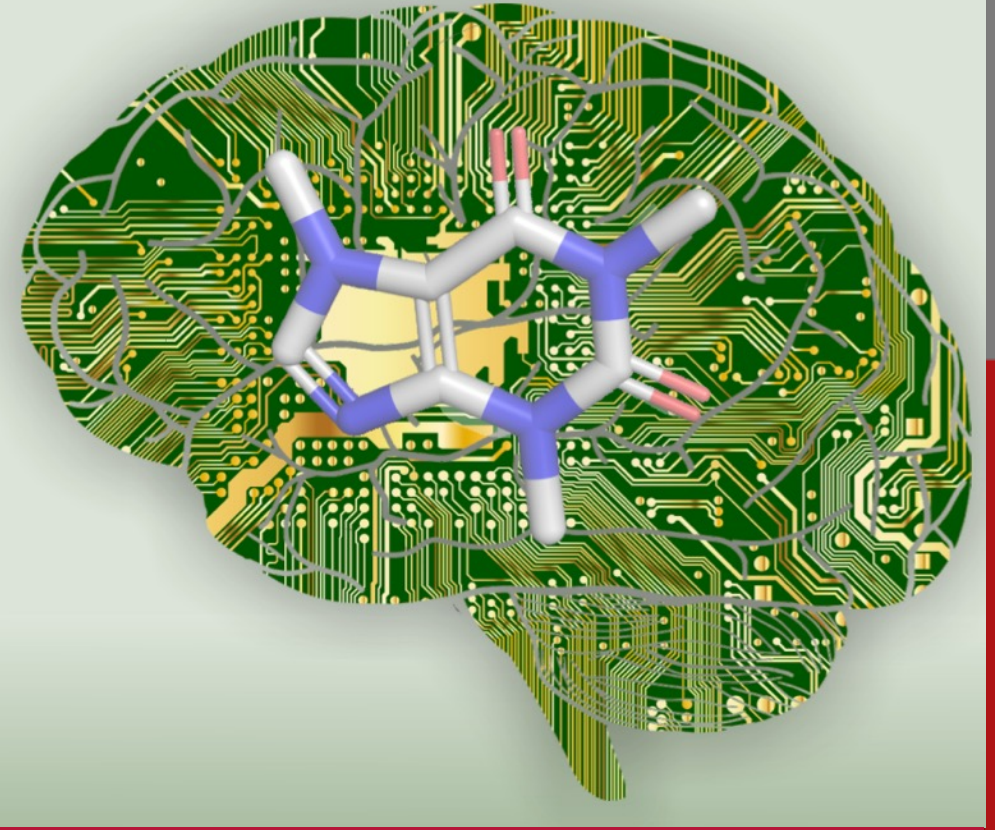


David Winkler | La Trobe Institute for Molecular Science (LIMS) | Monash Institute of Pharmaceutical Sciences (MIPS) | School of Pharmacy, University of Nottingham

AI and Machine Learning for Next Generation Drugs and Materials

LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

MONASH University Institute of Pharmaceutical Sciences

The University of Nottingham

HORIZON 2020

MARIE CURIE ACTIONS

latrobe.edu.au

Collaborators and acknowledgements

SARS-CoV-2: Nik Petrovsky, Sakshi Piplani and Puneet Singh (Vaxine), Peter Winn, Dennis Ward, Harinda Rajapaksha (Oracle Cloud Systems)
Cancer biomarkers: Sanduru Krishnan, Darren Creek, Dovile Anderson, David Rudd, Nico Voelcker (Monash) Ehud Hauben, Chandra Kirana, Guy Maddern, Kevin Fenix (Adelaide and Basil Hetzel Institute)
2D materials: Olexander Isayev (Carnegie Mellon), Joe Shapter (UQ), Amanda Ellis, Peter Sherrell, Nick Shepelin, Alexander Corletto (Melbourne), Marco Fronzi, Mike Ford (UTS)
OPVs and Photocatalysts: Haoxin Mai, Tu Le, Dehong Chen, Rachel Caruso (RMIT), Takashi Hisatomi (Shinshu University) , Kazunari Domen (Tokyo)
Biomaterials: Manuel Romero, Jeni Luckett, Grazziela Figueredo, Alessandro Carabelli, David Scurr, Andrew Hook, Jean-Frédéric Dubern, Amir Ghaemmmaghami, Morgan Alexander, Paul Williams (Nottingham), Aliaksei Vasilevich, Steven Vermeulen, Jan de Boer (Eindhoven) Aurélie Carlier (Maastricht), Dan Anderson, Bob Langer (MIT), Molly Stevens, (Imperial), Eileen Gentleman (Kings), Irene Yarovsky (RMIT)
Fluorescent polymers, perovskite solar cells, catalysts: Nas Meftahi, Tu Le, Andrew Christofferson, Salvi Russo and colleagues, Rachel Caruso, Hoaxin Mai and colleagues (RMIT)
Batteries and green corrosion inhibitors: Mikhail Zheludkevich, Christian Feiler, Sviatlana Lamaka, Tim Würger, Rolf Meißner (Helmholtz-Zentrum Hereon), Tony Hughes (CSIRO)
Surface methods: Paul Pigram, Wil Gardner, Sarah Bamford, Robert Maddiona and team (La Trobe), Ben Muir (CSIRO), Davide Ballabio (Milan)
QSAR, machine learning, regenerative medicine, materials: Tu Le, Frank Burden, Vidana Epa, Phuc Ung, Anna Tarasova, David Haylock, Susie Nilsson, Jacinta White Brian Dalrymple, Gene Wijffels and team (CSIRO)

Bernstein Strecker:

$$R\text{-}CH_2OH \xrightarrow{UV} R\text{-}\overset{O}{\overset{\|}{C}}H \xrightarrow[HCN\ \ NH_3]{UV} H_2N\text{-}\underset{\underset{R}{|}}{CH}\text{-}CN \xrightarrow{Hydrolysis} H_2N\text{-}\underset{\underset{R}{|}}{CH}\text{-}COOH$$

Woon Radical-Radical:

$$R\text{-}CN \xrightarrow[2H]{UV} R\text{-}CHNH \xrightarrow[H]{UV} R\text{-}CH_2NH_2 \xrightarrow[COOH]{UV} H_2N\text{-}\underset{\underset{R}{|}}{CH}\text{-}COOH$$

$$CH_3OH + H_2O \xrightarrow{UV} COOH$$

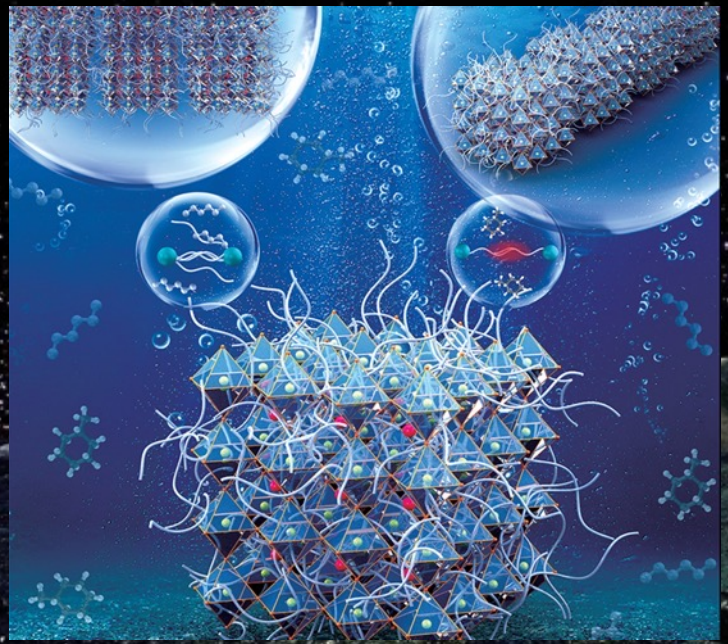
R=H methanimine and glycine
R=Me ethanimine and alanine

Elsila Modified Nitrile:

$$R\text{-}CN \xrightarrow[2H]{UV} R\text{-}CHNH \xrightarrow[HCN]{UV} NC\text{-}\underset{\underset{R}{|}}{CH}\text{-}NH_2 \xrightarrow{Hydrolysis} HOOC\text{-}\underset{\underset{R}{|}}{CH}\text{-}NH_2$$

Observable universe ~10^{80} particles of matter

Materials space ~10^{100}



“Prediction is difficult, especially the future.”

“But I think mathematics will greatly change its aspect.”

Niels Bohr

LIMS1 building, La Trobe University

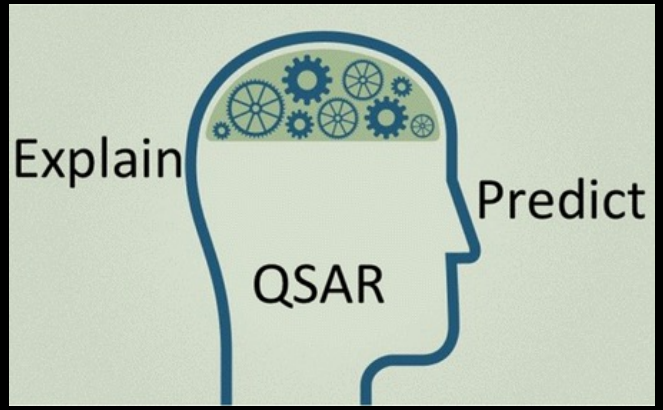
Outline

- Improving QSAR using machine learning – feature selection, Bayesian NN, RVM, validation, feature importance, overfitting
- Sparse methods – stem cell markers (Asymmetrex), Sr MSC (RepGen), CRC markers
- Tripeptide motifs – as design tools, novel antibiotics (Betabiotics), myelofibrosis drugs
- Molecular design – SARS-CoV-2 origin and COVID-19 drugs
- New applications –biomaterials, stem cell bioreactors topographical biomaterials, fluorescent polymers, surface chemistry analysis, 2D and porous materials, photovoltaics, catalysts, corrosion and battery technologies

Improving QSAR using machine learning

Frank Burden,
Burden Index

Hansch and Fujita



JOURNAL OF CHEMICAL INFORMATION AND MODELING
Perspective
pubs.acs.org/jcim

Understanding the Roles of the "Two QSARs"

Toshio Fujita[†] and David A. Winkler[#,‡,§,∥,⊥]

[†]Professor Emeritus at Kyoto University, 38-1 Iwakura-Miyakecho, Kyoto, Japan 606-0022
[‡]CSIRO Manufacturing, Bag 10, Clayton South MDC 3169, Australia
[§]Monash Institute of Pharmaceutical Sciences, 392 Royal Parade, Parkville 3052, Australia
[∥]Latrobe Institute for Molecular Science, Latrobe University, Bundoora 3086, Australia
[⊥]School of Chemical and Physical Sciences, Flinders University, Bedford Park 5042, Australia

ABSTRACT: Quantitative structure–activity relationship (QSAR) modeling has matured over the past 50 years and has been very useful in discovering and optimizing drug leads. Although its roots were in extra-thermodynamic relationships within small sets of chemically similar molecules focused on mechanistic interpretation, a second class of QSAR models has emerged that relies on machine learning methods to generate models from large, chemically diverse data sets for predictive purposes. There has been a tension between the two groups of QSAR practitioners that is unnecessary and possibly counter-productive. This paper explains the difference in philosophy and application of these two distinct, but equally important, classes of QSAR models and how they can work together synergistically to accelerate the discovery of new drugs or materials

Quantitative structure activity relationship (QSAR) modeling is now more than 50 years old, and its utility has been shown in numerous research publications and by the number of new drug and agrochemical entities that have been developed with its aid. The method has evolved very substantially since the seminal linear regression QSAR models published by Hansch and Fujita.[1,2] Many new types of molecular descriptors have been developed, new mathematical methods such as neural networks,[3−8] support vector machines,[9−11] kernel regression,[12] and random forest[13,14] have been applied to mapping structure to activity, and QSAR has now incorporated 3D structures using field based methods like CoMFA and CoMSIA,[15,16] conformation, and chirality.[17−19] The method has therefore evolved steadily since the 1960s as has been well-summarized in numerous recent reviews of the history of QSAR methods.[20−24]

scientific meetings and is largely unpublished, but there have been a number of publications in the past decade or two that have also carried this debate. For example, Zefirov and Palyulin[26] discussed the general problem of descriptive versus predictive QSAR arguing that high quality correlations are not necessarily predictive. Trospha et al. subsequently summarized work by several QSAR practitioners who emphasized that "one of the most important aspects of QSAR modelling is the ability to interpret the models in physico-chemical and/or mechanistic sense" (pure or classical QSAR modellers).[27] However, some of these studies did not rigorously validate these mechanistically focused models, an essential step in good QSAR modeling. Tropsha et al. made the important point that QSPR models must be validated for predictive power before they are applied to predict, let alone explain, the structure−property relationships of biological, pharmaceutical, environmental, or any other

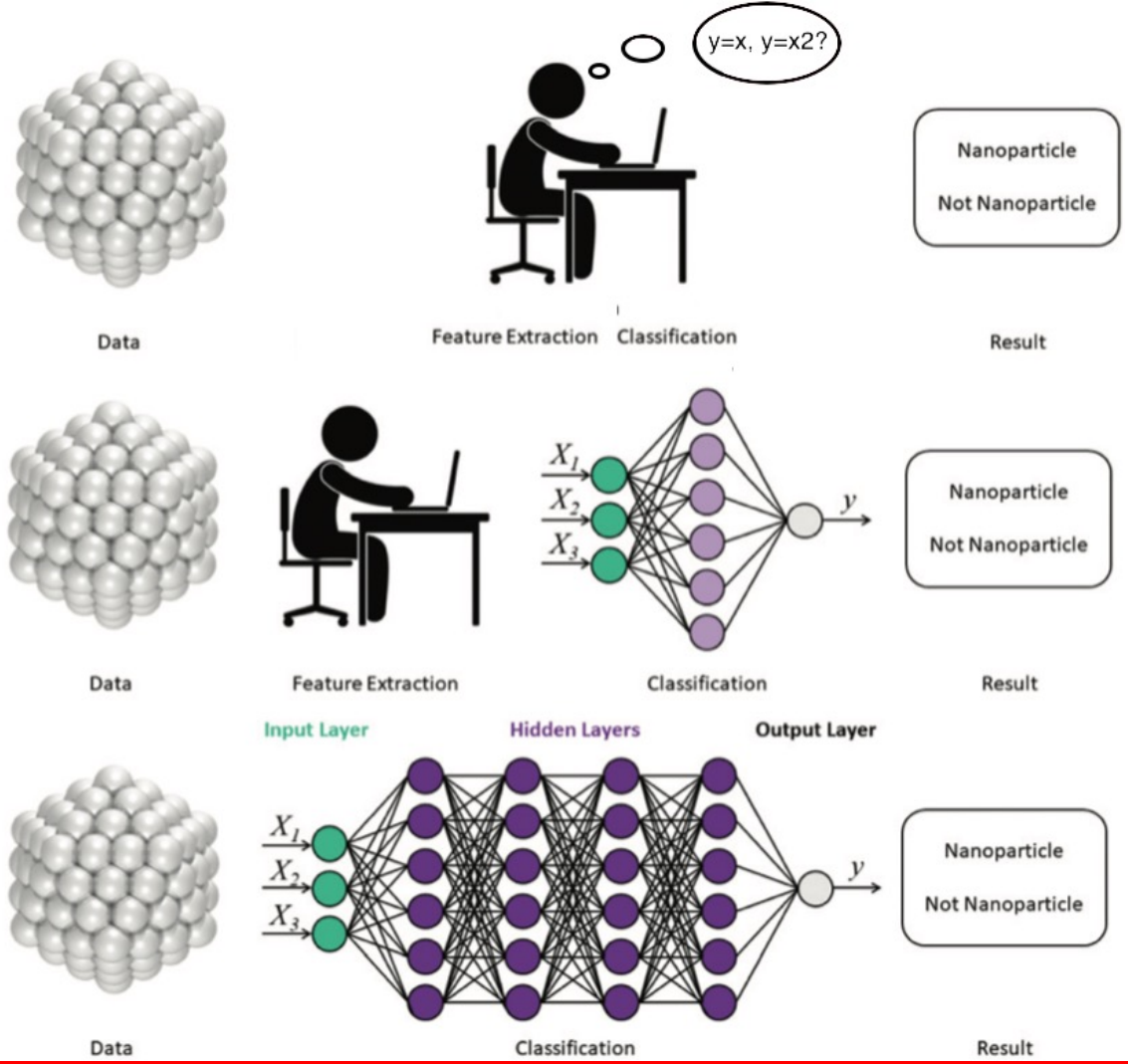
1960s
JACS 1964, 86, 1616–1626 (4000 citations)
JACS 1964, 86, 5175–5180 (2000 citations)
Nature 194, 178-180 (1300 citations)
2016
J. Chem. Inf. Model. 2016, 56, 269–274 (150 citations)

