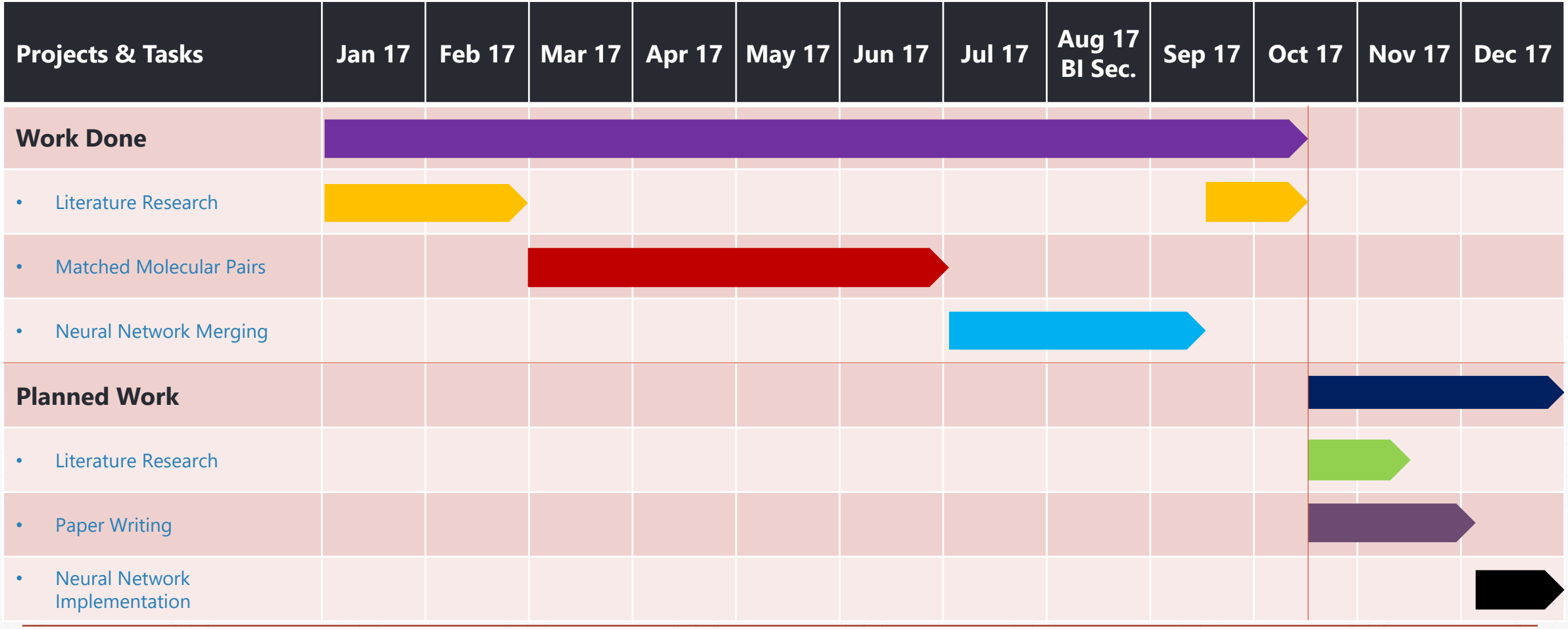
- To find optimal strategies for sharing of chemical information by combining predications of models developed using data from individual collaborating partners.
- To investigate the surrogate data approach for model sharing
- To apply linear multiparty secure computation approaches for model development
- To compare pros and cons of different methods.

BIGCHEM Project ESR10 – Deliverables

- D4.1 – Overview of Strategies for data sharing (in progress)
  - The report will summarize the strategies for secure sharing of data that will be developed and validated during the project.

- D4.2 – Comparison of performances of different data sharing approaches (future)
  - Report will assess the performance of different data sharing strategies.

Project Progress ESR10



Achievements / Publications ESR10

- Successful publication of an article during 1st year of project.

Withnall M; Chen H; Tetko I

Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective.

