Using Data Science techniques to put molecules in context

Dr Aileen Day
Data Science, Royal Society of Chemistry

What do we work on?

Tech Development

- Data processing pipeline
- Term extraction from literature

Cheminformatics

- Molecular characterisation
- Chemical similarity
- Molecule recommender

Applications

- Citation velocity
- Recommending papers

Business analytics

- Lead generation
- Data dashboards

Recommender Systems

amazon

NETFLIX

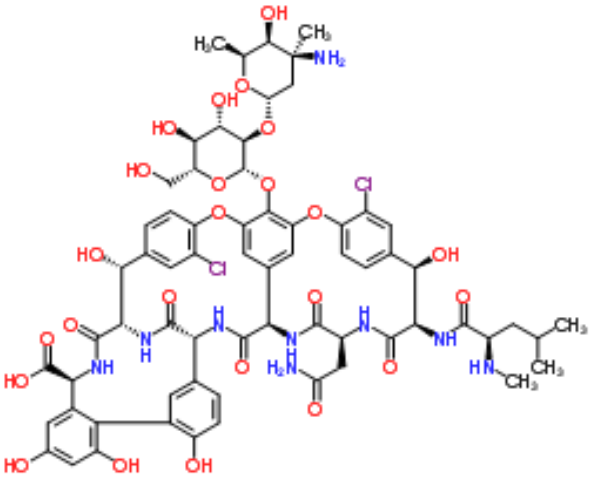
facebook

SIGMA-ALDRICH

nature

ScienceDirect®

What about a molecule recommender?



Vancomycin

What other molecules are “related” to vancomycin?

Use Cheminformatics fingerprinting…

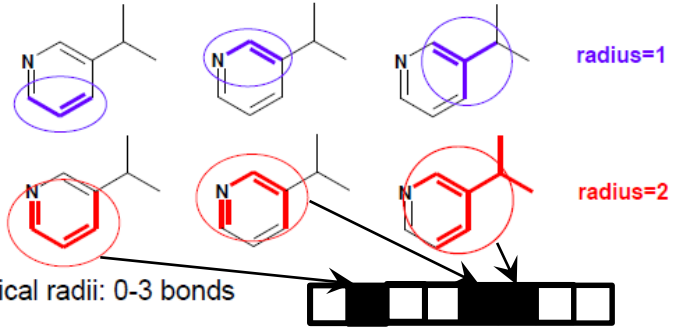
Cheminformatics fingerprinting methods

But which way?

e.g.

- Method 1: Morgan (radius=2) fingerprint
- and Dice coefficient similarity
- Use RDKit http://www.rdkit.org/

- Similarity fingerprint
- Atom types :
  - Connectivity: (Element, #heavy neighbors, #Hs, charge, isotope, inRing)
  - Chemical features: Donor, Acceptor, Aromatic, Halogen, Basic, Acidic
- Fingerprint takes into account the neighborhood of each atom:



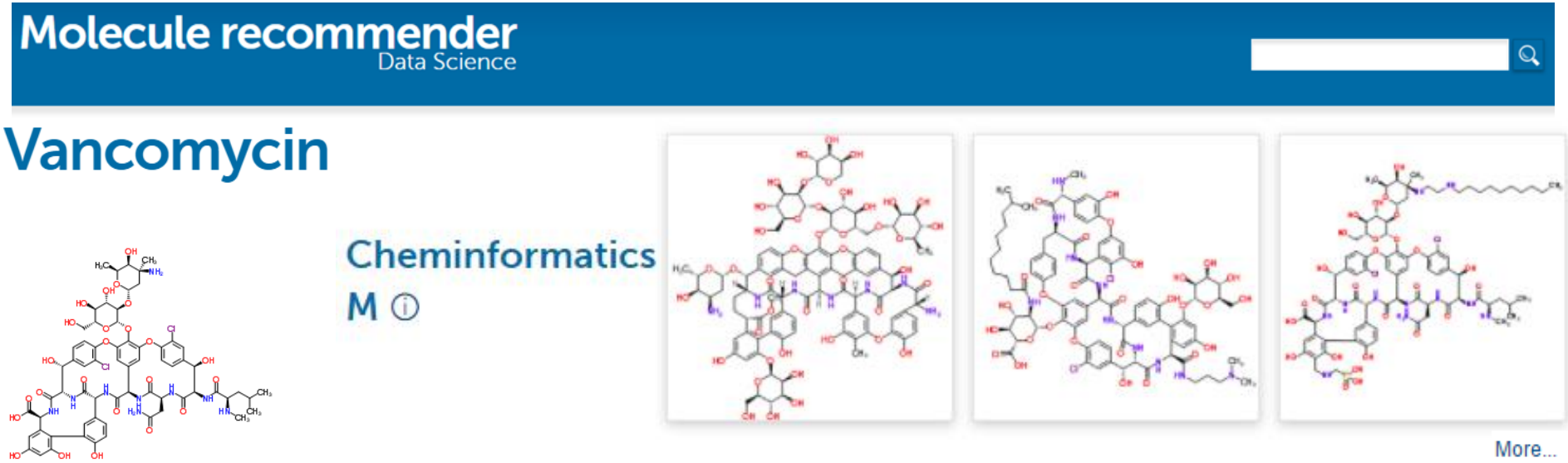
- Typical radii: 0-3 bonds

Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742-754 (2010).

NOVARTIS

What molecules are related to…

Method 1: Cheminformatics - Morgan (radius=2)