How modelling strategies have changed

1960-1990

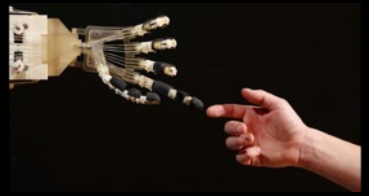
1990-2020

2010-now



Fujita; Winkler, Understanding the roles of the “two QSARs”, J. Chem. Inf. Mod. 2016 56 (2), pp 269; Barnard et al. Nanoscale, 2019, 11, 19190

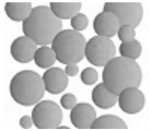
QSAR/machine learning unit operations



- Ideally acquire large, chemically diverse, high quality data sets
- Generate features (descriptors), mathematical representations of chemical entities
- Select relevant subsets of features in context-dependent way
- Generate the model linking features to desired propert(ies)
- Validate the model, and quantify its predictivity and domain of applicability
- Deploy the model – mechanisms, new predictions and designs, virtual screening

Our contribution to QSAR methods

LA TROBE INSTITUTE FOR MOLECULAR SCIENCE



Descriptors → Feature selection → Machine learning → Validation → Predictions

- Image-based
- Signature
- Information

- Sparse L1
- MLREM
- LASSO

- BRANN (Gaussian)
- ANN (Laplacia
- Relevanc Vector M

- Separate test set
- Cross va ion
- Exte test set

- Domain of applicability
- De novo
- Information

Burden and Winkler. Optimum QSA Sparse Bayesian Methods, *QSAR*

Burden and Winkler. Robust Artificial Neural Networks, *J. M*

Alexander et al. Beware of R^2: correct statistical usage in QSAR and QSPR studies, *J. Chem. Inf. Model*. 2015; **55**: 1316

Effective and context-aware feature selection

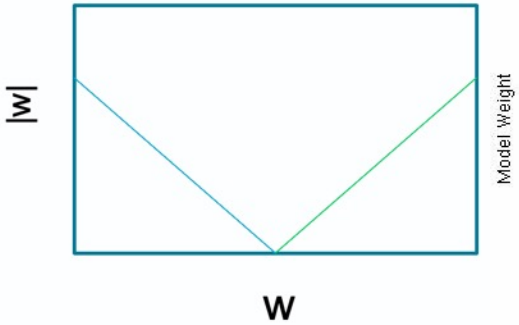
- Distribution of weights: Least squares (MLR) has a Gaussian prior

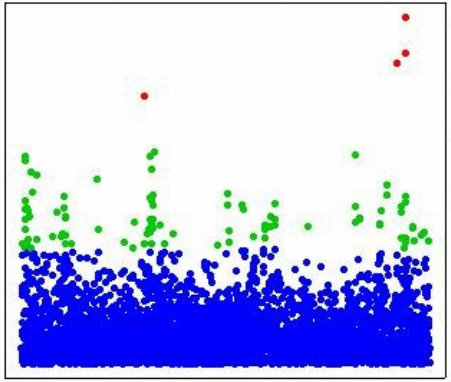
$$p(w \mid \alpha) = \prod_{i=1}^{N_V} \frac{\alpha}{2} \exp(-\alpha w_i^2)$$

- This can be replaced with a Laplacian prior

$$p(w \mid \alpha) = \prod_{i=1}^{N_V} \frac{\alpha}{2} \exp(-\alpha |w_i|)$$

- which effects the removal of uninformative weights by driving them to zero,





Burden, **Winkler**. *QSAR Comb Sci.* 2009; **28**: 645-653

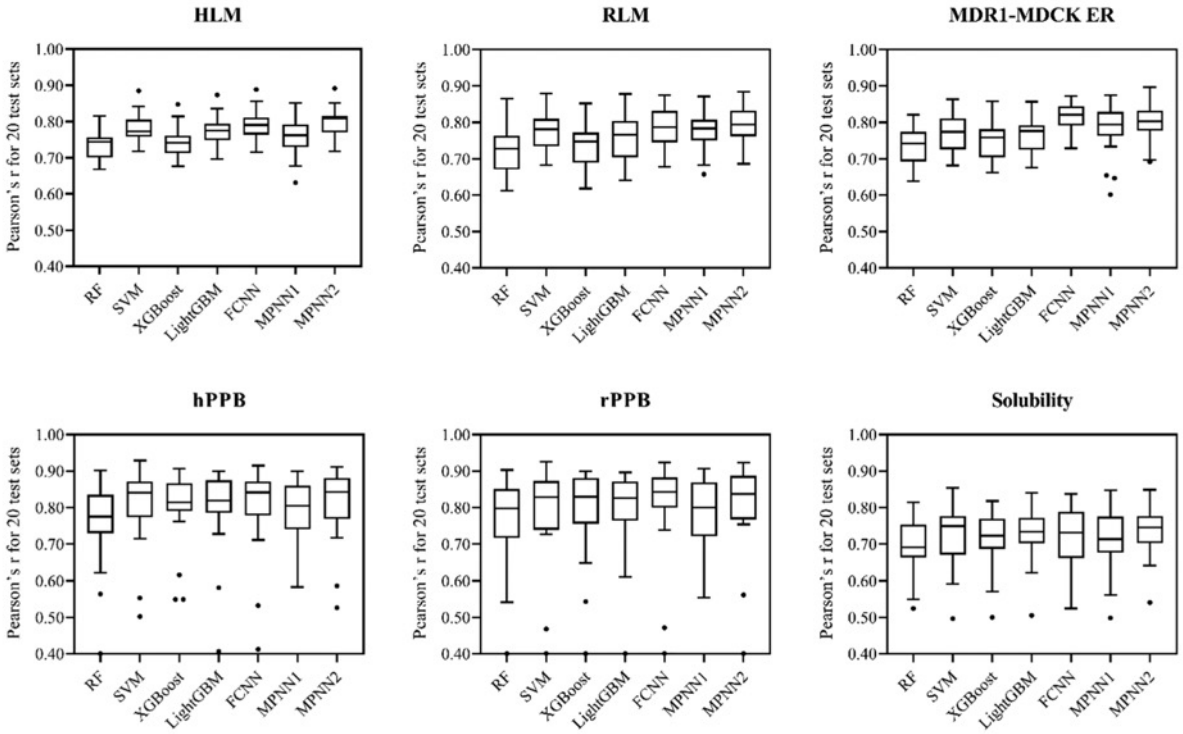
Figure 1. Frequency of a chance correlation with a r^2_{CV} value greater than 0.25, as a function of the numbers of rows and columns containing random data, using PLS. The figure is based on data in Table 1. See the text for further discussion.

- Oneto, et al. Do we real … *Neurocomputing* **2023**, *543*, 126227

Which machine learning algorithm is best?

Table 1. Comparison o
Green: BNN better than

Data set	D
3A4	C
CB1	b
DPP4	ir
HIVINV	ir
HIVPROT	ir
LOGD	lo
METAB	%
NK1	ir
OX1	ir
OX2	ir
PGP	tr
PPB	lo
RATF	lo



tter than random forest.

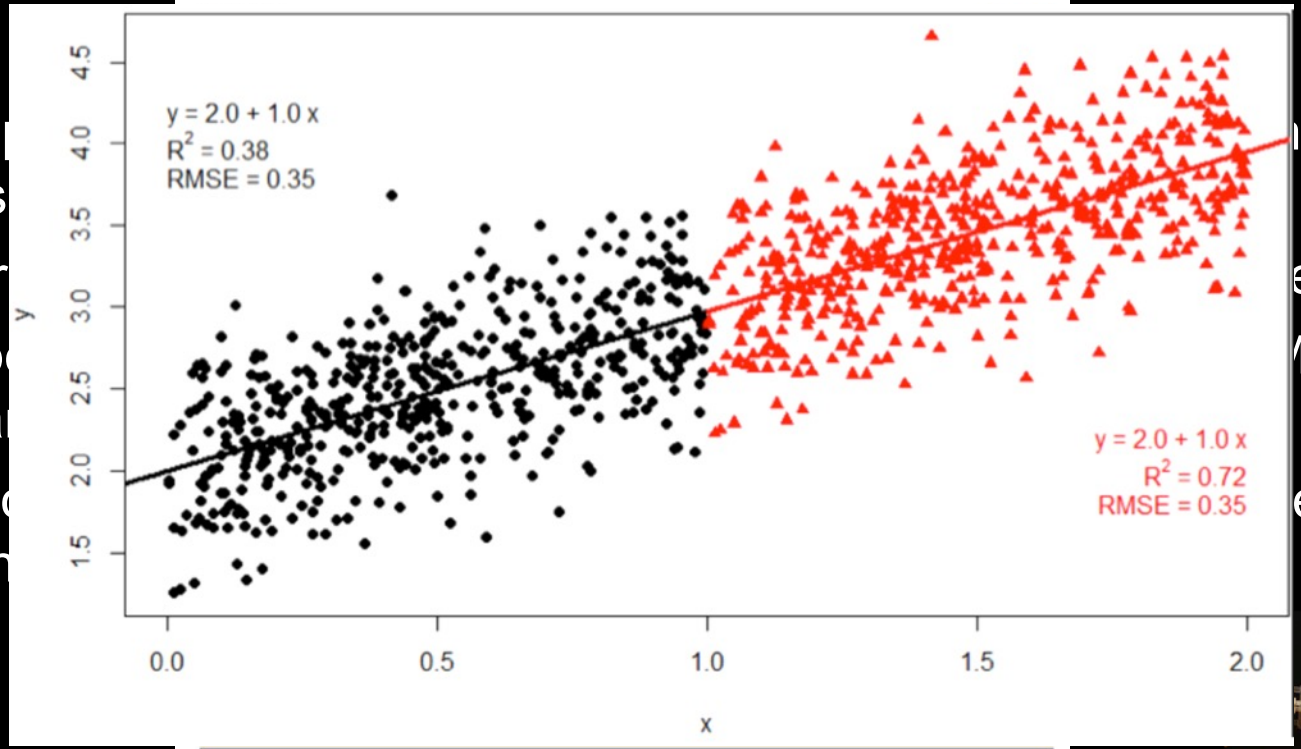
Test set SEP	
DNN	BNN
0.48	0.50
1.25	1.14
1.30	1.27
0.44	0.46
1.66	1.04
0.51	0.53
21.78	23.89
0.76	0.72
0.73	0.79
0.95	1.08
0.36	0.40
0.56	0.58
0.54	0.49

Fang, et al. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *J. Chem. Inf. Mod. 2023* **Article ASAP** DOI: 10.1021/acs.jcim.3c00160

D. A. Winkler, T. C. Le, *Mol. Inf.* **2017**, *36*, 1600118

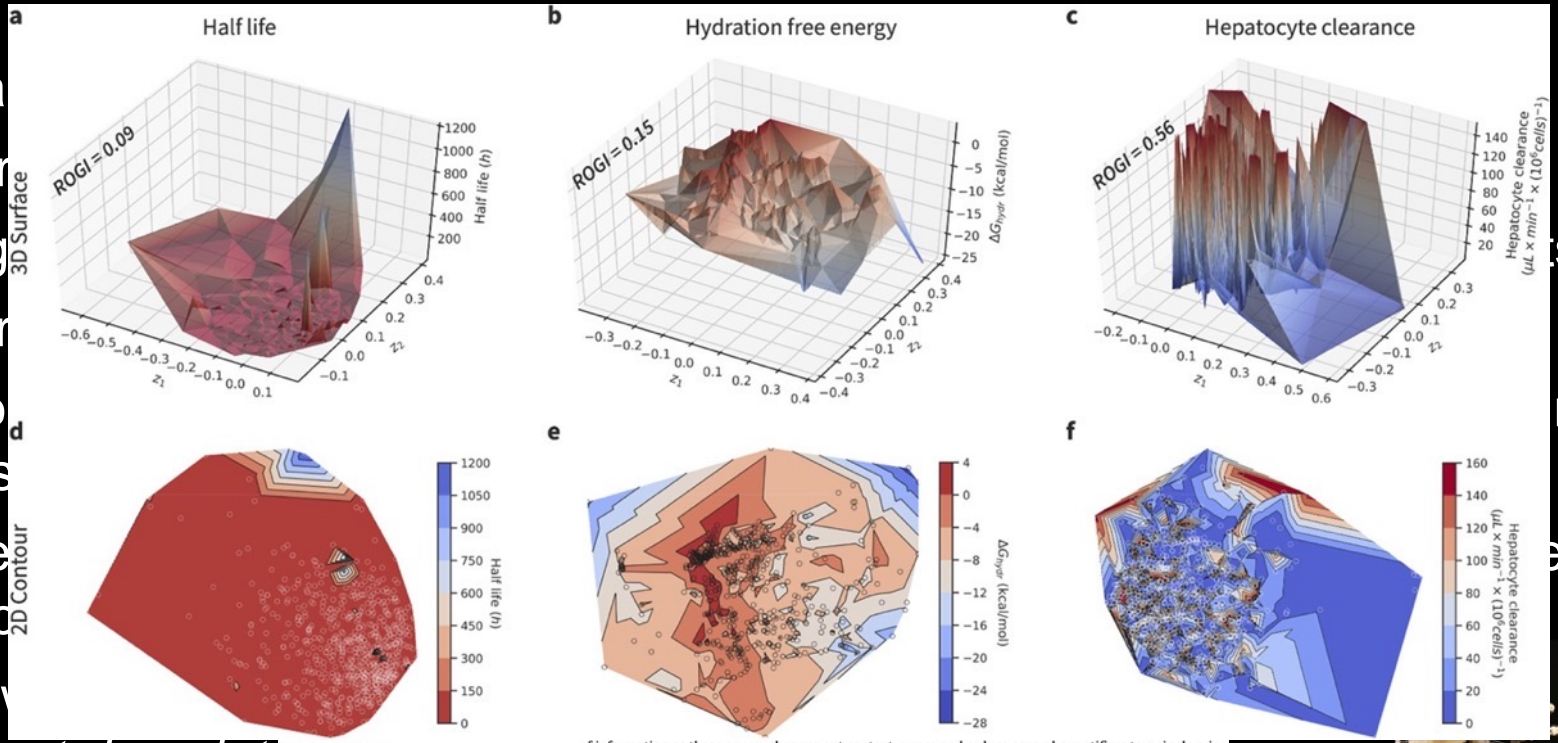
How to best validate models

- For small
- As data a , the method is
- For larger est
- Test set p MAE rather tha
- Ideally, m ents (work with



Assessing and using feature importance

- Ma
- For
- Sig ts
- For
- Mo nt for this
- We e more info
- Nev *about descript*

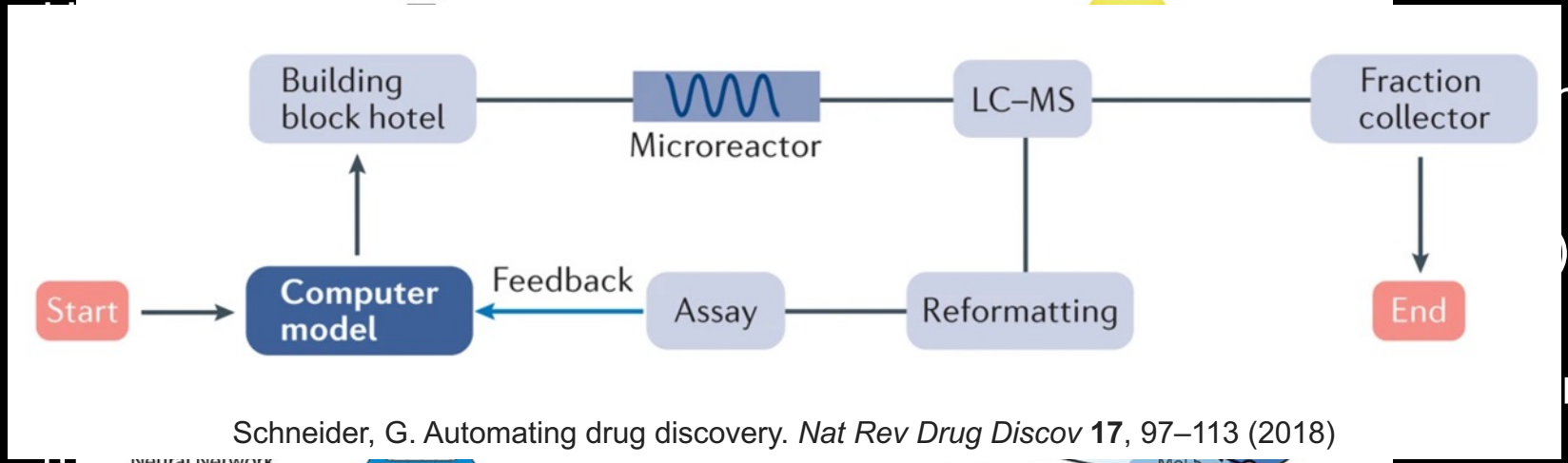


Interpretability
Machine learning
Responsible AI

of information as these approaches are not context-aware and reduce several quantifiers to a single crisp output. More importantly, their representation of "importance" as coefficients may be difficult to comprehend by end-users and decision makers. Here we show how the use of fuzzy data fusion methods can overcome some of the important limitations of crisp fusion methods by making the importance of features easily understandable.