ChemMedChem, 2017, 12, 1-9.

doi:10.1002/cmdc.201700303 (Open Access)

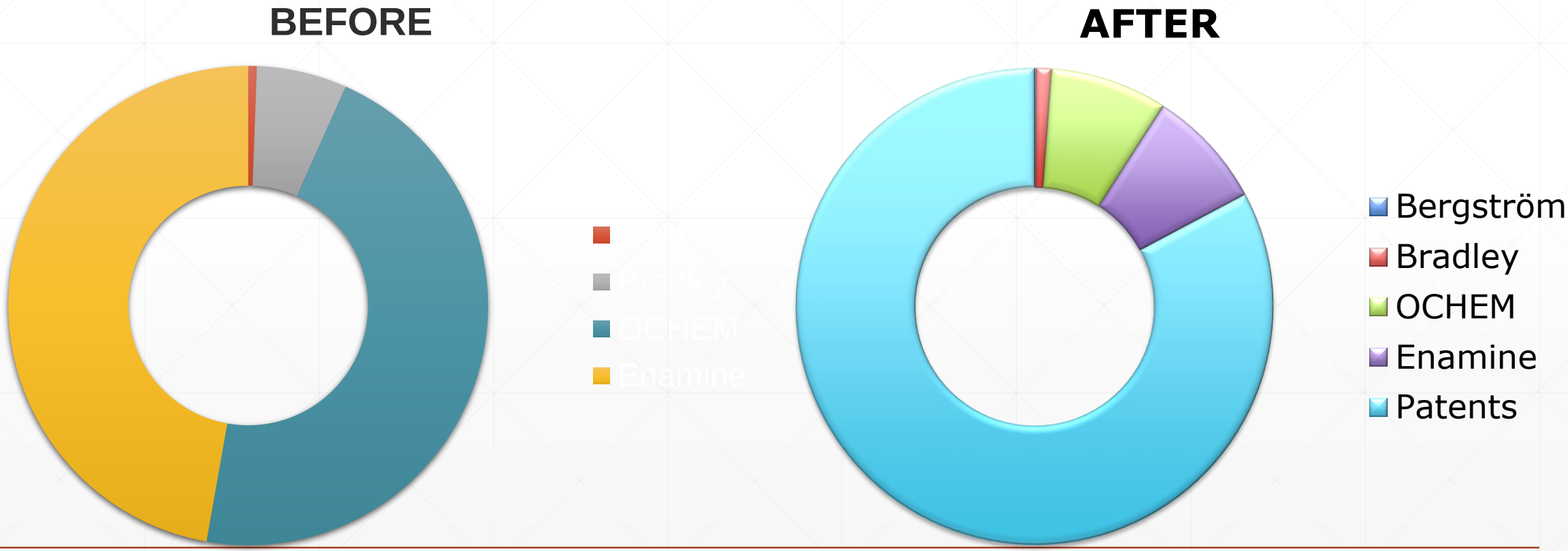
Dataset

- **275,008 molecules after filtering**
  - incomplete records
  - compounds with a molecular weight >1000 Da.
- **Primarily in the drug-like range (50–250°C)**
- **Chemaxon:**
  - Standardised
  - Neutralised
  - Salts were removed
  - Structures were cleaned
- **Melting Points ranging from –199°C to 517°C**

- **Patents**
- Research papers published by:
  - Bradley
  - Bergström
- Enamine
- The existing OCHEM database