Cheminformatics similarity methods

Or…

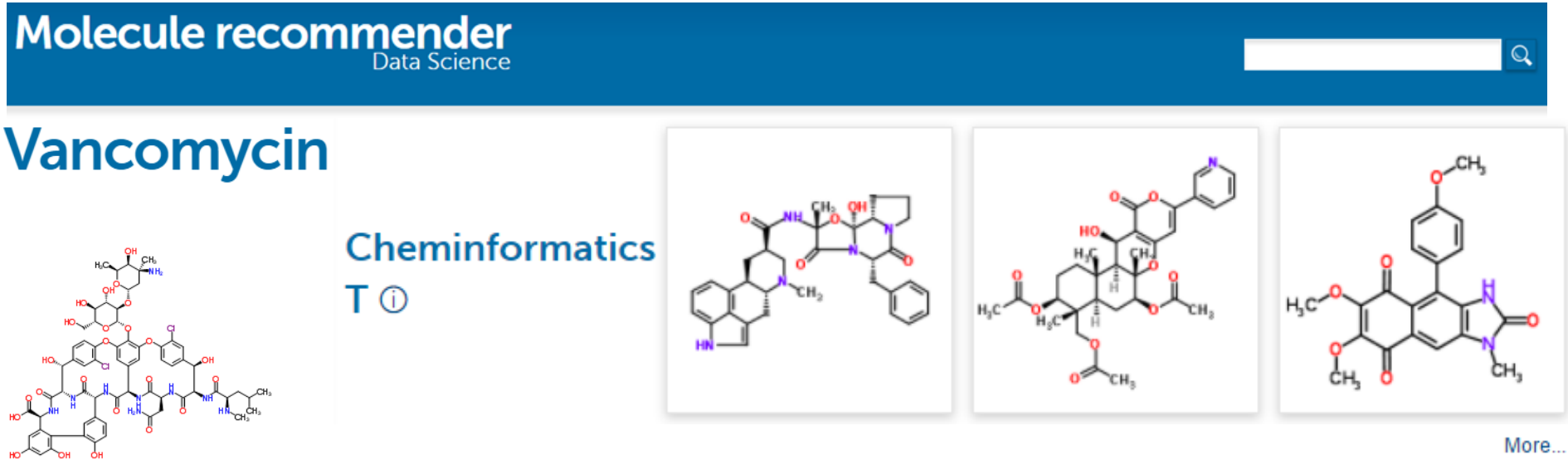
e.g.

- Method 2: Topology
- and Dice coefficient similarity
- Use RDKit http://www.rdkit.org/

- Identifies and hashes topological paths (along bonds) to make fingerprints
- then folded

What molecules are related to…

Method 2: Cheminformatics - Topology



But what do we mean by “related”?

- Researchers have different, more specific questions behind “What molecules are related to vancomycin?”

For example Amazon…

amazon

Frequently bought together

What other items do customers buy after viewing this item?

Your recently viewed items and featured recommendations

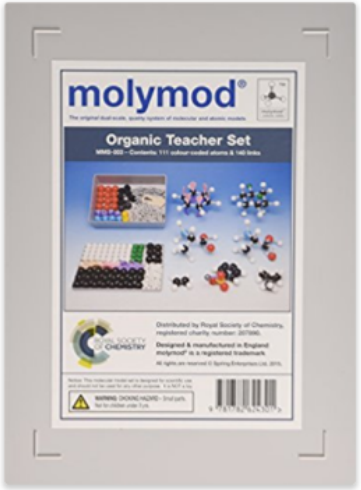
See personalised recommendations

Sign in

New customer? Start here.

For example Amazon…

amazon



See this image

Molymod Mms-003: Organic Teacher 111 Atom Set Paperback – 10 Mar 2015

by Royal Society of Chemistry (Author)

★★★★★ 2 customer reviews

See all formats and editions

Paperback
£40.74

5 Used from £33.53
20 New from £29.60

Want it delivered by Friday, 9 June? Order within 23 hrs 52 mins and choose **One-Day Delivery** at checkout. Details

Note: This item is eligible for **click and collect.** Details

These popular molecular modelling sets can be used to make many different molecules. Designed for teachers, this set contains 111 colour-coded atoms and 140 links. The medium links can be used for single bonds, while the longer, flexible links can be used for double or triple bonds. Short links can be used to create compact models. Using molecular models can help students to visualise concepts such as isomerism through hands-on learning. The models can also be used to learn about balancing equations and molecular geometry. Molymod is a registered trade mark of the EU (and other places)

Read more

For example Amazon…

amazon

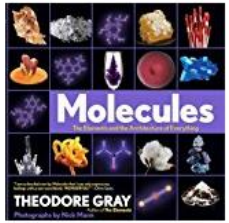
Customers who bought this item also bought

Page 1 of 2

Molymod MMS-004 Molecular Model Teacher set for Inorganic & Organic Chemistry
21

Molymod MMS-002 Molecular Model Set for Advanced Level Chemistry
51
£20.00 Prime

Top Trumps Chemistry
Royal Society of...
27
Cards
£7.20 Prime



Molecules: The Elements and the Architecture of Everything
Nick Mann
23
Hardcover
£19.99 Prime

Molymod MMS-008 50 Atom Molecular Model Set for Organic Chemistry
30
£16.85 Prime

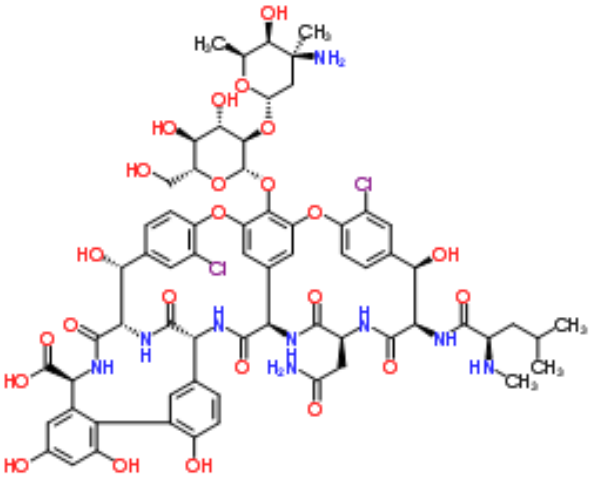
Molymod MMS-007 Molecular Model Teacher Set for Biochemistry
2
£64.00 Prime

DNA Model - Advanced miniDNA 12 Layer (Base pair) Model Kit
13
£20.00 Prime

What molecules are related to vancomycin?

- What could I use as a drug molecule instead of this?
- What other molecules are about the same size and shape?
- What other molecules have the same functional groups?
- What else could I use for this application?
- What else could I use with similar or better properties?
- What else could I replace this molecule with in this reaction?
- What can I synthesise this molecule from?
- What can I use as a solvent for this molecule?
- What other molecules might pack together the same as this when crystallised?
- What shall I work on next?
- What are my colleagues (competitors) working on?
- Or sometimes just “surprise me!”
- …

What about a molecule recommender?