Deploying the model

- U
-
-
- nt 1)
-
-) to
-
- netic
-
- D s for discovery



Schneider, G. Automating drug discovery. *Nat Rev Drug Discov* **17**, 97–113 (2018)

Gómez-Bombarelli et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, ACS Cent. Sci. 2018

Applying QSAR/ML to new areas

ROYAL SOCIETY OF CHEMISTRY

Chem Soc Rev

REVIEW ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Soc. Rev.*, 2020, **49**, 3525

QSAR without borders

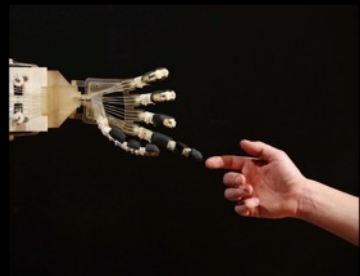
Eugene N. Muratov,[ab] Jürgen Bajorath,[c] Robert P. Sheridan,[d] Igor V. Tetko,[e] Dmitry Filimonov,[f] Vladimir Poroikov,[f] Tudor I. Oprea,[ghi] Igor I. Baskin,[jk] Alexandre Varnek,[j] Adrian Roitberg,[l] Olexandr Isayev,[a] Stefano Curtalolo,[m] Denis Fourches,[n] Yoram Cohen,[o] Alan Aspuru-Guzik,[p] David A. Winkler,[qrst] Dimitris Agrafiotis,[u] Artem Cherkasov*[v] and Alexander Tropsha*[a]

Prediction of chemical bioactivity and physical properties has been one of the most important applications of statistical and more recently, machine learning and artificial intelligence methods in chemical sciences. This field of research, broadly known as quantitative structure–activity relationships (QSAR) modeling, has developed many important algorithms and has found a broad range of applications in physical organic and medicinal chemistry in the past 55+ years. This Perspective summarizes recent technological advances in QSAR modeling but it also highlights the applicability of algorithms, modeling methods, and validation practices developed in QSAR to a wide range of research areas outside of traditional QSAR boundaries including synthesis planning, nanotechnology, materials science, biomaterials, and clinical informatics. As modern research methods generate rapidly increasing amounts of data, the knowledge of robust data-driven modelling methods professed within the QSAR field can become essential for scientists working both within and outside of chemical research. We hope that this contribution highlighting the generalizable components of QSAR modeling will serve to address this challenge.

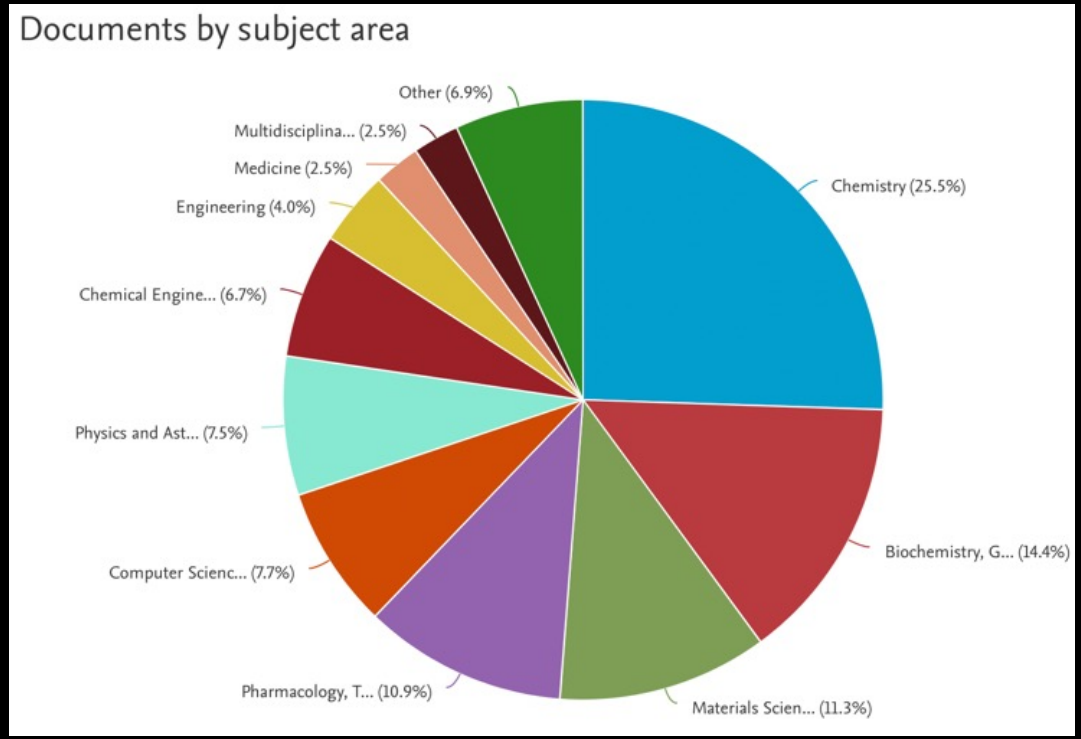
Received 7th February 2020

DOI: 10.1039/d0cs00098a

rsc.li/chem-soc-rev



Machine learning is very widely applicable

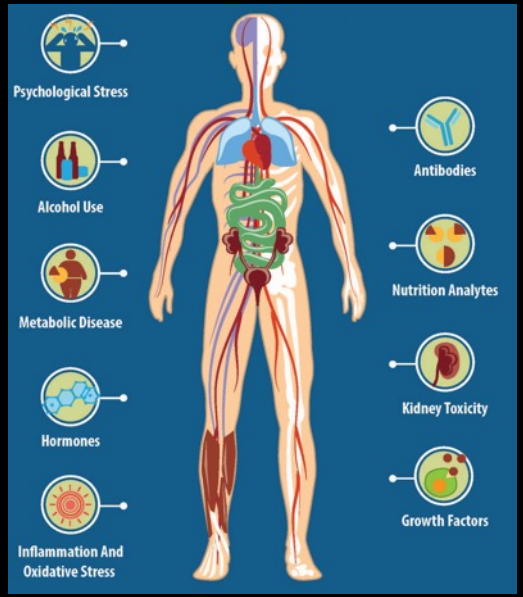


Recent projects

- Nanosafety
- 2D materials
- Porous materials
- Biomaterials
- Stem cells
- Corrosion/batteries
- Pandemics – vaccines/drugs
- OPVs
- Perovskite solar cells
- Photocatalysts
- Fluorescent polymers
- Surface analysis
- Cancer - drugs/biomarkers
- ‘Atomic’ drugs

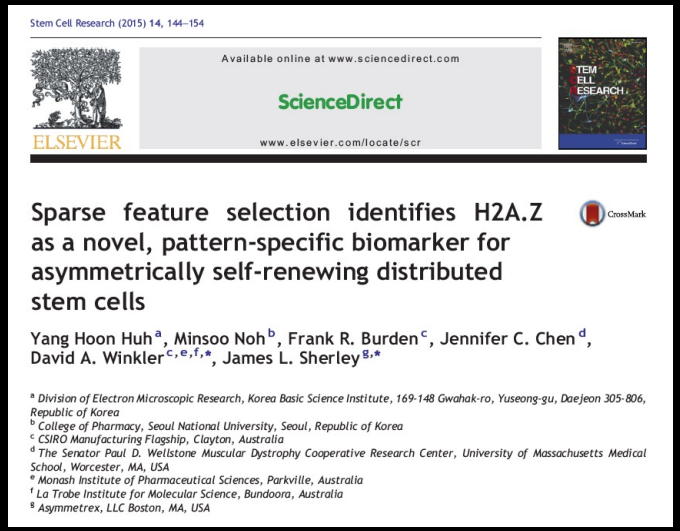
Recent applications of sparse methods

- Stem cell biomarkers (Asymmetrex)
- Sr-induced mesenchymal stem cell differentiation (RepRegen/Stronbone)
- Colorectal cancer biomarkers



Discovering new stem cell markers

- A long-standing challenge in stem cell biomedicine to identify and count tissue stem cells.
- No biological markers specific for adult tissue stem cells.
- With James Sherley, identified biomarkers for symmetry of stem cell division
- Asymmetrex now sells biomarkers that allow monitoring of tissue stem cell number and quality for regenerative medicine



Stem Cell Research (2015) 14, 144–154

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/scr

ELSEVIER

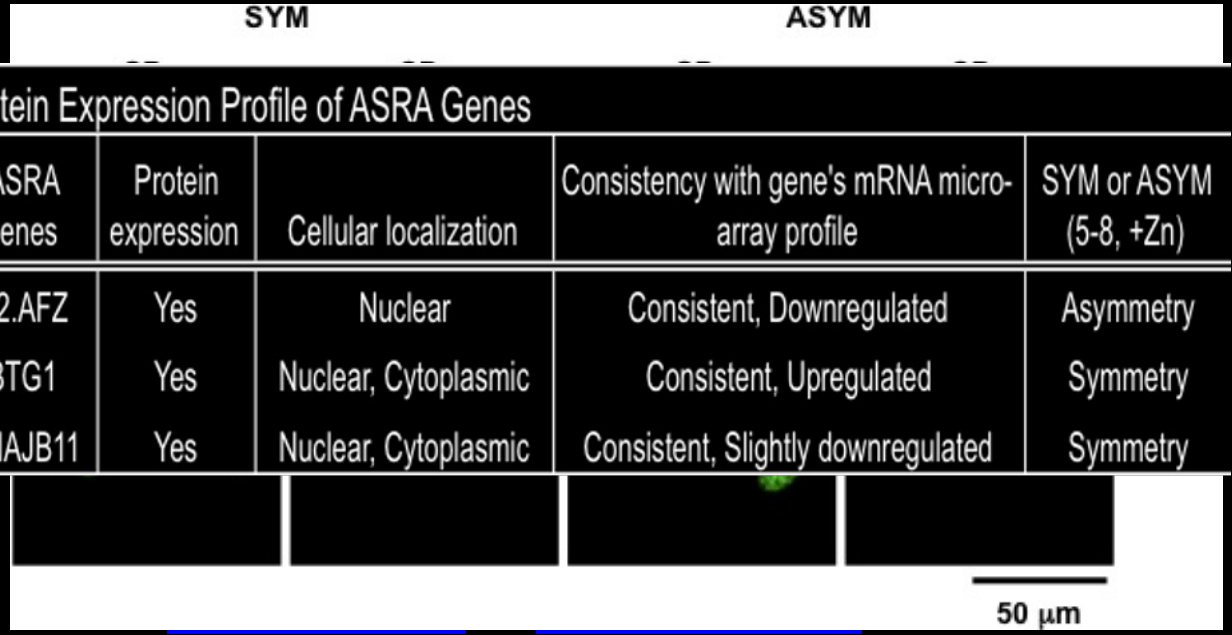
Sparse feature selection identifies H2A.Z as a novel, pattern-specific biomarker for asymmetrically self-renewing distributed stem cells

Yang Hoon Huh[a], Minsoo Noh[b], Frank R. Burden[c], Jennifer C. Chen[d], David A. Winkler[c,e,f,*], James L. Sherley[g,*]

[a] Division of Electron Microscopic Research, Korea Basic Science Institute, 169-148 Gwahak-ro, Yuseong-gu, Daejeon 305-806, Republic of Korea
[b] College of Pharmacy, Seoul National University, Seoul, Republic of Korea
[c] CSIRO Manufacturing Flagship, Clayton, Australia
[d] The Senator Paul D. Wellstone Muscular Dystrophy Cooperative Research Center, University of Massachusetts Medical School, Worcester, MA, USA
[e] Monash Institute of Pharmaceutical Sciences, Parkville, Australia
[f] La Trobe Institute for Molecular Science, Bundoora, Australia
[g] Asymmetrex, LLC Boston, MA, USA

Discovering new stem cell markers

Model Systems for Orthogonal-Intersection Gene Microarrays for Studying Genes Associated with Asymmetric Self-Renewal



Protein Expression Profile of ASRA Genes

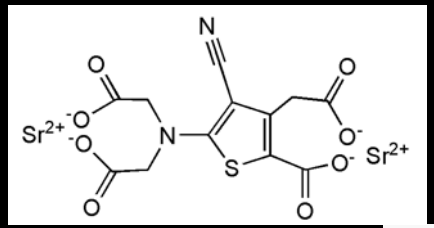
ASRA genes	Protein expression	Cellular localization	Consistency with gene's mRNA micro-array profile	SYM or ASYM (5-8, +Zn)
H2.AFZ	Yes	Nuclear	Consistent, Downregulated	Asymmetry
BTG1	Yes	Nuclear, Cytoplasmic	Consistent, Upregulated	Symmetry
DNAJB11	Yes	Nuclear, Cytoplasmic	Consistent, Slightly downregulated	Symmetry

Recent applications of sparse methods

- Stem cell biomarkers (Asymmetrex)
- Sr-induced mesenchymal stem cell differentiation (RepRegen/Stronbone)
- Colorectal cancer biomarkers

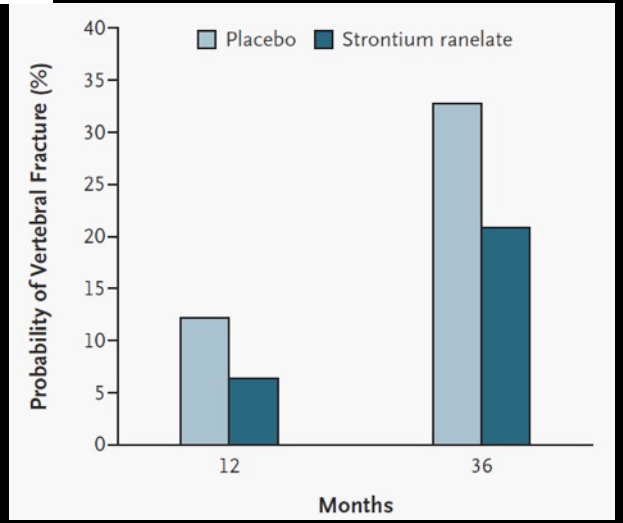
EXIT

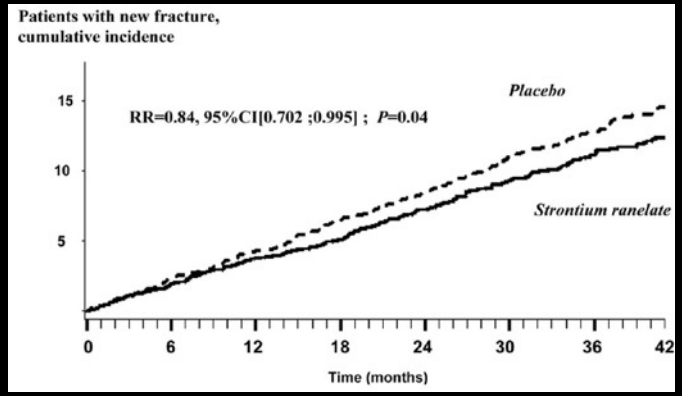
Sr-induced osteogenic differentiation of MSCs



Strontium ranelate (Protelos®) approved in EU for the treatment and prevention of osteoporosis – strontium is the active component. Reduces risk of vertebral and non-vertebral fractures in post-menopausal women. Although controversial, reported to have an anabolic AND anti-catabolic effect on bone

Strontium ion's mechanism of action is not fully understood, but it is thought to up-regulate differentiation of osteoprogenitors or to stimulate bone formation