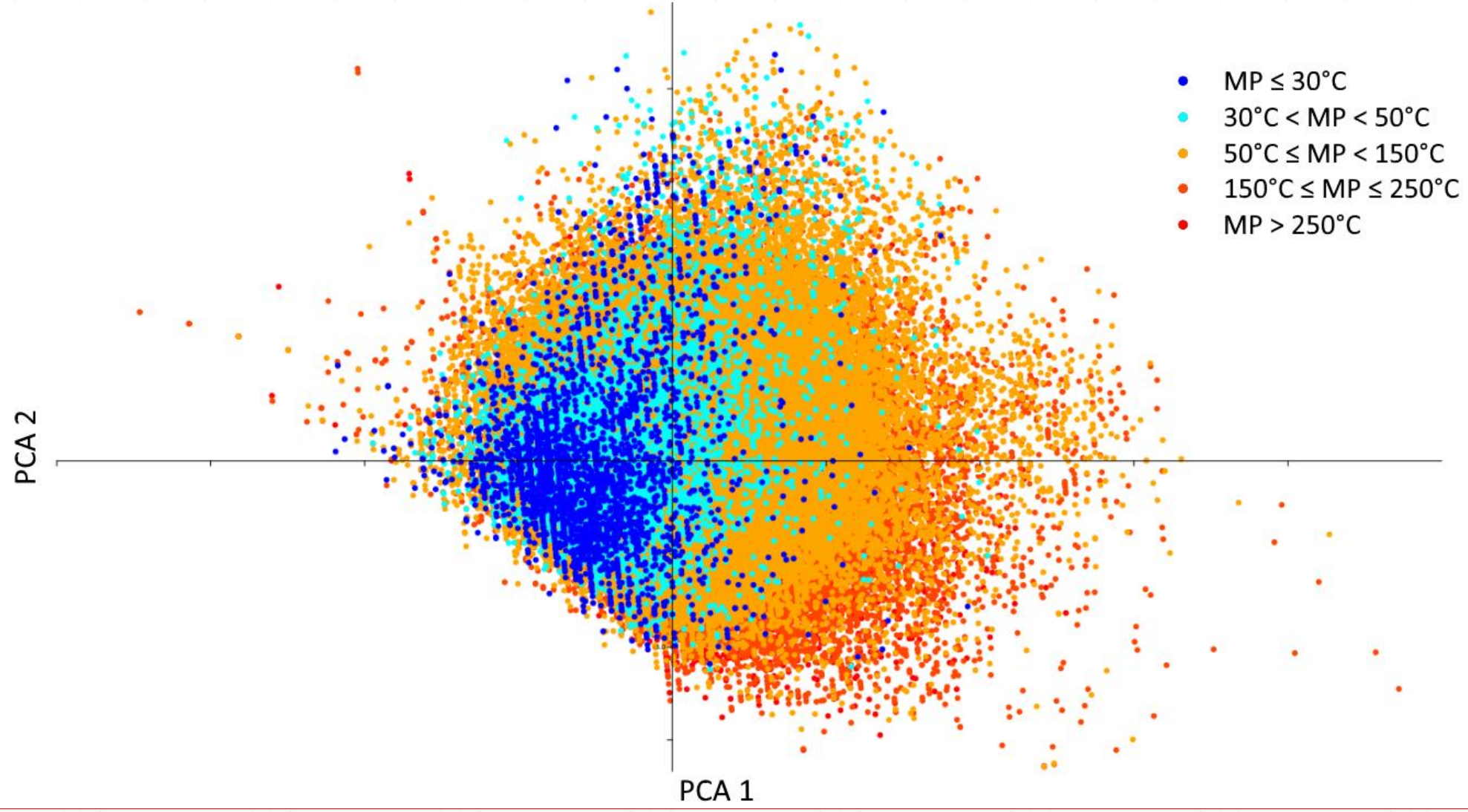
Dataset - Patents

- Majority of chemical data was taken from publicly available patents.
- Added another *ca.* 225,000 compounds to dataset



Reproduced with permission from the Tetko and Kazan presentation about melting point prediction using OCHEM.
I. V. Tetko, D. M. Lowe, A. J. Williams, J. Cheminformatics 2016, 8, DOI 10.1186/s13321-016-0113-y

Dataset - Complete



Method

- **Investigate pairs where only a single descriptor has changed.**

Molecule	Desc.1	Desc.2	Desc.3	Desc.4	Desc.5	Desc.6	Desc.7	Desc.8
Mol1	0	1	17	0	2	0	2	2
Mol2	0	4	18	1	2	0	0	3
Mol3	0	7	14	0	2	0	4	4
Mol4	0	4	13	1	2	0	0	3
Mol5	0	3	17	1	1	0	2	2

Method

- **Investigate pairs where only a single descriptor has changed.**

Molecule	Desc.1	Desc.2	Desc.3	Desc.4	Desc.5	Desc.6	Desc.7	Desc.8
Mol1	0	1	17	0	2	0	2	2
Mol2	0	4	18	1	2	0	0	3
Mol3	0	7	14	0	2	0	4	4
Mol4	0	4	13	1	2	0	0	3
Mol5	0	3	17	1	1	0	2	2

Iterate through list of MMPs:

- Dictionary of molecules and descriptors for fast lookup and lower memory footprint
- Add ‘hits’ to list of objects with relevant information for later processing

Method

- **Investigate pairs where only a single functional group, or pair thereof, has changed.**

Molecule	FG.1	FG.2	FG.3	FG.4	FG.5	FG.6	FG.7	FG.8
Mol1	0	1	0	1	0	0	1	0
Mol2	0	1	1	0	0	0	1	0
Mol3	0	0	1	0	1	1	0	1
Mol4	0	1	1	0	1	0	1	0
Mol5	1	1	1	1	1	1	1	0

- **e.g. Mol2 ➔ Mol4 transformation results in an aryl chloride**
- **e.g. Mol1 ➔ Mol2 transformation changes carboxylic acid to carboxylic acid ester**

Results – 2D Descriptors

Descriptor Changed	Number of Samples	Mean Descriptor Change	$\Delta T_m/\Delta$Descriptor (°C)	Standard Error of Mean (°C)
Fluorine atoms	17,297	1.29	1.2	±0.3
Chlorine atoms	9,893	1.04	6.2	±0.4
Bromine atoms	2,804	1.02	14	±0.8
Iodine atoms	400	1.02	20	±2.2
H-Donors	12,889	1.02	23	±0.5
H-Acceptors	24,358	1.16	11	±0.3
Rotatable bonds	68,531	1.27	–7.3	±0.2
$logP_{calc}$	24,818	0.92	4.6	±0.4