Vancomycin

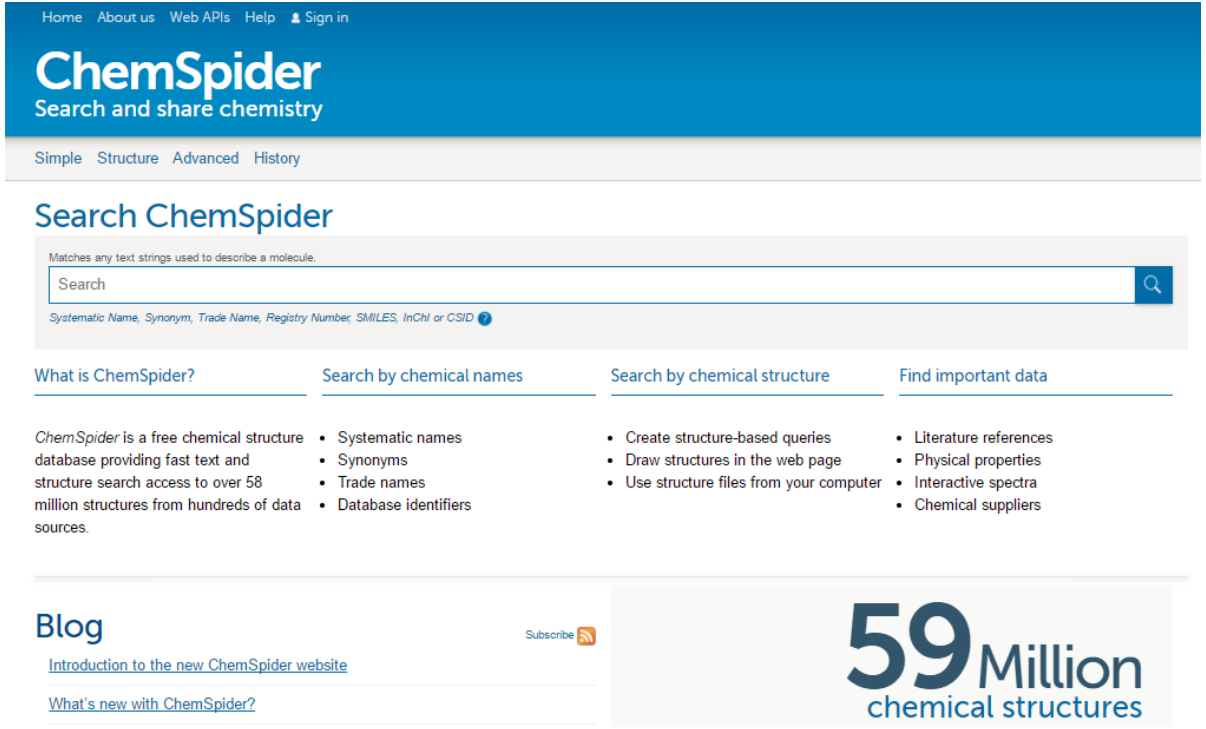
What other molecules are “related” to vancomycin?

Use Data Science…

RSC Data Science

We have access to:

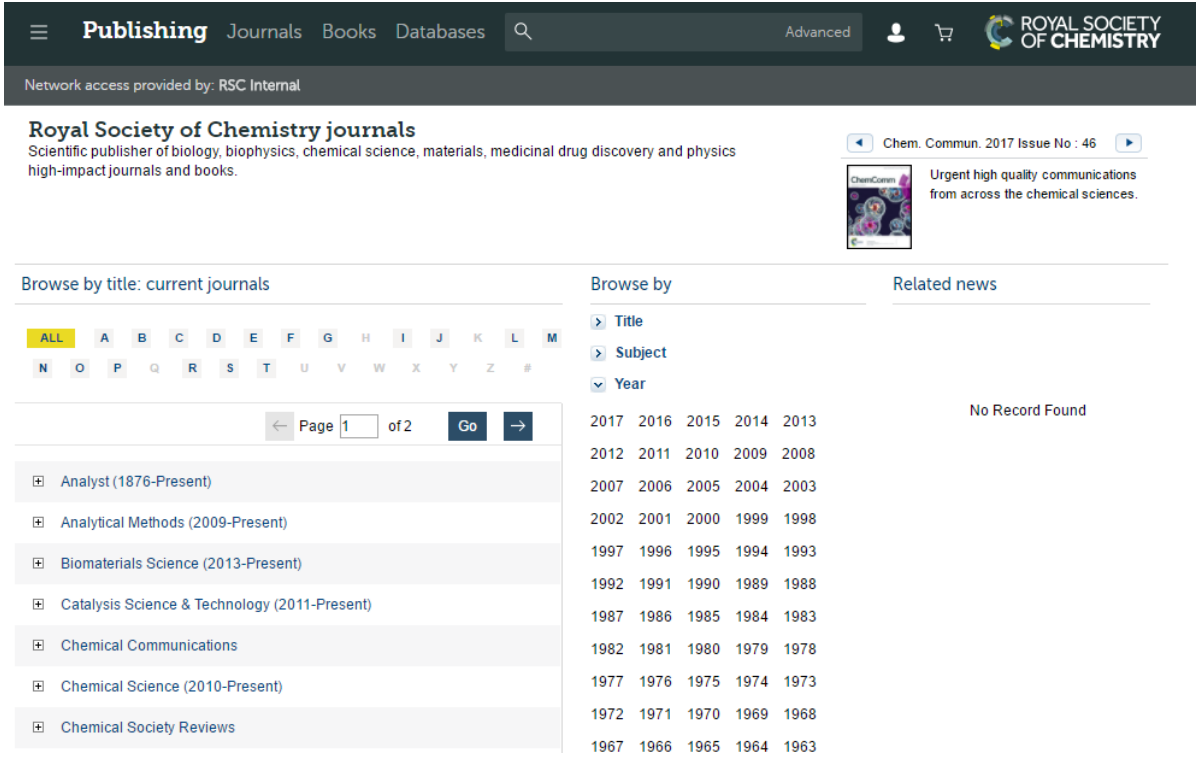
- ChemSpider



RSC Data Science

We have access to:

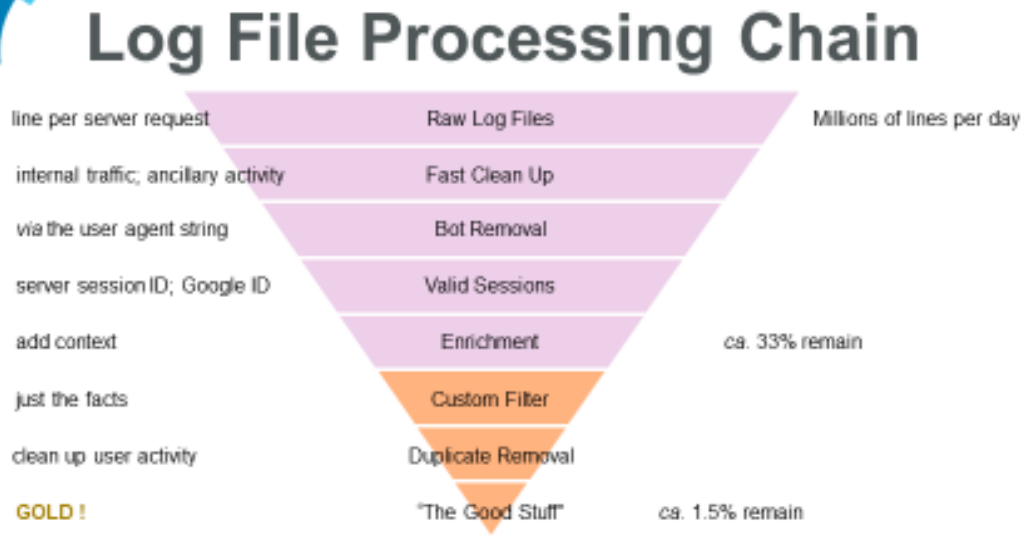
- ChemSpider
- RSC publishing



RSC Data Science

We have access to:

- ChemSpider
- RSC publishing
- logs




```
2016-06-24 00:05:07 192.168.0.1 pubs.rsc.org – GET /en/content/articlepdf/2007/sm/b704827k - - - XXX.XXX.XXX.XXX -
Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/50.0.2661.102+Safari/537.36
ShowEUCookieLawBanner=true;+X-Mapping-hhmaobcf=5EFF013F0F2EB5C7479A967277AFB2F4;+ASP.NET_SessionId=tzmjdkojxv2jcqh25omqerui;
+Branding=50000XXX;+AuthSystemSessionId=261e0a91-7d73-4fd7-9380-e73e298d6047;+__utmt=1;+__utma=1.2022872114.1464909160
.XXXXXXXXXX.XXXXXXXXXX.X;+__utmb=X.X.X.XXXXXXXXXXXX;+__utmc=1;+__utmz=X.XXXXXXXXXX.X.X.utmcsr=google|utmccn=(organic)|utmcmd
=organic|utmctr=(not%20provided);+iislog-host=pubs.rsc.org;+iislog-s-ip=172.30.229.101
http://pubs.rsc.org/en/Content/ArticleLanding/2007/SM/b704827k - 200 - - - 409353 - -
```

What molecules are related to X

- Cheminformatics similarity

  “X has structural features in common with...”

- Human behaviour

  “users who looked at X also viewed...”

- Published literature

  “papers mentioning X also mentioned...”

What molecules are related to X

- Cheminformatics similarity

  “X has structural features in common with...”

- Human behaviour

  “users who looked at X also viewed...”

- Published literature

  “papers mentioning X also mentioned...”

Data sets

We take privacy very seriously!

- Behaviour:
  - ChemSpider web logs (2015-2016)
  - molecules grouped by user IDs
  - anonymised, aggregated
- Literature:
  - RSC corpus (2000-2012)
  - text-mined for chemical compounds
  - molecules grouped by article
- Combine:
  - Must appear twice in both sets
  - Total of ca. 20K molecules

Methods

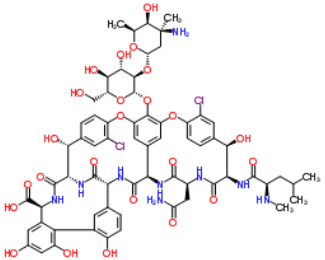
- Distance measures for pairs of molecules:
  - Fingerprinting: Dice Coefficient
  - Literature and Behaviour: Mean-square contingency coefficient φ
- Clusters using Affinity Propagation
  - Number of clusters decided by the process
  - Each cluster has exemplar – the “best example”
  - Implemented with Concurrent_AP Python package
- Display clusters
  - Interface using Django Python package

What molecules are related to…

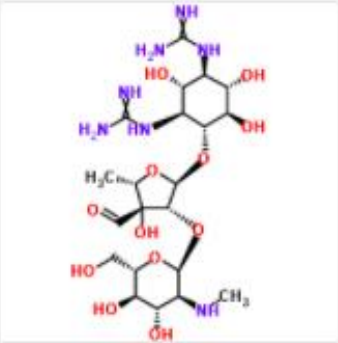
Method 3: Behaviour

Molecule recommender
Data Science

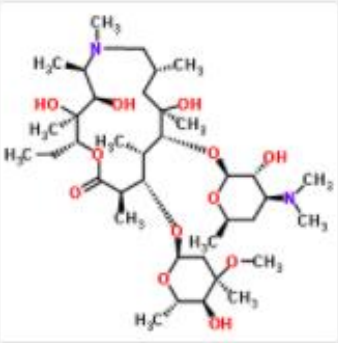
Vancomycin



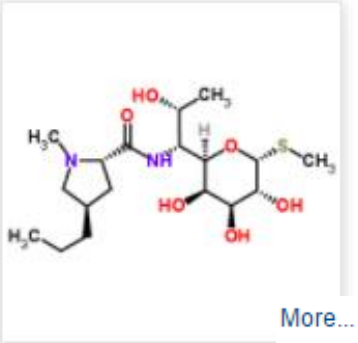
Behaviour ⓘ



streptomycin



azithromycin



lincomycin

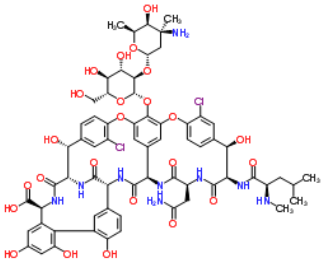
(antibiotics)

What molecules are related to…

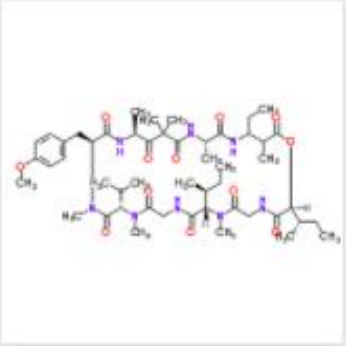
Method 4: Literature

Molecule recommender
Data Science

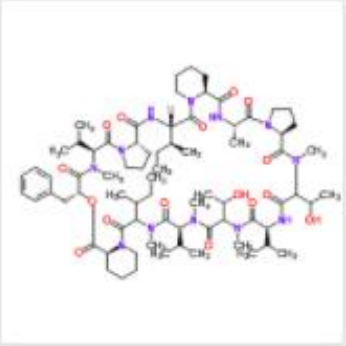
Vancomycin



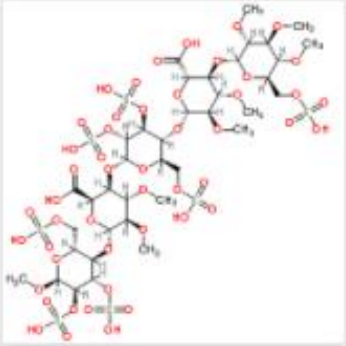
Literature ⓘ



majusculamide C



petriellin A



idraparinux

More...