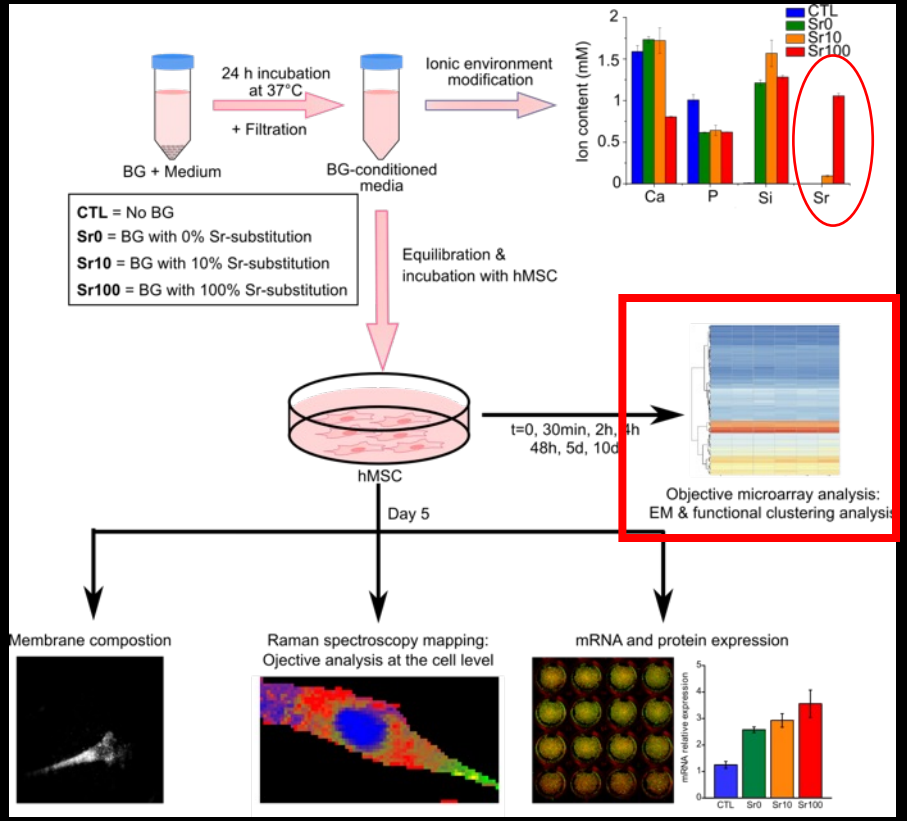




Reginster *et al. J Clin Endocrinol Metab*. 2005; Meunier *et al. NEJM* 2004

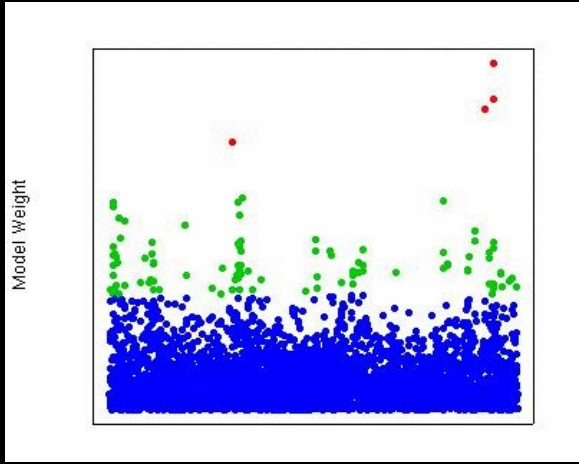
Sr-induced osteogenic differentiation of MSCs

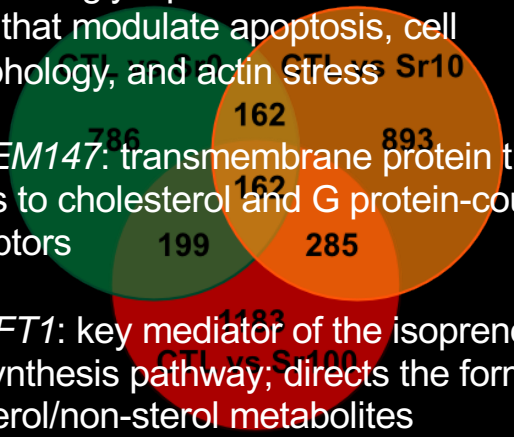
Evaluate the genome-wide response of human mesenchymal stem cells (hMSC) to strontium-substituted bioactive glasses (BG) using a combination of unsupervised biological and physical science techniques



Autefage, Gentleman, Winkler, Burden, Stevens, **PNAS** 2015

Whole genome gene expression analysis





- *PMP22*: glycoprotein associated with lipid rafts that modulate apoptosis, cell morphology, and actin stress
- *TMEM147*: transmembrane protein that binds to cholesterol and G protein-coupled receptors
- *FDFT1*: key mediator of the isoprenoid biosynthesis pathway; directs the formation of sterol/non-sterol metabolites

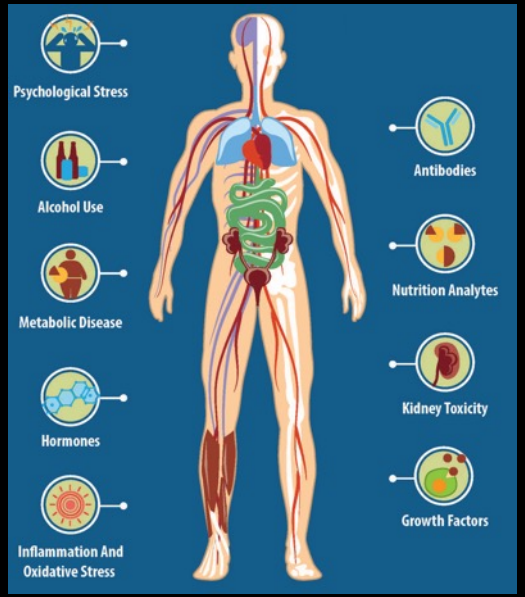
Ex

Gene symbol	Gene name	GeneBank accession n°	Contribution factor	*p*-value
PMP22	peripheral myelin protein 22	NM_000304	2.2+/-0.9	0.01
TMEM147	transmembrane protein 147	NM_032635	2.7+/-1.4	0.04
FDFT1	farnesyl-diphosphate farnesyltransferase 1	NM_004462	0.8+/-0.5	0.09

REPREGEN™
The Smart Biomaterials Company

Recent applications of sparse methods

- Stem cell biomarkers (Asymmetrex)
- Sr-induced mesenchymal stem cell differentiation (RepRegen/Stronbone)
- Colorectal cancer biomarkers



EXIT

Biomarkers for colorectal cancer detection

Metabolomics (2023) 19:84
https://doi.org/10.1007/s11306-023-02049-z

ORIGINAL ARTICLE

Staging of colorectal cancer using lipid biomarkers and machine learning

Sanduru Thamarai Krishnan[1,2,3] · David Winkler[4,5,6] · Darren Creek[1,7] · Dovile Anderson[1,7] · Chandra Kirana[8,9] · Guy J Maddern[8,9] · Kevin Fenix[8,9] · Ehud Hauben[8,9] · David Rudd[1,3] · Nicolas Hans Voelcker[1,3,10]

Received: 22 August 2022 / Accepted: 7 September 2023

Abstract

Introduction Colorectal cancer (CRC) is the third most commonly diagnosed cancer worldwide. Alteration in lipid metabolism and chemokine expression are considered hallmark characteristics of malignant progression and metastasis of CRC. Validated diagnostic and prognostic biomarkers are urgently needed to define molecular heterogeneous CRC clinical stages and subtypes, as liver dominant metastasis has poor survival outcomes.

Objectives The aim of this study was to integrate lipid changes, concentrations of chemokines, such as platelet factor 4 and interleukin 8, and gene marker status measured in plasma samples, with clinical features from patients at different CRC stages or who had progressed to stage-IV colorectal liver metastasis (CLM).

Methods High-resolution liquid chromatography-mass spectrometry (HR-LC-MS) was used to determine the levels of candidate lipid biomarkers in each CRC patient's preoperative plasma samples and combined with chemokine, gene and clinical data. Machine learning models were then trained using known clinical outcomes to select biomarker combinations that best classify CRC stage and group.

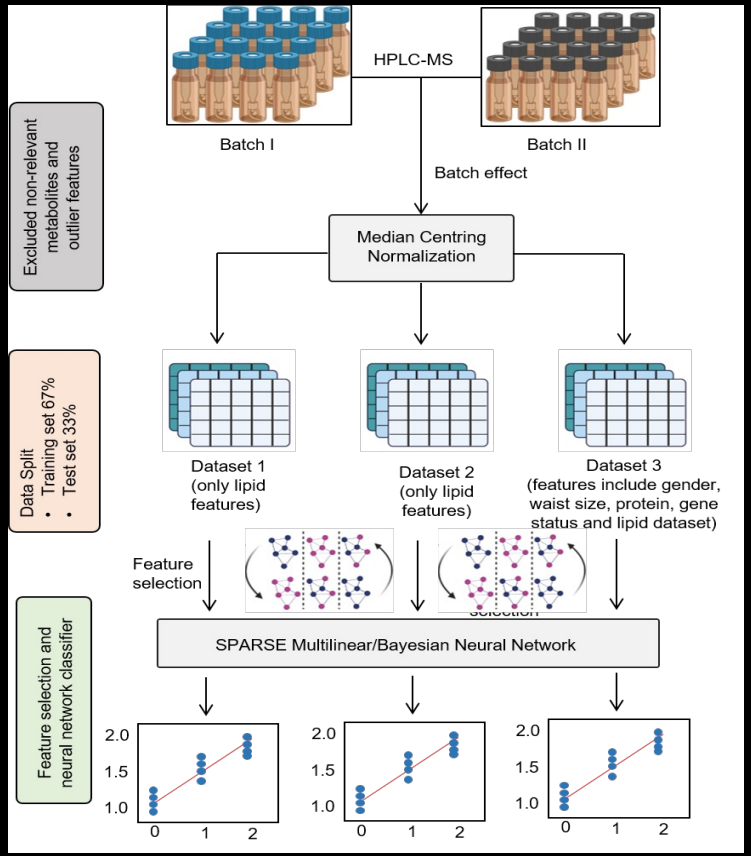
Results Bayesian neural net and multilinear regression-machine learning identified candidate biomarkers that classify CRC (stages I-III), CLM patients and control subjects (cancer-free or patients with polyps/diverticulitis), showing that integrating specific lipid signatures and chemokines (platelet factor-4 and interluken-8; IL-8) can improve prognostic accuracy. Gene marker status could contribute to disease prediction, but requires ubiquitous testing in clinical cohorts.

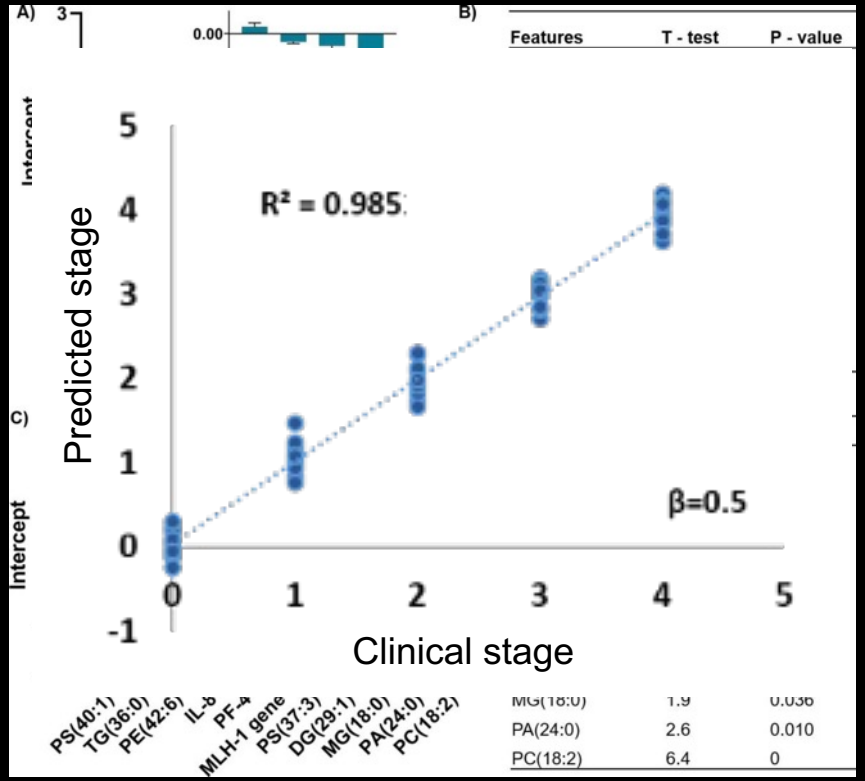
Conclusion Our findings demonstrate that correlating multiple disease related features with lipid changes could improve CRC prognosis. The identified signatures could be used as reference biomarkers to predict CRC prognosis and classify stages, and monitor therapeutic intervention.

Keywords Metastatic colorectal cancer classification · Biomarker · Multi-omics · Machine learning · Cancer Subtypes · Lipidomics

- CRC is the third most common cancer.
- Altered lipid metabolism and chemokine expression during CRC progression and metastasis.
- Diagnostic and prognostic biomarkers urgently needed to define clinical stages and subtypes. Liver dominant metastasis has poor outcomes
- Used sparse modelling to identify set of most relevant lipids for CRC

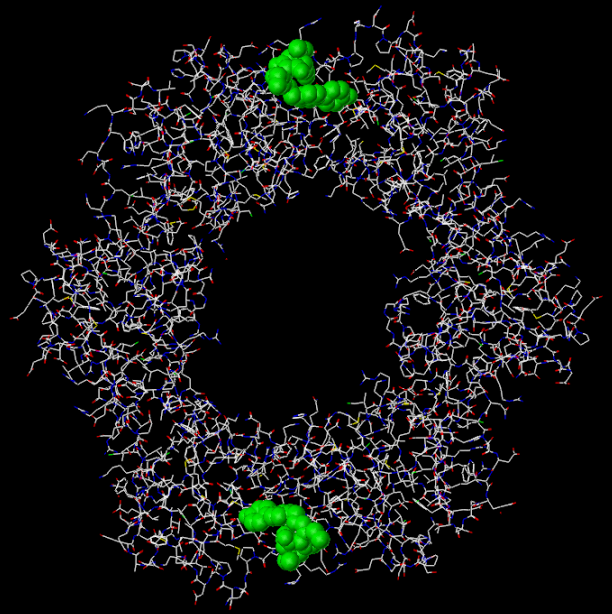
Biomarkers staging colorectal cancers (CRC)





Tripeptide motifs

- Drug design tool
- Novel antibiotics (Betabiotics)
- Lead first-in-class myelofibrosis drug



EXIT

Biologically-relevant tripeptide motifs

SKL peroxisomal targeting

DHP stimulates pituitary gland controlling thyroid-stimulating hormone secretion

R of thrombosis

P otor

G g

D

E otif

D l replis

L

F natory,

G

G n

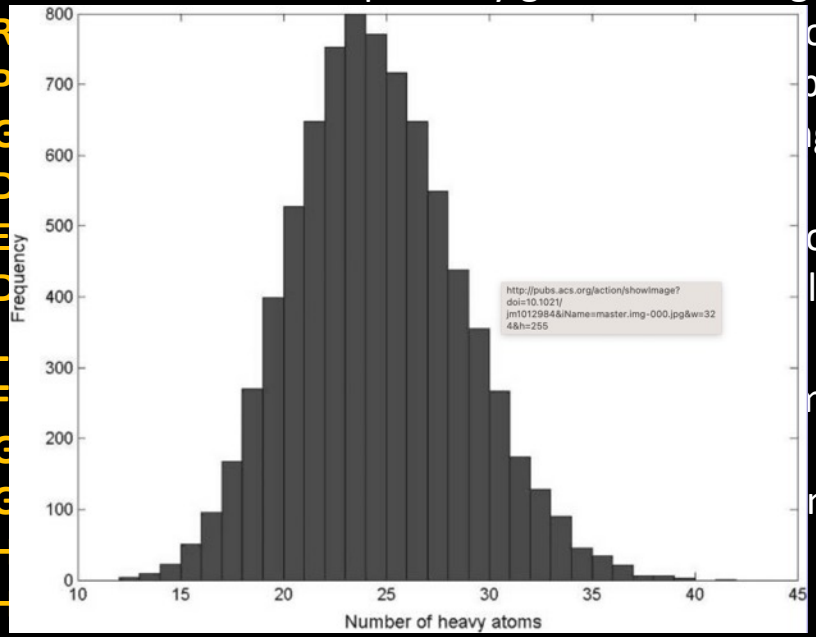
H

H

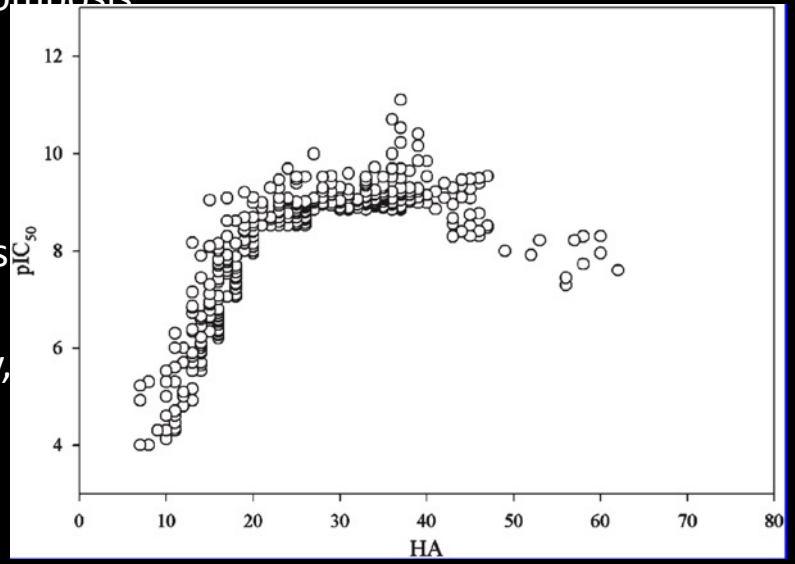
KPV anti-inflammatory properties

HAV cadherin recognition sequence

Distribution of heavy atoms in 8000 tripeptides



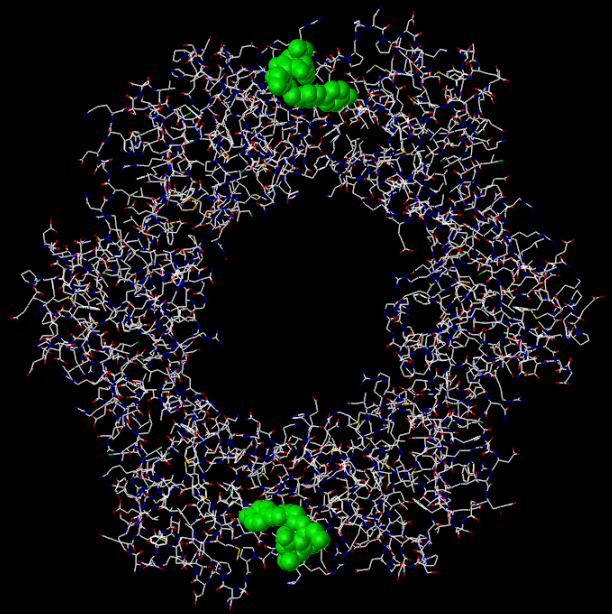
Inhibition versus number of heavy atoms in ligand