- All p-values < 0.0001
- logP classified as unchanging when Δ(logP) < 0.5

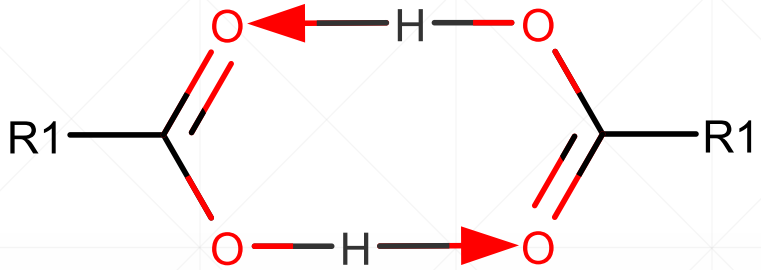
Results – 2D Descriptors

Descriptor Changed	Number of Samples	Mean Descriptor Change	$\Delta T_m/\Delta$Descriptor (°C)	Standard Error of Mean (°C)
Fluorine atoms	17,297	1.29	1.2	±0.3
Chlorine atoms	9,893	1.04	6.2	±0.4
Bromine atoms	2,804	1.02	14	±0.8
Iodine atoms	400	1.02	20	±2.2
H-Donors	12,889	1.02	23	±0.5
H-Acceptors	24,358	1.16	11	±0.3
Rotatable bonds	68,531	1.27	–7.3	±0.2
$logP_{calc}$	24,818	0.92	4.6	±0.4

- Similar results to the original study by Schultes *et al.*
- Largest observed effect from hydrogen bond donors
- Rotatable bond contribution decreases MP

Results – 2D Descriptors

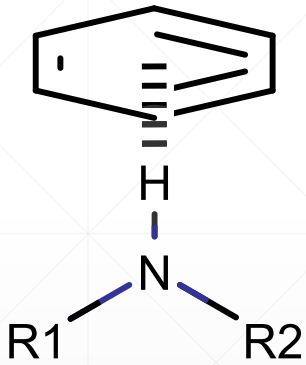
- Hydrogen Bond Acceptors
- Intermolecular Interactions ↑
- Crystal Lattice ↔ Crystal Lattice

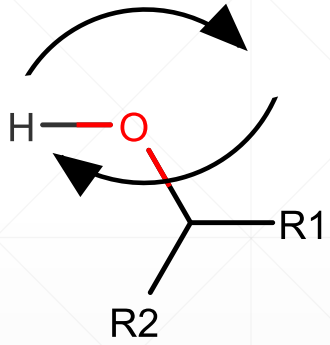


Hydrogen Bond Donors

Intermolecular Interactions ↑

Crystal Lattice ↔ Crystal Lattice





Substantial number of H-bond donors are amines:

This can result in protonation, and strong ionic lattice interactions.

Halides

Correlates well with the known intermolecular halogen bonding series:

MP increasing down the series

Rotatable Bonds

Increase in number of DoF ➔ higher molecule flexibility ➔ higher melting entropy

In some cases, also leads to less efficient crystal packing.

Conferences / Outreach

- *“Drug Innovation in Academia”* – Heidelberg – Poster Presentation
- STC *Symposium on Theoretical Chemistry* – Basel – Talk
- RICT2017 – *In Absentia* – Poster Presentation
- GCC (Planned) – Poster Presentation

- TUM Open Day participation (2016, 2017) Outreach

Secondments

- Completed:
  - Secondment to Boehringer Ingelheim in Biberach to work on model-building with existing pharmaceutical data

- Planned:
  - Secondment to CWI in Amsterdam for training on existing and in developing cryptographic solutions to secure multi-party computation

BIGCHEM Training and Schools

- Fortnightly online courses on Big Data and Pharmacological Design running since the start of the fellowship [first course 29 September 2016]

- Participation in 3 BIGCHEM Schools
  - Winter School 2016 – Munich
  - Summer School 2017 – Barcelona
  - Winter School 2017 – Modena

Planned Future Work

- Currently writing a review paper, comparing existing techniques for achieving the goals outlined by this project.
- Several techniques
  - Metadata (general)
  - Cryptographic (general)
  - Specific Implementation
  - *Database searching (general)*

Acknowledgements

- The BIGCHEM fellows
- The Tetko group
- Everyone who’s helped from AZ and BI
- Marie Curie Actions