Comparing methods

- We have lots of related molecules by different methods – do we need to display all of them?
- Compare similarities of clusters and rankings…

Compare rankings: Mantel permutation test

	Behaviour	Literature	Morgan	Topology
Behaviour	—	0.044	0.015	0.011
Literature	0.044	—	0.036	0.030
Morgan	0.015	0.036	—	0.110
Topology	0.011	0.030	0.110	—

- Some correlations are significant but none are strong:
  - methods are contextually distinct
- Cheminformatics fingerprinting methods correlated most significantly (expected)
- Literature more loosely correlated with Behaviour and Cheminformatics methods
- Behaviour most distinct from Cheminformatics methods

Comparing methods

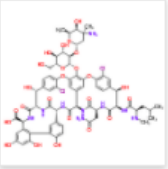
- We have lots of related molecules by different methods – do we need to display all of them?

- Yes!
- They're all contextually distinct – best used in combination?
- Investigate via user testing

Molecule recommender

Data Science

Search



ChemSpider 14253

Authors

Michèle R. Prinsep Dudley H. Williams John W. Blunt Murray H. G. Munro Brent R. Copp Peter T. Northcote Bing Xu Jonathan B. Spencer Hirokazu Arimoto

Keywords

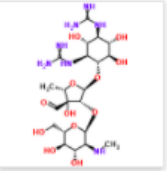
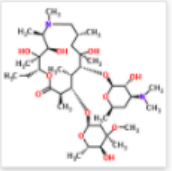
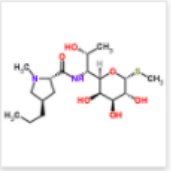
antibiotic antibiotics aureus methicillin staphylococcus bacteria bacterial resistant antibacterial

Categories

Microbiology Natural products Bioorganic chemistry Drug Discovery Natural Products Biotechnology Pharmacology Bioorganic

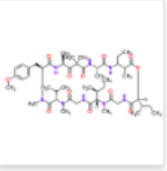
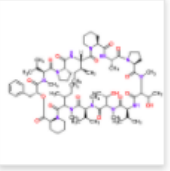
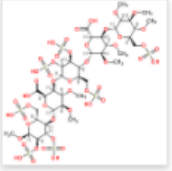
Other molecules related by:

Behaviour ⓘ

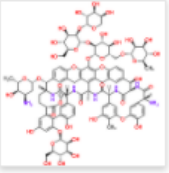
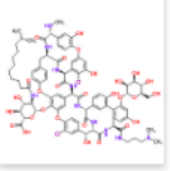
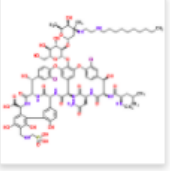
More...

Literature ⓘ

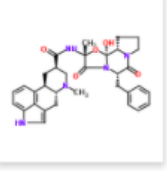
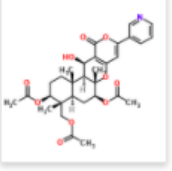
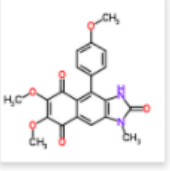
More...

Cheminformatics M ⓘ

More...

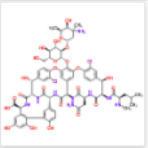
Cheminformatics T ⓘ

More...

Beyond chemical contexts

Vancomycin



ChemSpider 14253

Authors

Michèle R. Prinsep Dudley H. Williams John W. Blunt Murray H. G. Munro Brent R. Copp Peter T. Northcote Bing Xu Jonathan B. Spencer Hirokazu Arimoto

Keywords

antibiotic antibiotics aureus methicillin staphylococcus bacteria bacterial resistant antibacterial

Categories

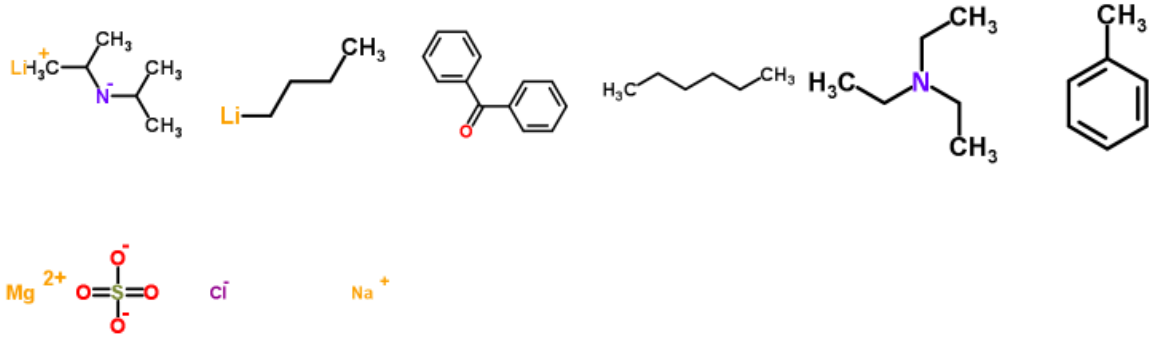
Microbiology Natural products Bioorganic chemistry Drug Discovery Natural Products Biotechnology Pharmacology Bioorganic

Molecule recommender

Data Science

Search

Jonathan Goodman



Authors

Jonathan M. Goodman

Keywords

www.ch.cam.ac.uk 6-31g** opls aa gromacs mm2* situ31p jaguar 180(2) macromodel

Categories

Biophysics Total synthesis Simulations Stereochemistry Catalysis Quantum and Theoretical