Ung & Winkler, J. Med. Chem. 2011, 54, 1111

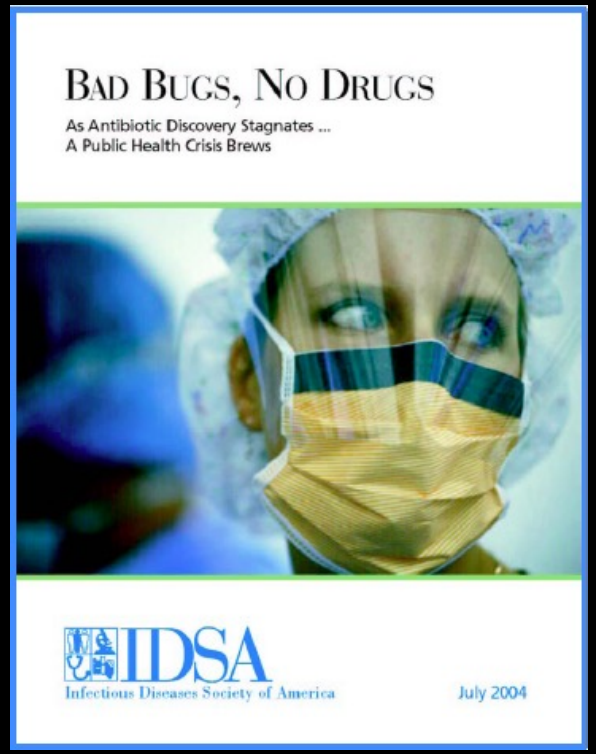
Tripeptide motifs

- Drug design tool
- Novel antibiotics (Betabiotics)
- Lead first-in-class myelofibrosis drug



EXIT

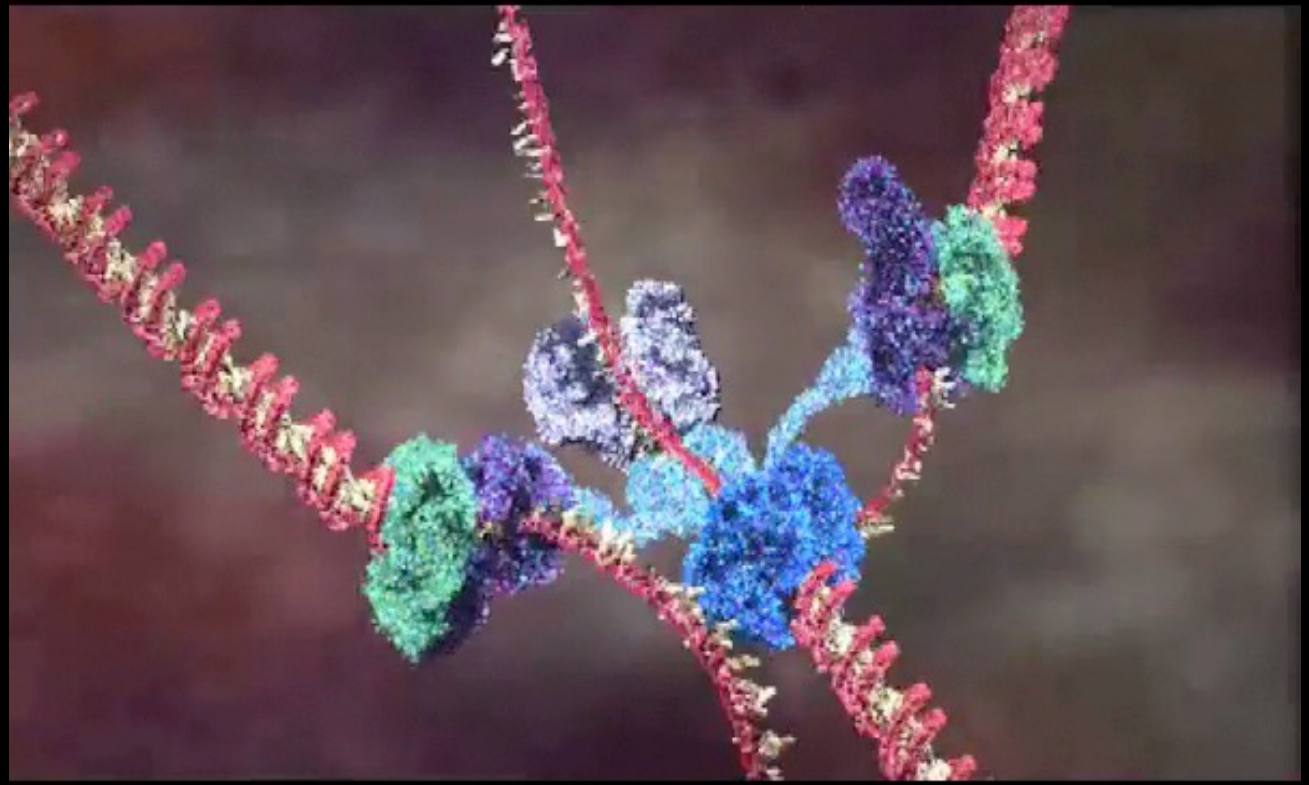
Application of tripeptide motifs: Global antibiotic crisis



- No new antibiotic classes 1970-2000 when Betabiotics began
- Resistance problems - MRSA, VRSA, VRE, TB (5 million deaths in 2019)
- US$43 billion market (2022 dollars)
- Innovation needed – new mode of action antibiotic
- Role for small companies – *even more important now with drug pipelines drying up*

wehi.edu.au

DNA polymerase beta protein



Alignments of pol B & C in eubacteria

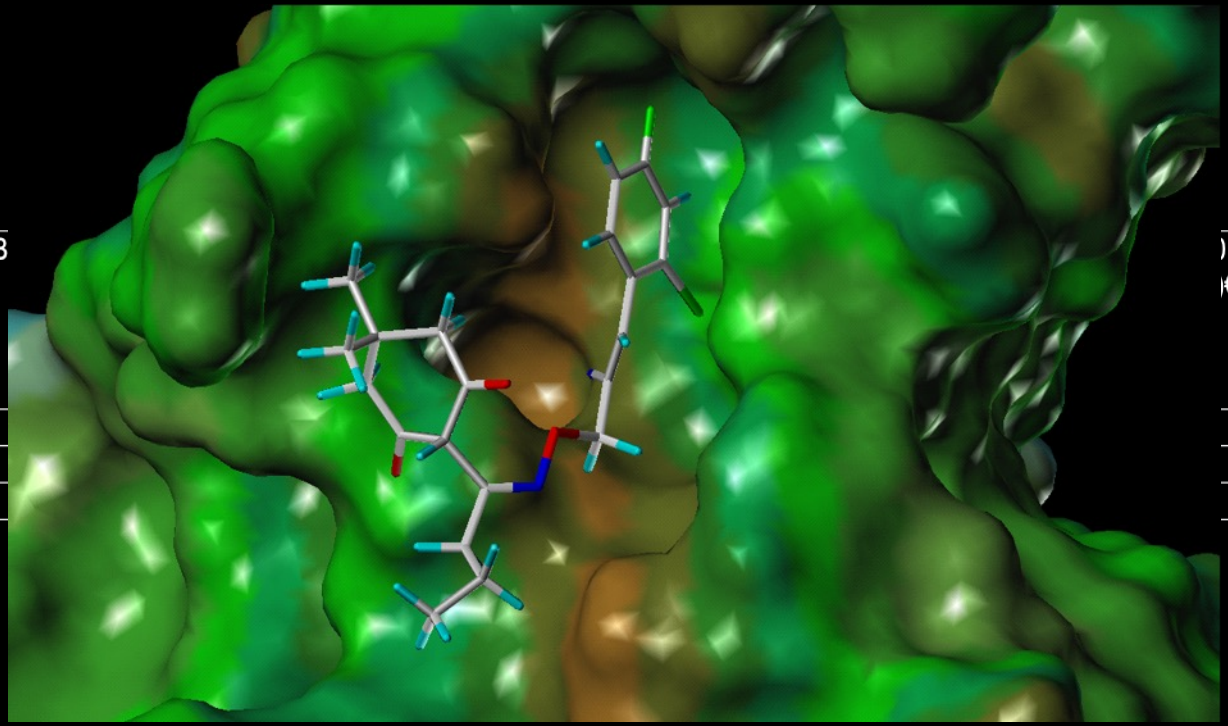
```
PolC - eubacteria
Clostridium difficile          TNHGSLENMSE-RN
Bacillus anthracis             DSQGCLGDLPD-QN
Bacillus subtilis              DRHGCLESLPD-QN
Bacillus stearothermophilus    ESRGCLDSLPD-HN
Staphylococcus aureus          DELGSLPNLPD-KA
Enterococcus faecalis          NENGVLKDLPD-EN
Streptococcus pyogenes         DEMGILGNMPE-DN
Streptococcus pneumoniae       DEMGILGNMPE-DN
Lactococcus lactis             TNMGVLEGMPD-DN
Ureaplasma urealyticum         RVLGVLDHLSE-TE
Mycoplasma pulmonis            KSMGIFEQIPE-TN
Mycoplasma genitalium          EQLQLFDEFEH-QD
Mycoplasma pneumoniae          TQMQLLDEFREQDN
Clostridium acetobutylicum     RKFGCLKGLPE-SD
Thermotoga maritima            KSLGVLGDLPE-TE
PolB - eubacteria
Escherichia coli               IE-DNFATLM--TG
Salmonella typhimurium         VE-DNFATLL--TG
Klebsiella pneumoniae          VN-DDFATIV--TG
Yersinia pestis                TQ-DDFTTLI--TG
Vibrio cholerae                IG-KQFDELI--AP
Pseudomonas aeruginosa         VG-DDFATLV--DR
Pseudomonas putida             VG-DDFARLT--DH
Shewanella putrefaciens        MK-LNYTNIA--SK
```

Wijffels et al., *J. Med. Chem.*, 2011.
Kurz et al., *J. Bacteriol.* 2004.
Wijffels et al., *Biochem.* 2004.
Dalrymple et al., *PNAS* 2001.

D/SLF –beta protein inhibitors design

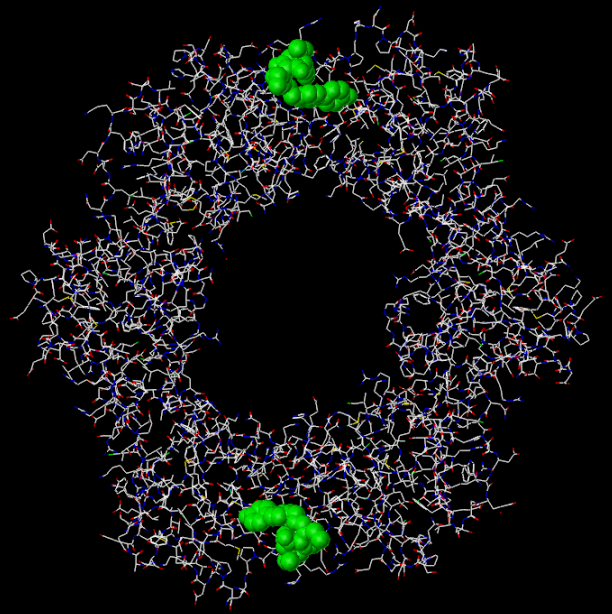
A database search examined conformations of all SLF and DLF motifs in the Protein database.

There was a surprising degree of conservation of 3D structure, providing a pharmacophore for virtual screening of small molecule libraries.



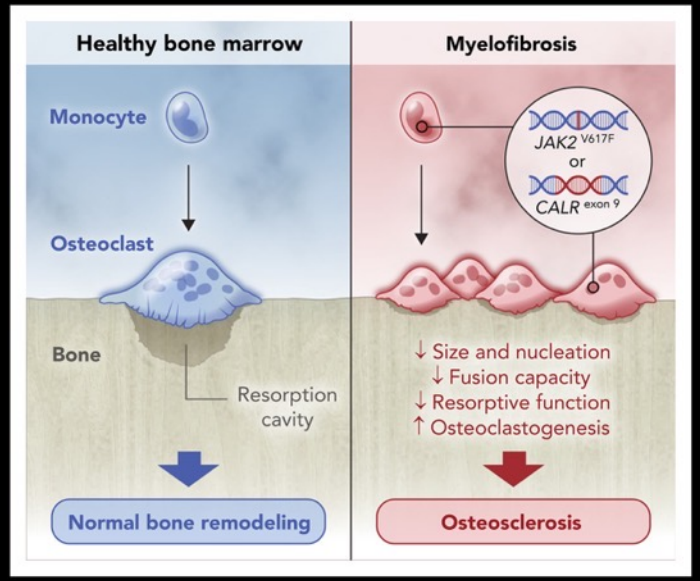
Tripeptide motifs

- Drug design tool
- Novel antibiotics (Betabiotics)
- Lead first-in-class myelofibrosis drug



EXIT

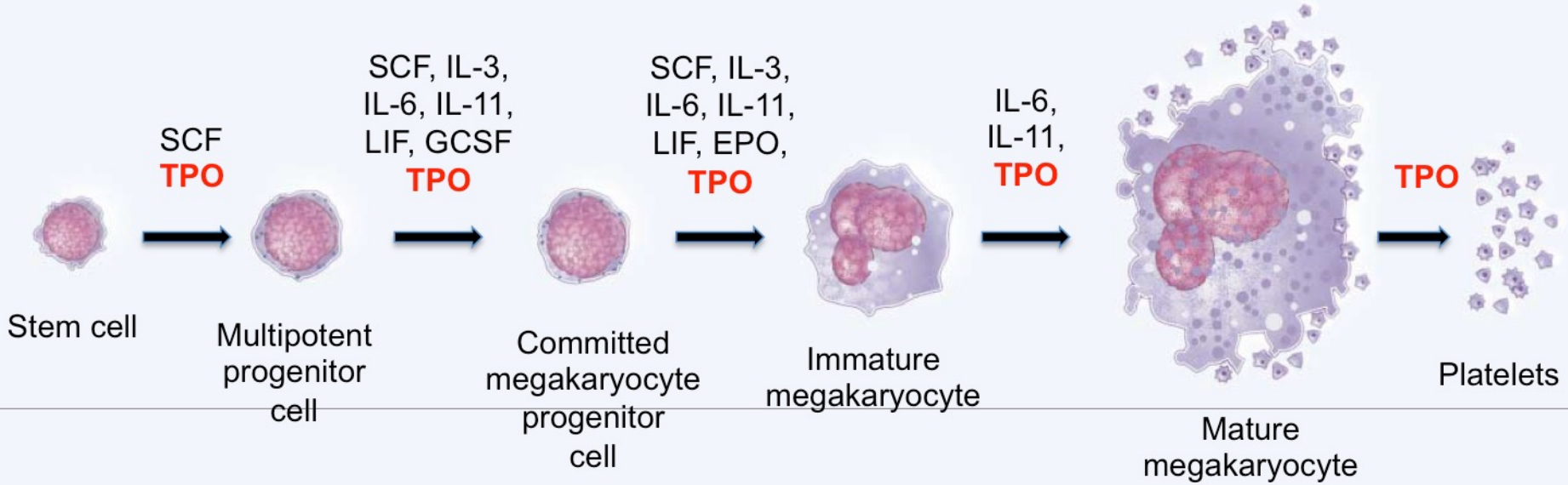
Myelofibrosis, an incurable blood cancer



- Myelofibrosis is an incurable blood cancer with no disease modifying treatment
- Doctors treating myelofibrosis have hit a ‘clinical brick wall’ due to lack of effective drug treatments
- Myelofibrosis cuts short lives of patients and causes very painful symptoms
- We developed the first drug lead that can change of course of the disease by exploiting a novel target in blood stem cells
- The only registered drug for myelofibrosis earns US$1.2Bn pa in sales, but it is only palliative (does not modify course of disease).