Further reading

Further reading

1. Suode Zhang, Samran Prabpai, Palangpon Kongsaeree and Per I. Arvidsson, 'Poly-N-methylated α-peptides: synthesis and X-ray structure determination of β-strand forming foldamers', *Chem. Commun.*, 2006, doi:10.1039/b513277k
2. Karen L. Sutton, Claudia A. Ponce de Leon, Kathryn L. Ackley, Richard M. C. Sutton, Apryll M. Stalcup and Joseph A. Caruso, 'Development of chiral HPLC for selenoamino acids with ICP-MS detection: application to selenium nutritional supplements', *Analyst*, 2000, doi:10.1039/a907847i
3. Karen L. Sutton, Richard M. C. Sutton, Apryll M. Stalcup and Joseph A. Caruso, 'A comparison of vancomycin and sulfated beta-cyclodextrin as chiral selectors for enantiomeric separations of selenoamino acids using capillary electrophoresis with UV absorbance detection', *Analyst*, 2000, doi:10.1039/a908558k
4. Young-Ger Suh, Dong-Yun Shin, Kyung-Hoon Min, Soon-Sil Hyun, Jae-Kyung Jung and Seung-Yong Seo, 'Facile construction of the oxaphenalene skeleton by peri ring closure. Formal synthesis of mansonone F', *Chem. Commun.*, 2000, doi:10.1039/b001859g
5. Karen Ochoa Lara, Carolina Godoy-Alcántar, Alexey V. Eliseev and Anatoly K. Yatsimirsky, 'Recognition of α-amino acid derivatives by N,N'-dibenzylated S,S-(+)-tetrandrine', *Org. Biomol. Chem.*, 2004, doi:10.1039/b402698e
6. Hefziba T. ten Brink, Dirk T. S. Rijkers, Johan Kemmink, Hans W. Hilbers and Rob M. J. Liskamp, 'Ring-closing metathesis for the synthesis of side chain knotted pentapeptides inspired by vancomycin', *Org. Biomol. Chem.*, 2004, doi:10.1039/B408820D
7. Robert A. Hill and Andrew Sutherland, 'Hot off the press', *Nat. Prod. Rep.*, 2004, doi:10.1039/b413749n
8. Robert A. Hill and Andrew Sutherland, 'Hot off the press', *Nat. Prod. Rep.*, 2004, doi:10.1039/b403197k
9. Wen-Yong Lou, Min-Hua Zong, Hong Wu, Ruo Xu and Ju-Fang Wang, 'Markedly improving lipase-mediated asymmetric ammonolysis of d,l-p-hydroxyphenylglycine methyl ester by using an ionic liquid as the reaction medium', *Green Chem.*, 2005, doi:10.1039/b502716k

Next step

- Molecular Recommender
  - User evaluation
    - come and find me and try it (especially if you've published in RSC publications)!
      - which method results do you find most useful?
      - how many results would you like to see – just one (I feel lucky)/ or lots
      - where would you like to see this tool?
      - what other features would you like to see? e.g. reactions that this molecule takes part in?
- Better chemical name extraction => ChemListem

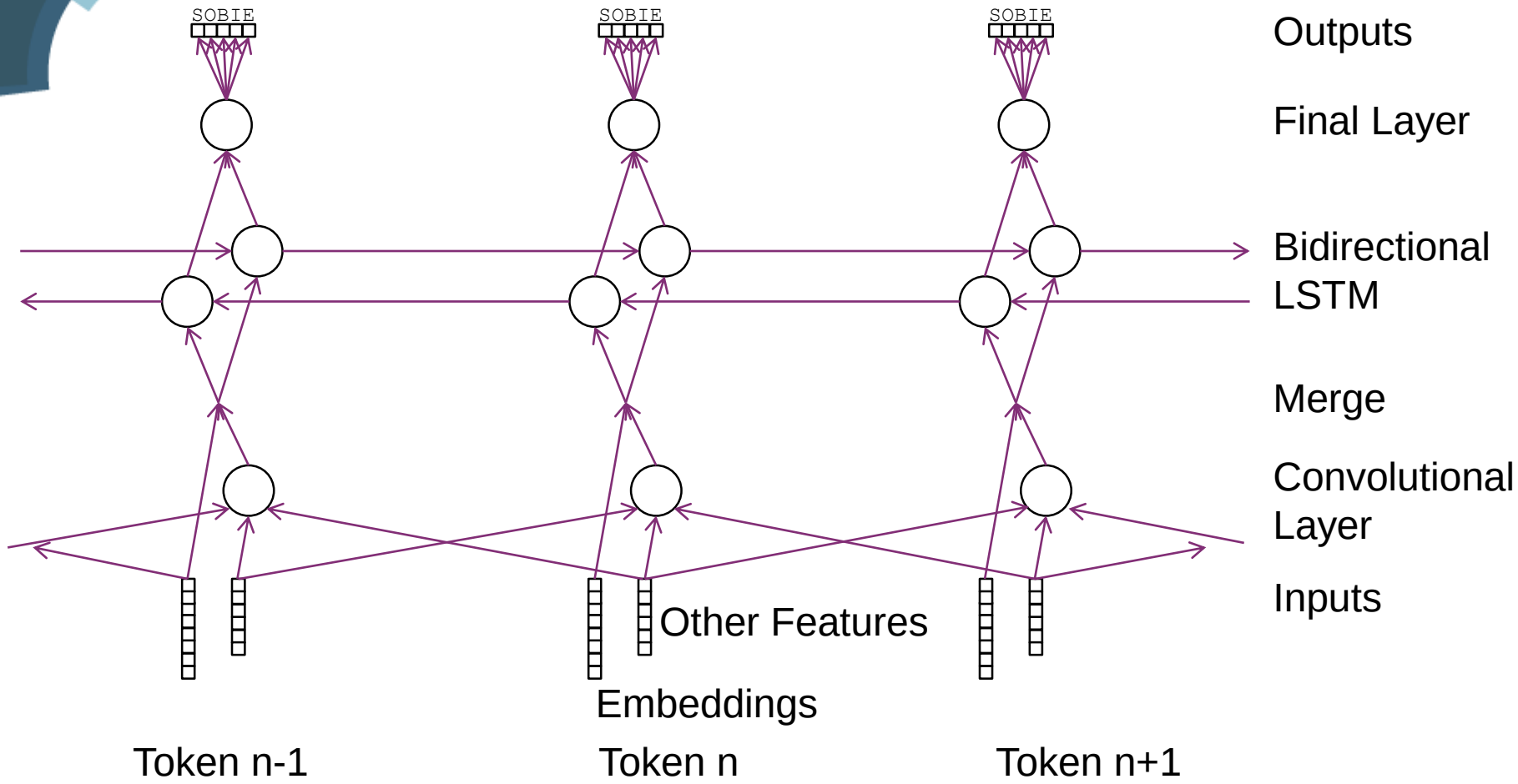
Chemlistem
Named Entity Recognition (NER)