Role of thrombopoietin (TPO) in haematopoiesis

http://www4.utsouthwestern.edu/huanglab/research.jpg

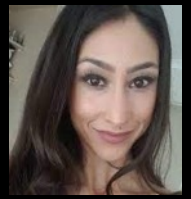


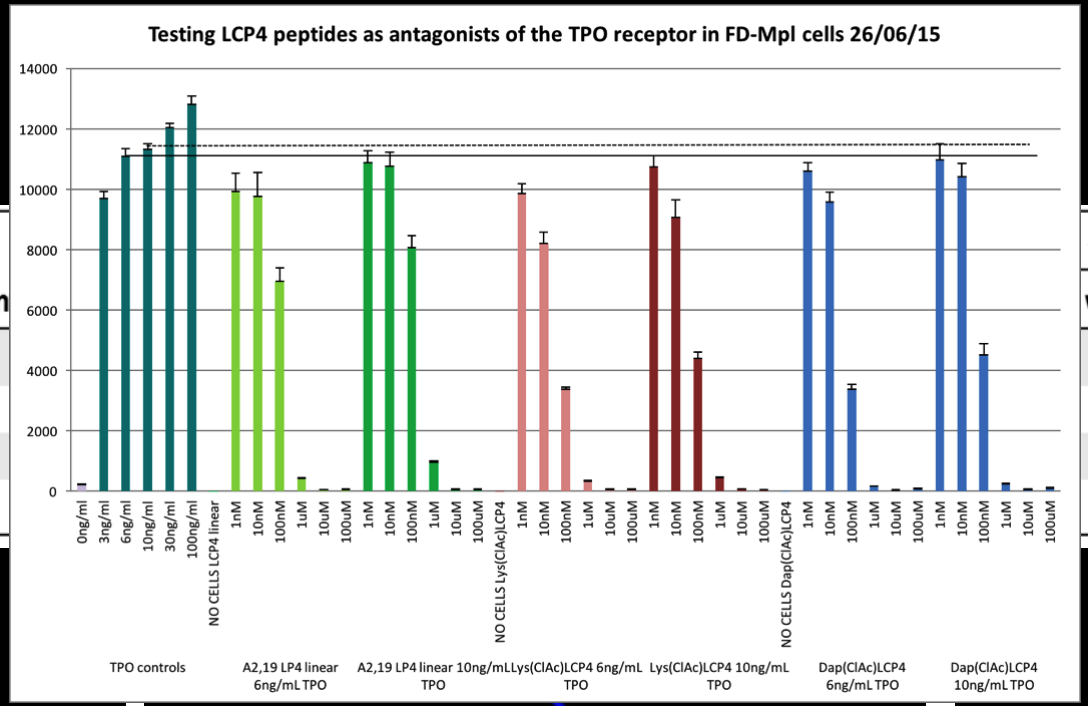
Myelofibrosis: tripeptide motif from phage display

Table 1. Family 2 binding peptides from phage display that identify the minimal motif specific to c-Mpl.[27]

Peptide	Sequence	IC_{50} [nM]
AF12192	GCTL**REW**LHGGFCGG	200
AF12193	GGCADGPTL**REW**ISFCGG	60
AF12359	GNADGPTL**RQW**LEGRRPKN	60
AF12434	LAIEGPTL**RQW**LHGNGRDT	20
AF12405	TIKGPTL**RQW**LKSREHTS	50
AF12505	IEGPTL**RQW**LAARA	2
AF13948	IEGPTL**RQW**LAARA(βAla)K | ARAALWQRLTPGEI	0.5

First small molecule TPO antagonist





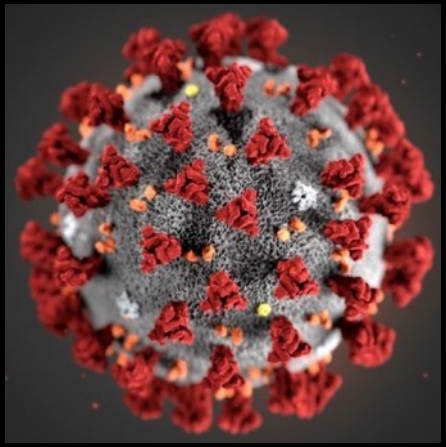
Selectively ablates MF HSCs

Patien	(%) with cytokines + LCP4 (100 nM)	% Reduction*
SP10	47.1	38.8
SP19	65.9	14.1
SP21	7.3	72.1
SP22	38.1	25.7

Adv. Therapeut. 2021
Blood 127 (2016)
ChemMedChem. 2013
Exp. Hematol. 2013
Mol. BioSyst. 2012
Cytokine Growth Fact. Rev. 2011
ChemMedChem. 2009.
11 patents

Recent molecular design research: –

- COVID drugs and origin of virus
- Antifibrotics/antihypertensives (Vectus)



EXIT

Lab Leak: A Scientific Debate Mired in Politics — and Unresolved

More than a year into the SARS-CoV-2 pandemic, some scientists say the possibility of a lab leak never got a fair look.

BY CHARLES SCHMIDT (HTTPS://UNDA AUTHOR/CHAR SCHMIDT/)

03.17.2021

Top: Security personnel stand guard outside the Wuhan Institute of Virology in Wuhan in February, as members of the World Health Organization (WHO) team investigating the origins of the Covid-19 coronavirus pay a visit. Visual: Hector Retamal / AFP via Getty Images

NIKOLAI PETROVSKY was scrolling through social media after a day on the ski slopes when reports describing a mysterious cluster of pneumonia cases

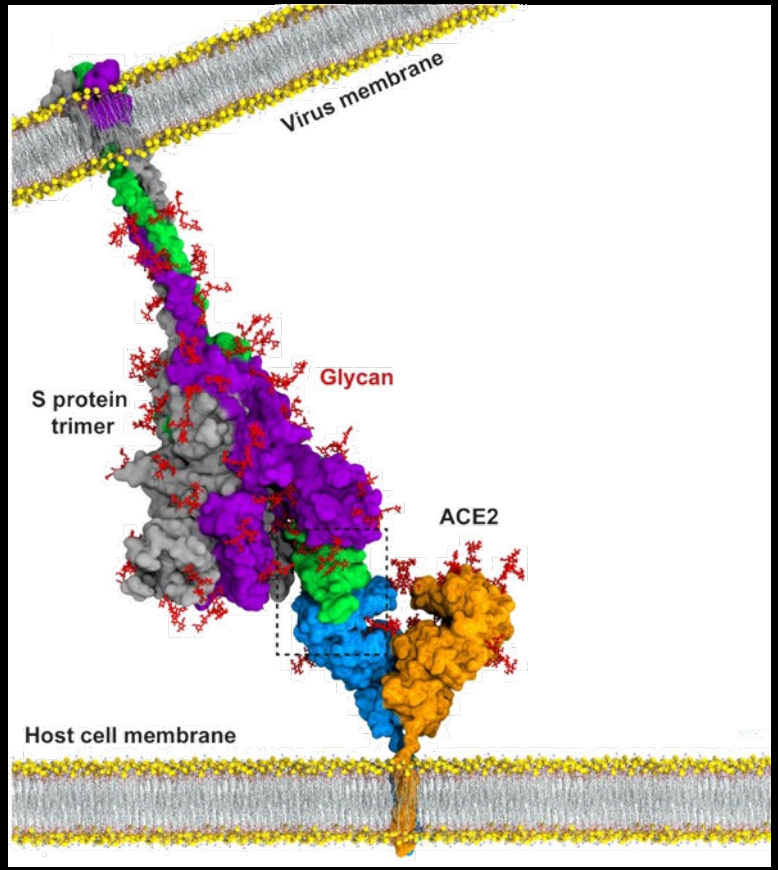
vaxine

LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

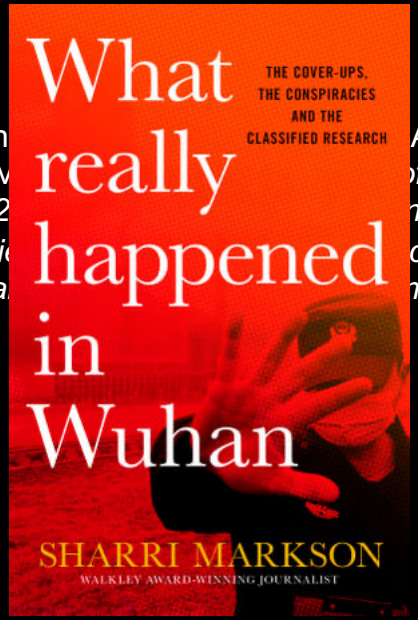
LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

vaxine

Infection-initiating event – spike interaction with ACE2



Piplan A.;
Petrov of SARS-
CoV-2 nding
affiniti ce for
anima n, **npj Sci Rep**.,

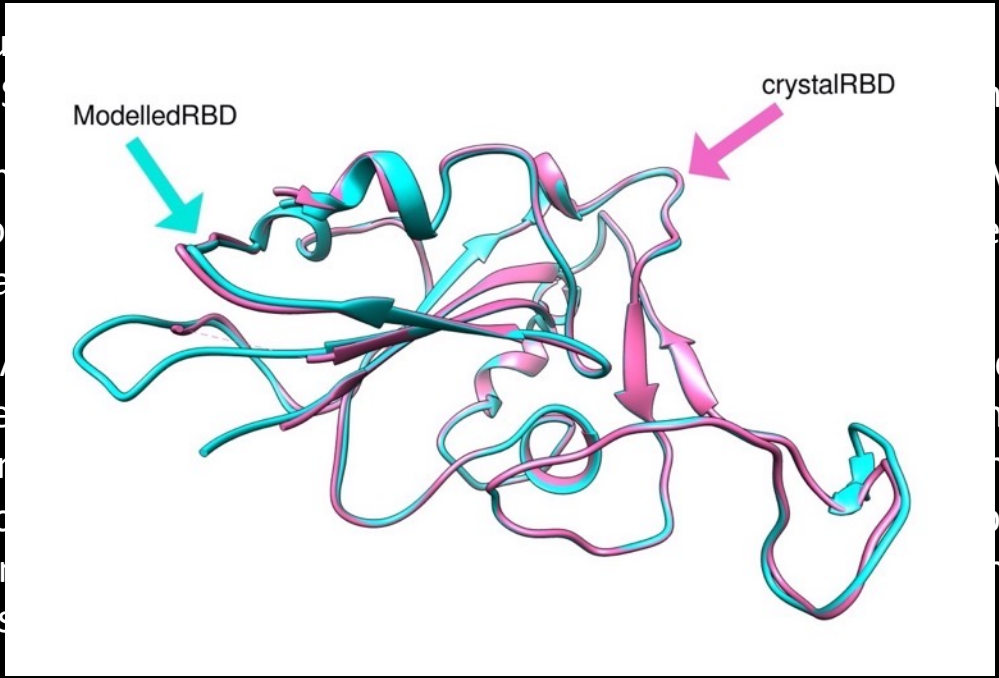


vaxine

LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

SARS-CoV-2 spike and ACE2 models

- The modelled structu Plot and molprobity scores in S e RBD
- Structure of the open VYB). Very high structural similarity o EM structures (PDB ID 6M0J (RBD) a
- Homology modelled A cture using hybrid HDOCK method. Pote human species because of use of hur nate possible bias, generated ACE2 struc endently by Modeller. Cα backbor 0.5-0.8 Å, showing very strong s



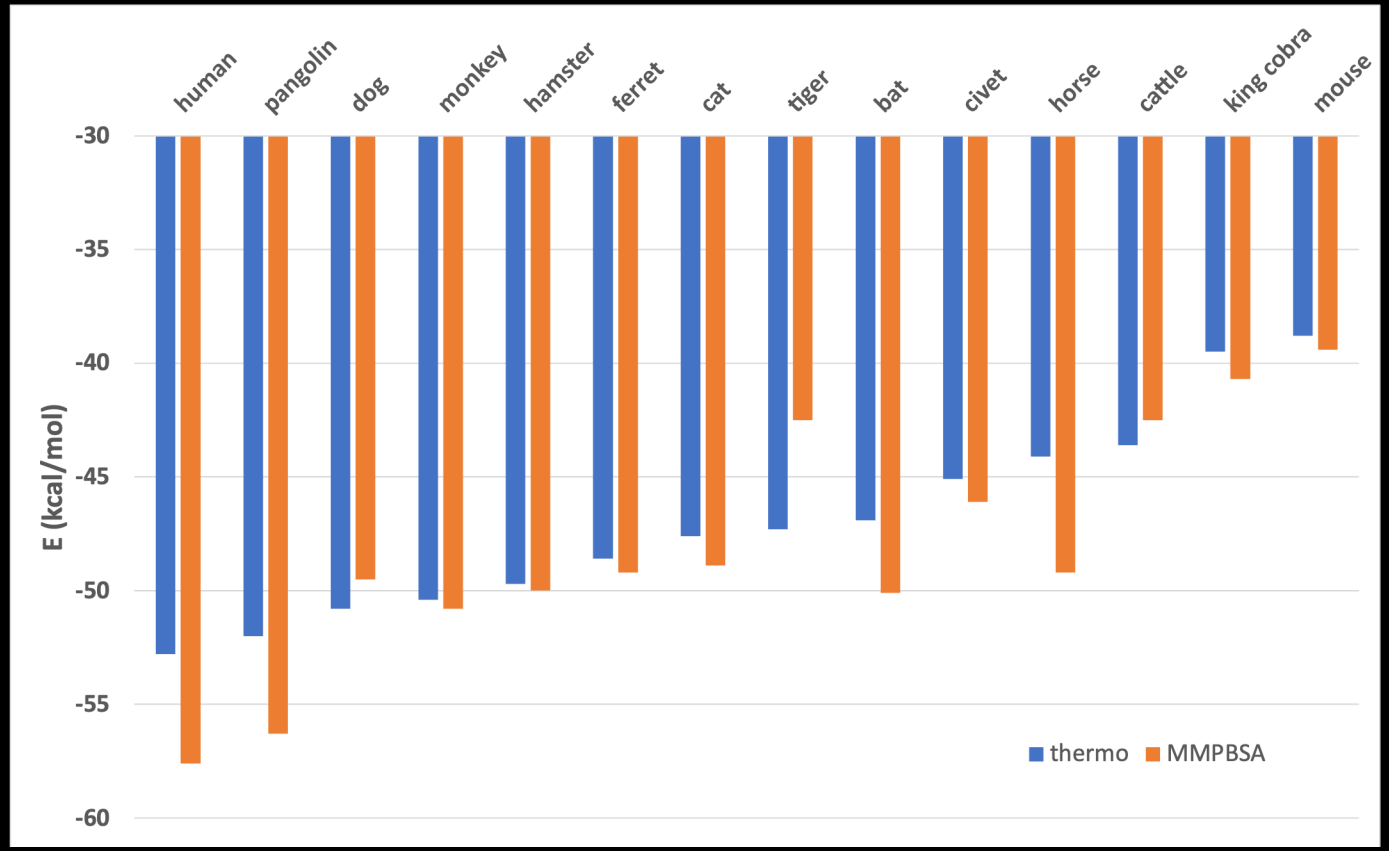
Spike–ACE2 binding free energies and observed infectivities