- Participated in *public, competitive evaluation* extracting chemical names from patents:
- **BioCreative** V.5 (Critical Assessment of Information Extraction in Biology) community-wide effort with the aim of evaluating biomedical text mining and information extraction tools, submitted and evaluated using **Becalm** platform
- **CEMP** (chemical entity mention in patents) task
- Using *deep learning* techniques – recurrent artificial neural networks

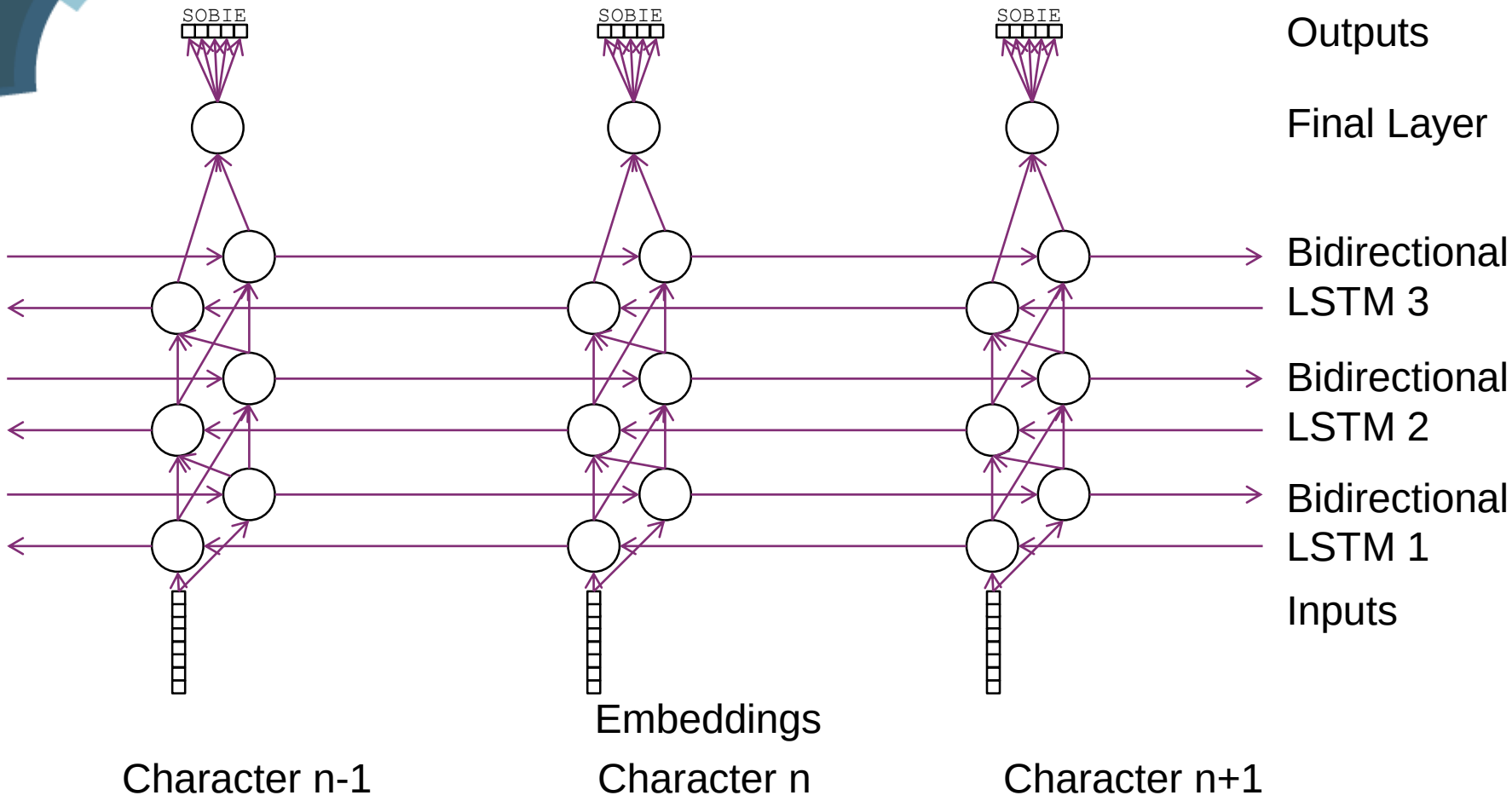
Chemlistem Methods

- Compared 3 methods:
  - “Traditional” Conditional Random Fields CRF translated to deep learning:
    - **Tokenises** using Oscar => words
    - Maps each word => GloVe “word **embeddings**” (n-dimensional vectors)
    - Rich per-token feature set
    - Uses external resources (e.g ChEBI and ChemSpider chemical name dictionaries)
    - Single recurrent bidirectional LSTM (Long Short-Term Memory) layer
  - Minimalist approach:
    - Character level – no tokeniser
    - Character embeddings only
    - No features
    - No external resources
    - Three recurrent bidirectional LSTM (Long Short-Term Memory) layers
  - Ensemble combination of previous two methods:
    - Run Traditional and Minimalist systems with a low threshold => generate 2 lists of entities
    - Combine scores of entities in lists and apply threshold of 0.475

“Traditional” neural network



“Minimalist” neural network



Example SOBIE output - Traditional

	in	methyl	ethyl	ketone	and
S (singleton)	0.002	0.250	0.040	0.001	0.001
O (other)	0.998	0.008	0.010	0.300	0.999
B (beginning)	0.0	**0.700**	0.150	0.004	0.0
I (inside)	0.0	0.040	**0.550**	0.045	0.0
E (end)	0.0	0.002	0.250	**0.650**	0.0