Species	ΔG_{eqn1} (kcal/mol)	ΔG_{MMPBSA} (kcal/mol)	SARS-Cov-2 infectivity
Homo sapiens (human)	-52.8	-57.6 ± 0.25	Permissive, high infectivity, severe disease in 5-10%,
Manis javanica (pangolin)	-52.0	-56.3 ± 0.4	Permissive [23,24]
Canis luparis (dog)	-50.8	-49.5	Permissive, low/mod infectivity, no overt disease [25,26]
Macaca fascicularis (monkey)	-50.4	-50.8	Permissive, high infectivity, lung disease [11]
Mesocricetus auratus (hamster)	-49.7	-50.0	Permissive, high infectivity, lung disease [27,28]
Mustela putorius furo (ferret)	-48.6	-49.2	Permissive, moderate infectivity, no overt disease [28-30]
Felis catus (cat)	-47.6	-48.9	Permissive, high infectivity, lung disease [26,29,31]
Panthera tigris (tiger)	-47.3	-42.5	Permissive, overt disease, RNA positive[26]
Rhinolophus sinicus (bat)	-46.9	-50.1 ± 1.0	Not permissive [11]
Paguma larvata (civet)	-45.1	-46.1	No reported infection
Equus ferus caballus (horse)	-44.1	-49.2	No naturally occurring infections[26]
Bos taurus (cattle)	-43.6	-42.5	No naturally occurring infections[26]
Ophiophagus hannah (king cobra)	-39.5	-40.7 ± 1.2	No reported infection
Mus musculus (mouse)	-38.8	-39.4	Resistant to infection[28]

vaxine

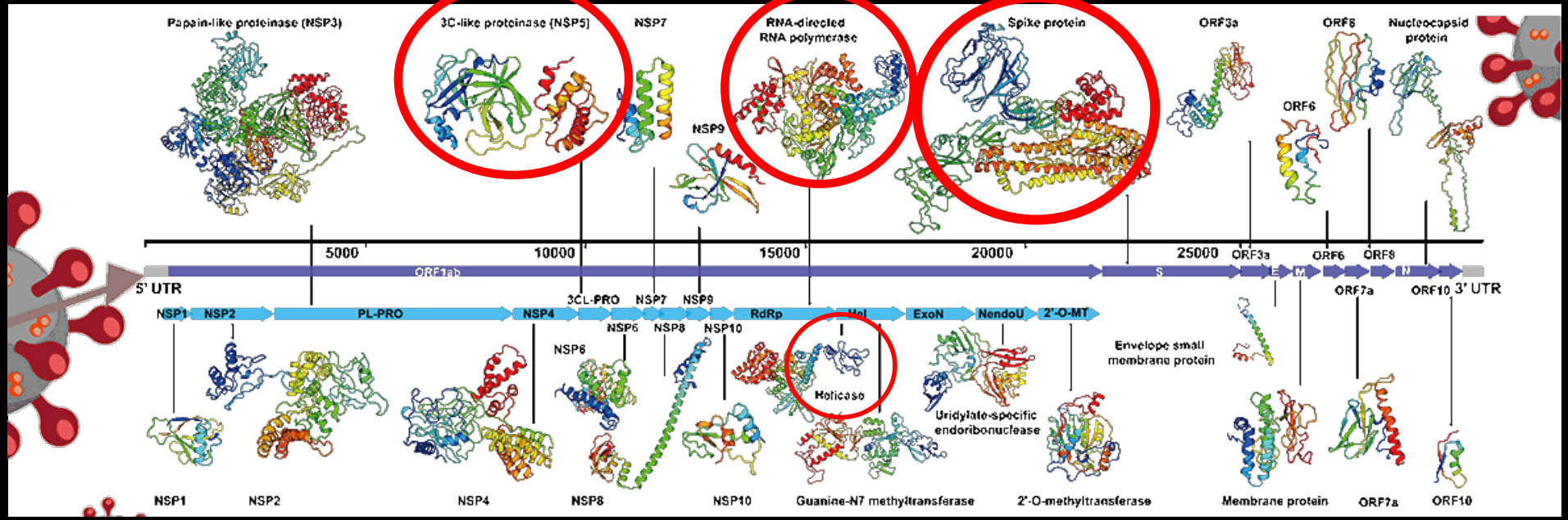
LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

Spike–ACE2 binding free energies and observed infectivities



LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

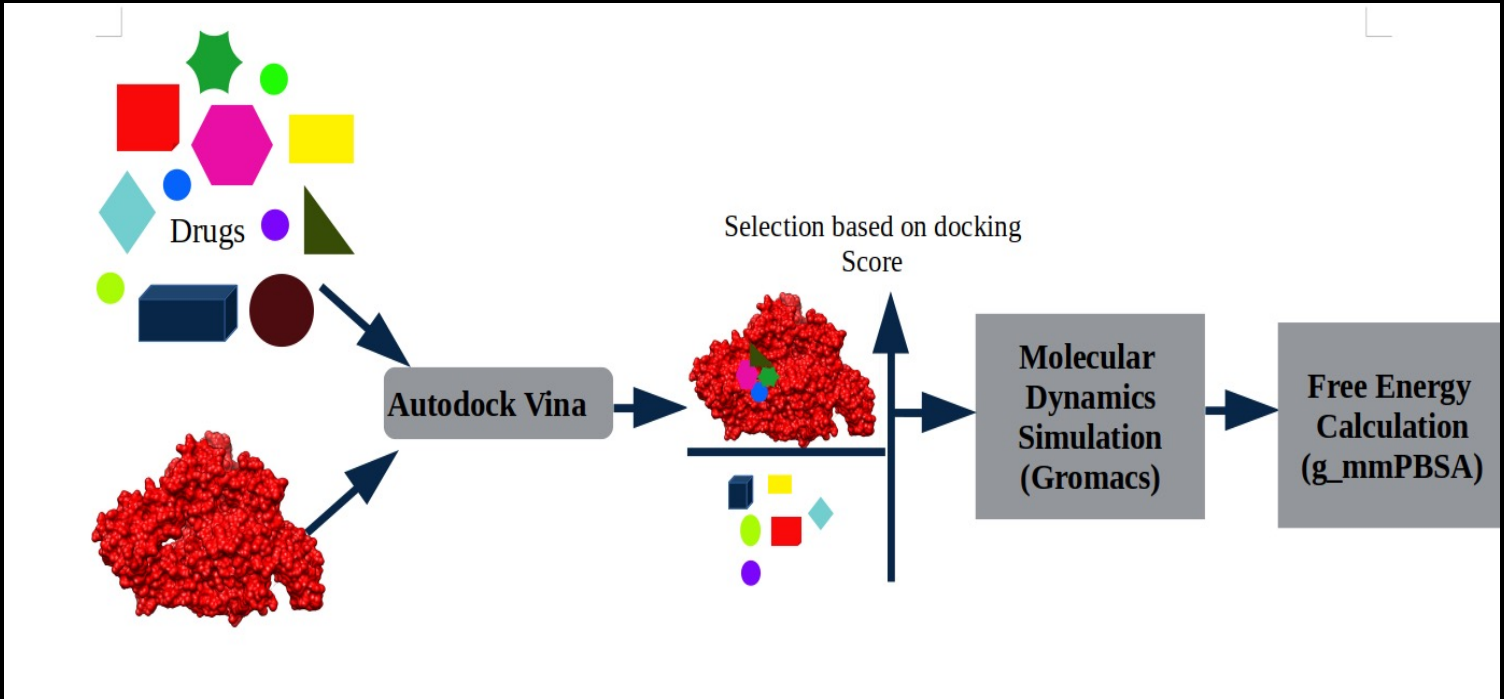
Drug repurposing targets – SARS-CoV-2 proteins (nsp)



Zhang Lab, UMich

LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

Computational screening of SARS-CoV-2 target proteins



Pilani et al., *Rational repurposing of drugs, clinical trials candidates, and natural products for SARS-Cov-2 therapy,* in Frontiers of COVID-19: Scientific and Clinical Perspectives of the Novel SARS-CoV-2, Adibi, Rajabifard, Islam, Ahmadvand (eds.), Springer Nature 2021.

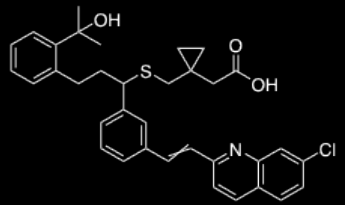
LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

Experimental validation of top 10 repurposed drugs for Mpro

Bemcentinib

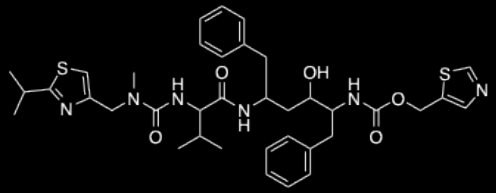
Phase 2 clinical trial, ED_{50} 0.1 (Huh7.5), 0.47 (Vero), 2.1 (Calu3) μM, predicted 2'-O-methyltransferase nsp16/nsp10 complex binding

Montelukast



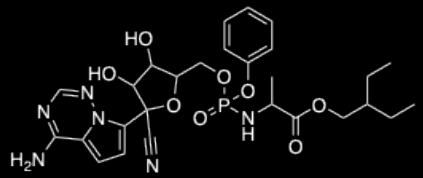
Significant reduction in SARS-CoV-2 infection in elderly asthmatic patients treated with MK. Several predicted M^{pro} binding studies

Ritonavir



In vitro EC_{50} 5.73 μM, Multiple single agent and combination human trials e.g.[27,76]. In vitro EC_{50} 26.63 μM.[77] Predicted M^{pro} binding

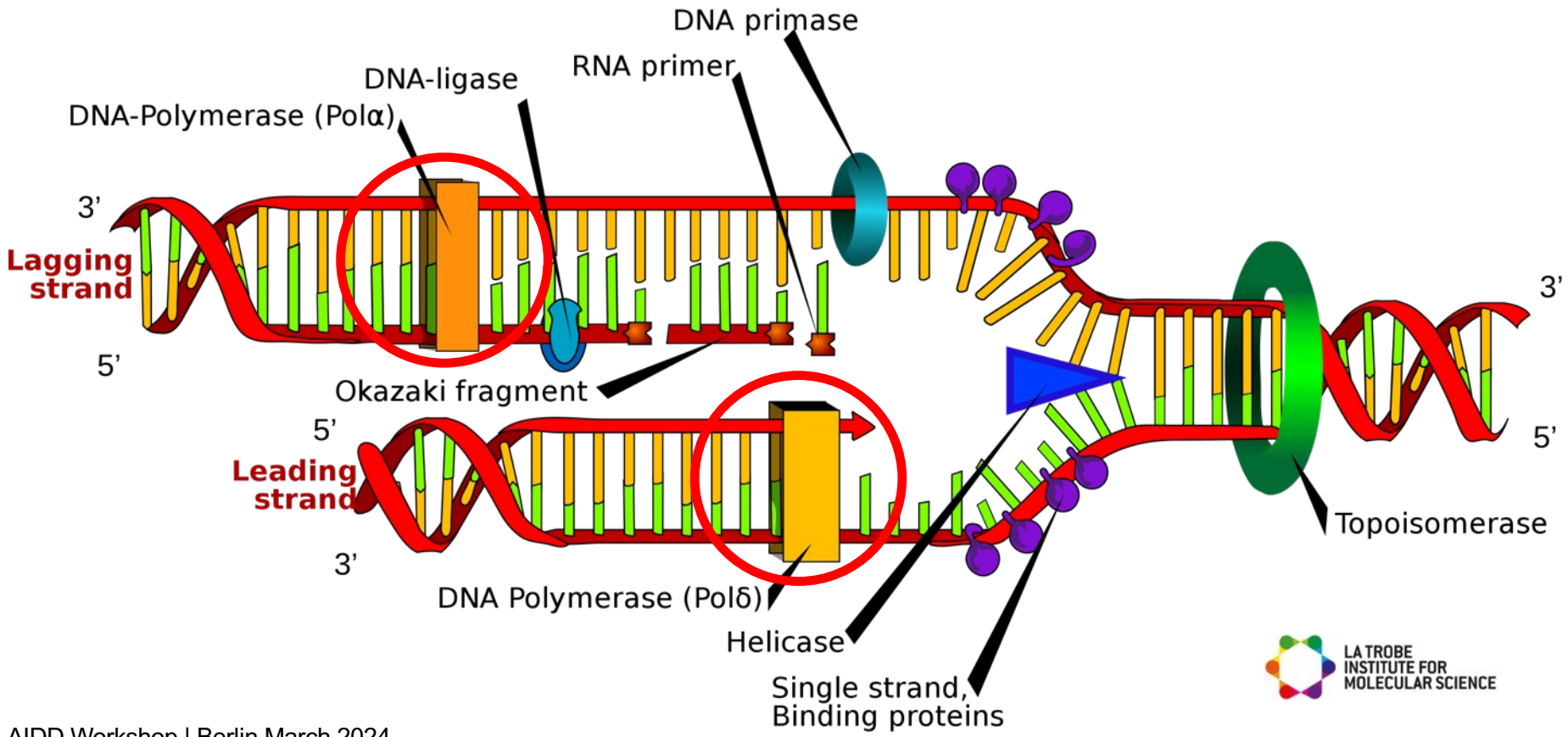
Remdesivir



Multiple human trials, in vitro EC_{50} 23.15 μM, predicted M^{pro} and RdRp binding

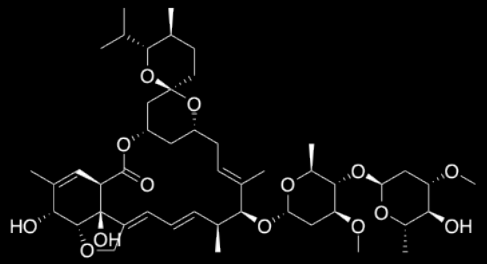
15% of 87 top predicted repurposing hits have experimental validation as of Sept 2020

RNA/DNA polymerases



Experimental validation of top 10 hit repurposed drugs

Ivermectin

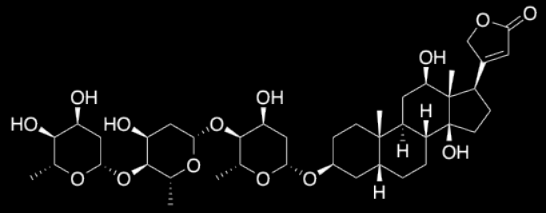


IC_{50} of 2.2 - 2.8 μM in monkey kidney cells.

Bemcentinib

10-40% protection at 50μM in Vero cells. IC_{50} 100nM and CC_{50} 4.7μM in human Huh7.5 cells and IC_{50} of 470nM and CC_{50} 0f 1.6μM in Vero cells, investigational treatment for COVID-19 (www.clinicaltrialsregister.eu, predicted to bind to Mpro.2

Digoxin



Predicted RdRP inhibitor,15 IC_{50} = 0.043 μM and CC_50 >10μM in Vero cells

Galidesvir

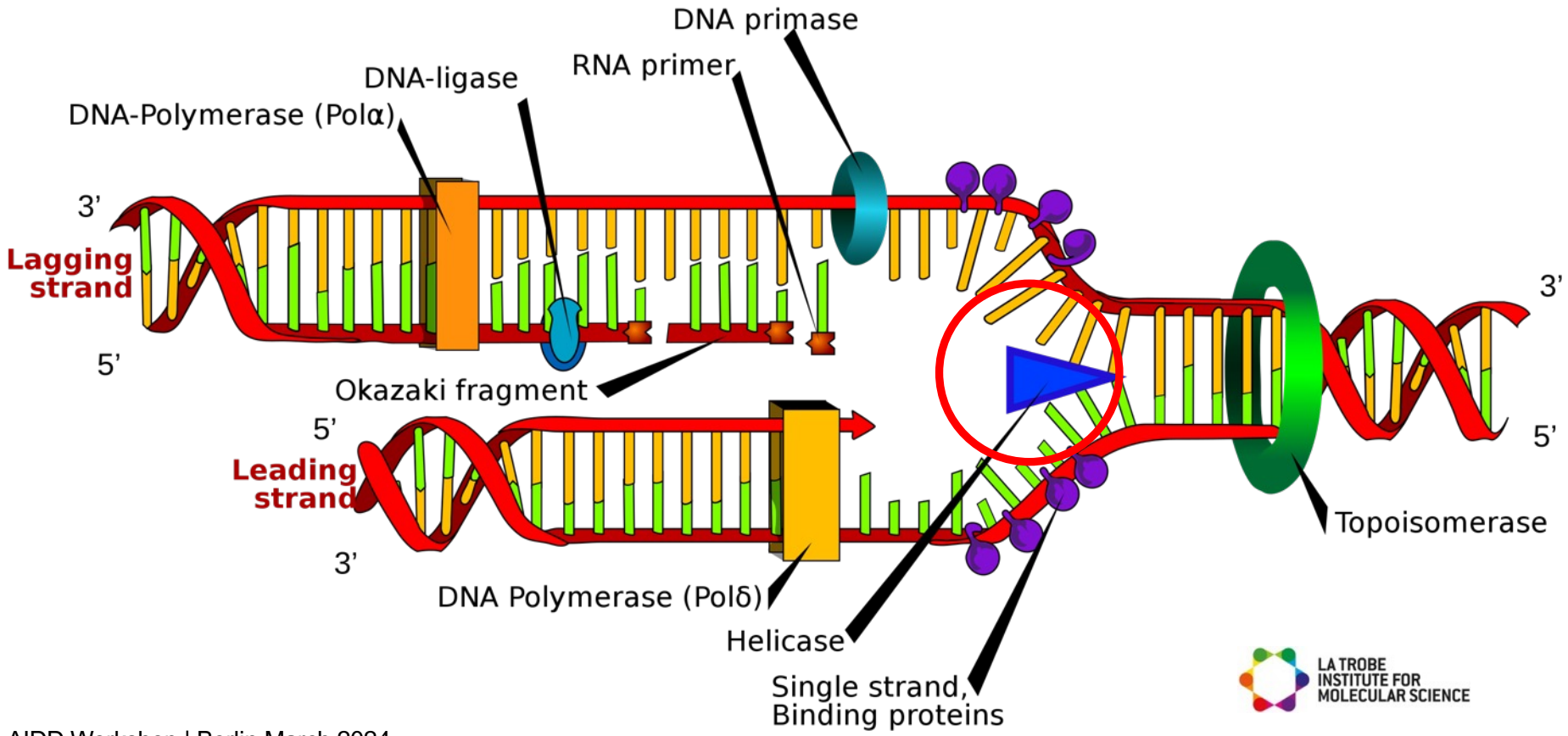
Clinical trials for COVID-19 and RdRP inhibitor