Results

System	Offical F-score	Official Precision	Official Recall	Internal F-score	Internal Precision	Internal Recall
Trad	.8919	.8867	.8971	.8703	.8648	.8758
Minimal	.8901	.8865	.8936	.8664	.8479	.8858
Ensemble	**.9032**	**.9002**	**.9062**	**.8807**	**.8646**	**.8976**

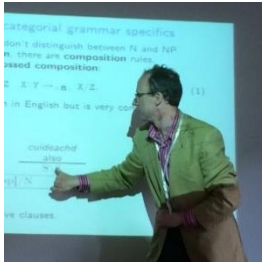
- Participating in *public, competitive evaluation* (BioCreative V.5 Becalm)
  - 0.9006 precision, 0.9062 recall, .9032 F
  - 3rd place out of 17 (0.1% off 1st, “differences in the top three weren’t statistically significant”)
  - inter-annotator agreement studies on manual annotators were at 90% (human level)

Chemlistem

- Peter Corbett, John Boyle. “Chemlistem - chemical named entity recognition using recurrent neural networks” (2017) http://www.biocreative.org/media/store/files/2017/BioCreative_V5_paper8.pdf
- Open source:
  - http://bitbucket.org/rscapplications/chemlistem
  - `pip install chemlistem`

Acknowledgements

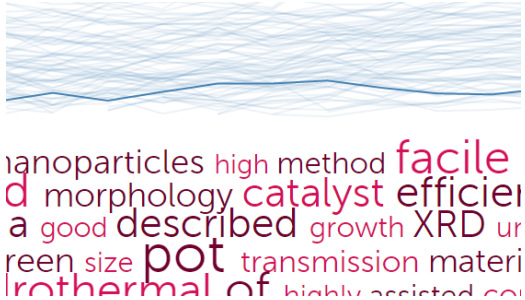
Colin Batchelor – Molecule Recommender development

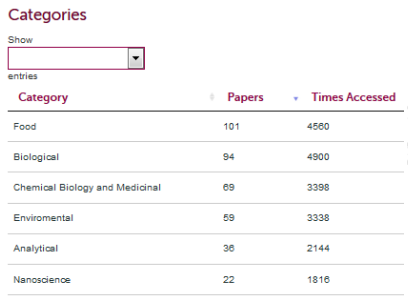


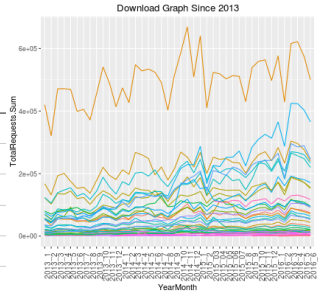
Peter Corbett – Chemlistem development

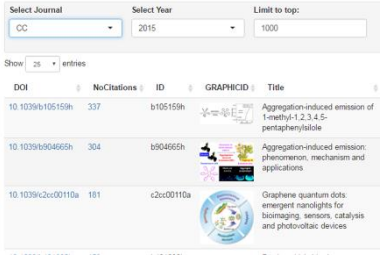
RSC Data Science Team

www.rsc.org/data-science



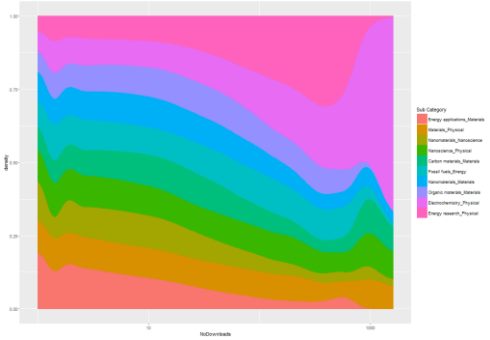


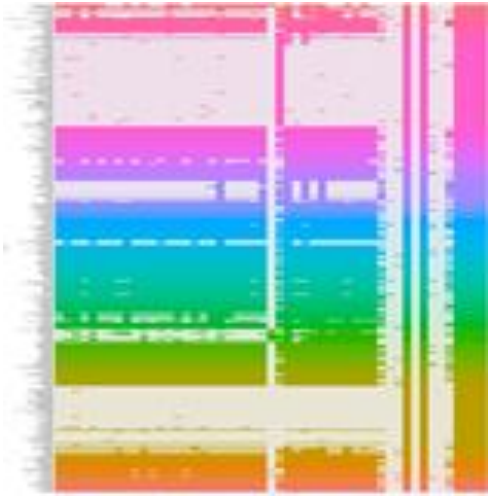


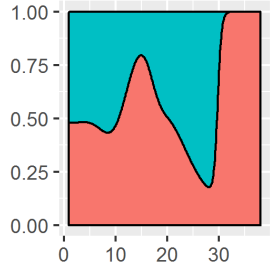


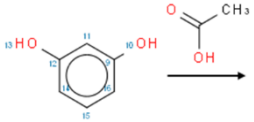
References

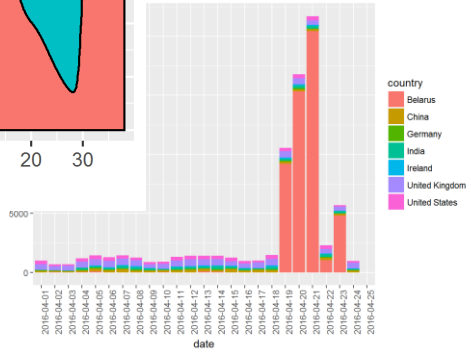
1. Y. Wang and L. Chen, *Nanomed.: Nanotechn*
2. A. P. Alivisatos, *Science*, 1996, **271**, 933–937
3. A. Priyam, D. E. Blumling and K. L. Knappe
4. F. Guo, Y. Zhu, X. Yang and C. Li, *Mater. Ch*
5. P. Wang, Y. Zhu, X. Yang, C. Li and H. L. Du
6. Y. Li, Y. Zhu, X. Yang and C. Li, *Cryst. Grov*

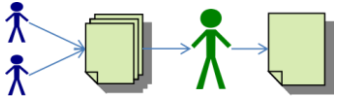


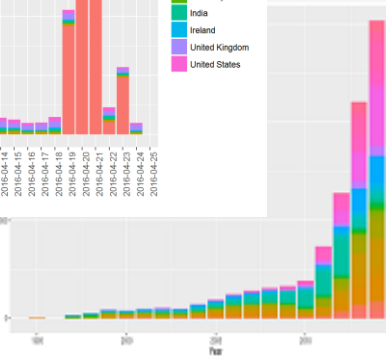












Any questions?