>30% of 80 top predicted repurposing hits have experimental validation as of Jan 2021

vaxine

LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

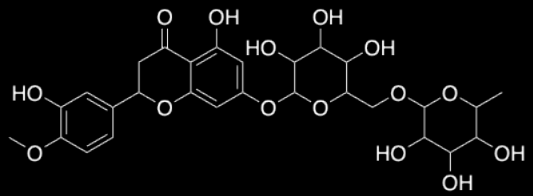
RNA/DNA helicases



LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

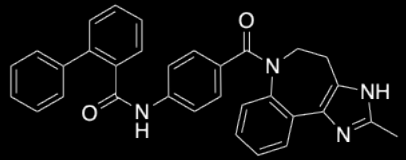
Experimental validation of top 10 hit repurposed drugs

Hesperidin (citrus flavanone glycoside)



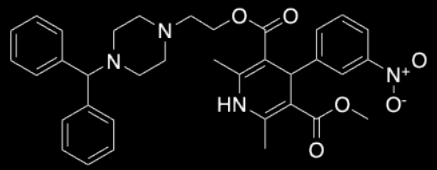
SARS-Cov-2 Mpro inhibition IC_{50} = 8.3 μM

Conivaptan (vasopressin inhibitor)



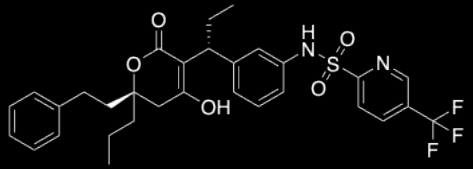
SARS-CoV-2 EC_{50} 10μM CC_{50} 13μM in HEK-293T cells.3 EC_{50} = 12.2 μM against HCoV-OC43

Manidipine (Ca channel blocker, anti-hypertensive)



IC_{50} 10μM against SARS-CoV-2 Mpro ; 14μM against PLpro. Apparent SARS-CoV-2 EC_{50} = 15±1 μM in plaque reduction assay. Kinetic Mpro IC_{50} = 4.8 μM. SARS-CoV-2 activity in in HUH7 cells (IC_{50}=2μM) and Vero cells (IC_{50}=7.5μM).

Tipranavir (antiviral protease inhibitor)



Inhibits replication of SARS-CoV-2 in VeroE6 cells, but low SI (EC_{50} = 13 μM, CC_{50} = 77 μM, SI = 6).

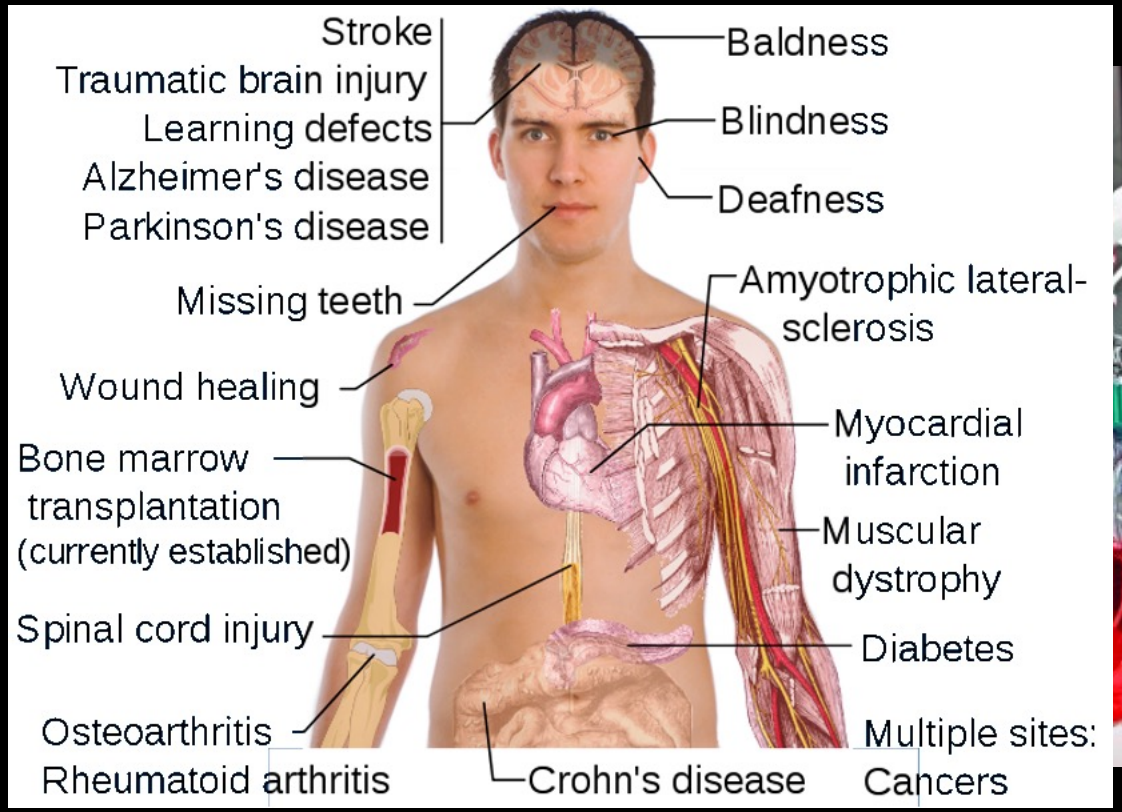
~30% of 87 top predicted repurposing hits have experimental validation at Mar 2021

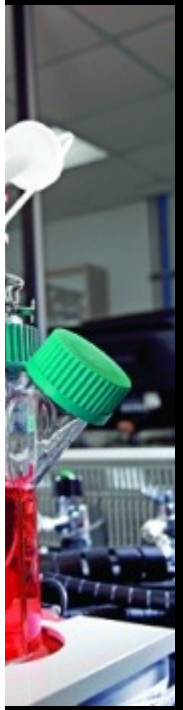
New applications: –

- stem cell bioreactors
- biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology

EXIT

Polymers to grow stem cells

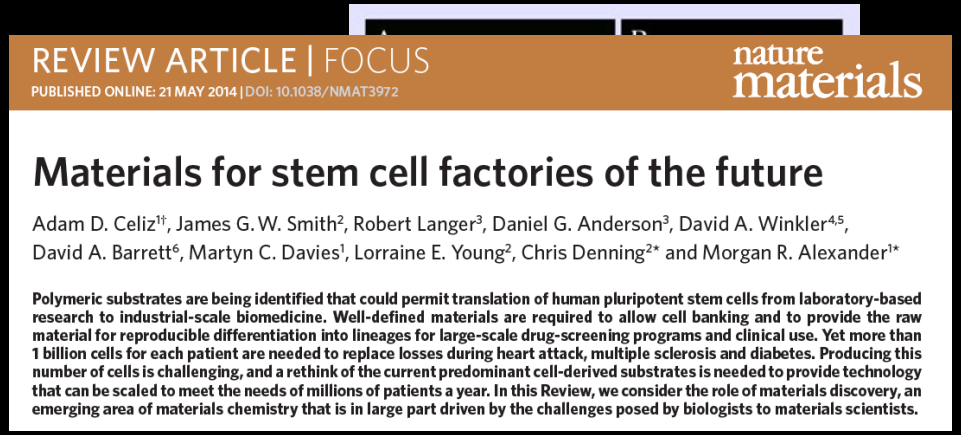




The University of Nottingham

Polymers to grow stem cells

- Present culture methods rely upon **animal-derived** products now under scrutiny.
- Future cell factories will need **chemically defined, serum-free, feeder-free synthetic** substrates and media to support robust self-renewal of pluripotent cells.
- Need new synthetic materials to control morphology, motility, gene expression and differentiation of stem and progenitor cells.
- Important surface properties that have been identified include: -
  - surface chemistry
  - surface
  - wettability
  - topography
  - elastic modulus



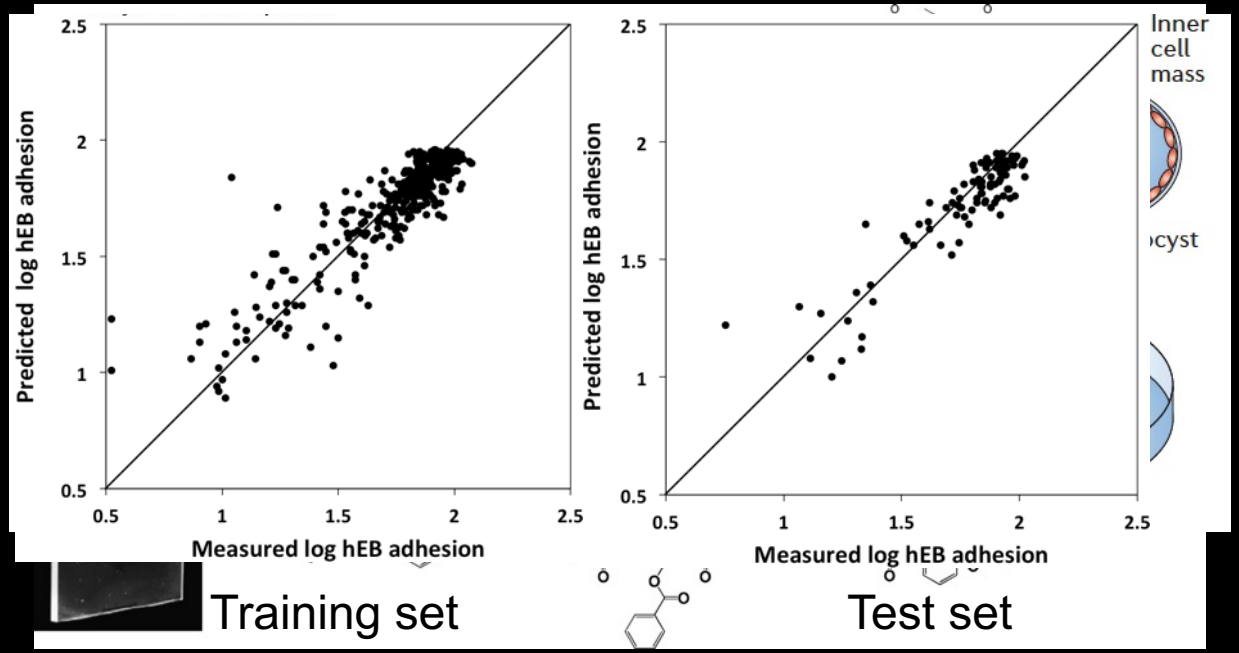
REVIEW ARTICLE | FOCUS
PUBLISHED ONLINE: 21 MAY 2014 | DOI: 10.1038/NMAT3972
nature materials

Materials for stem cell factories of the future

Adam D. Celiz[1†], James G. W. Smith[2], Robert Langer[3], Daniel G. Anderson[3], David A. Winkler[4,5], David A. Barrett[6], Martyn C. Davies[1], Lorraine E. Young[2], Chris Denning[2*] and Morgan R. Alexander[1*]

Polymeric substrates are being identified that could permit translation of human pluripotent stem cells from laboratory-based research to industrial-scale biomedicine. Well-defined materials are required to allow cell banking and to provide the raw material for reproducible differentiation into lineages for large-scale drug-screening programs and clinical use. Yet more than 1 billion cells for each patient are needed to replace losses during heart attack, multiple sclerosis and diabetes. Producing this number of cells is challenging, and a rethink of the current predominant cell-derived substrates is needed to provide technology that can be scaled to meet the needs of millions of patients a year. In this Review, we consider the role of materials discovery, an emerging area of materials chemistry that is in large part driven by the challenges posed by biologists to materials scientists.

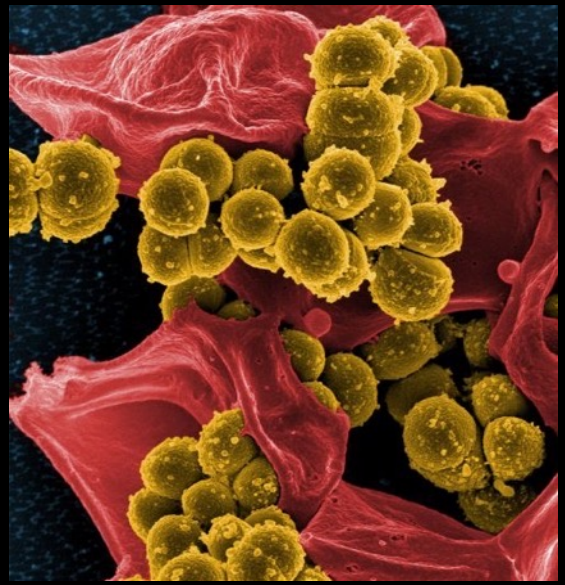
Polymers to grow stem cells



Epa, et al. *J Mat. Chem*. 2012; **22**: 20902-20906. RSC hot paper

New applications: –

- stem cell bioreactors
- next generation biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology

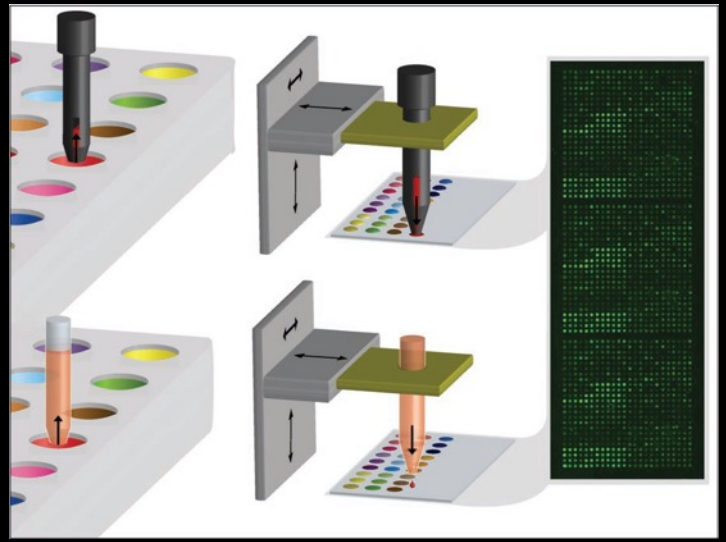


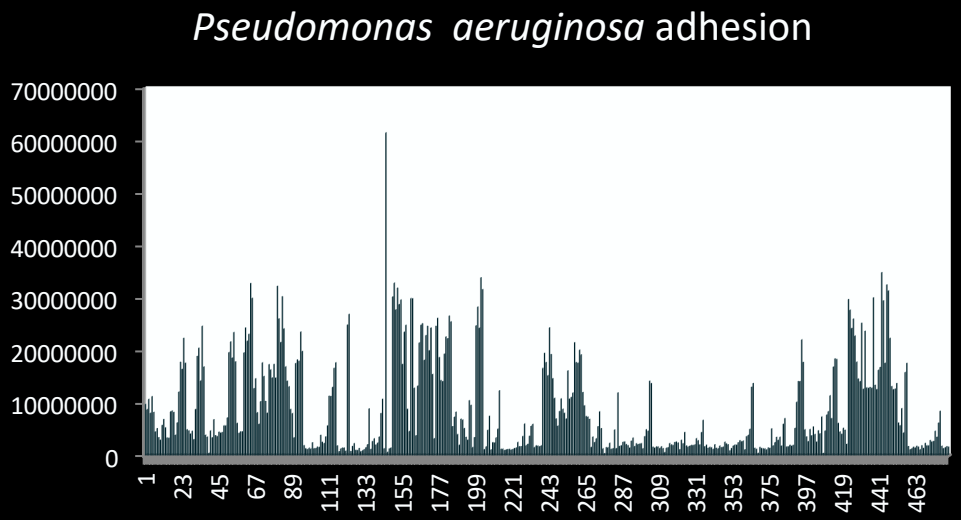
EXIT

Pathogen Attachment to Polymer Surfaces

- Bacterial adhesion and growth on biomaterial surfaces such as joint prostheses, heart valves, shunts, vascular and urinary catheters, intraocular lenses is a serious problem.
- Alexander et al. (Univ. of Nottingham) have studied adhesion of bacteria to a combinatorial library of polymer substrates.

Experimental details

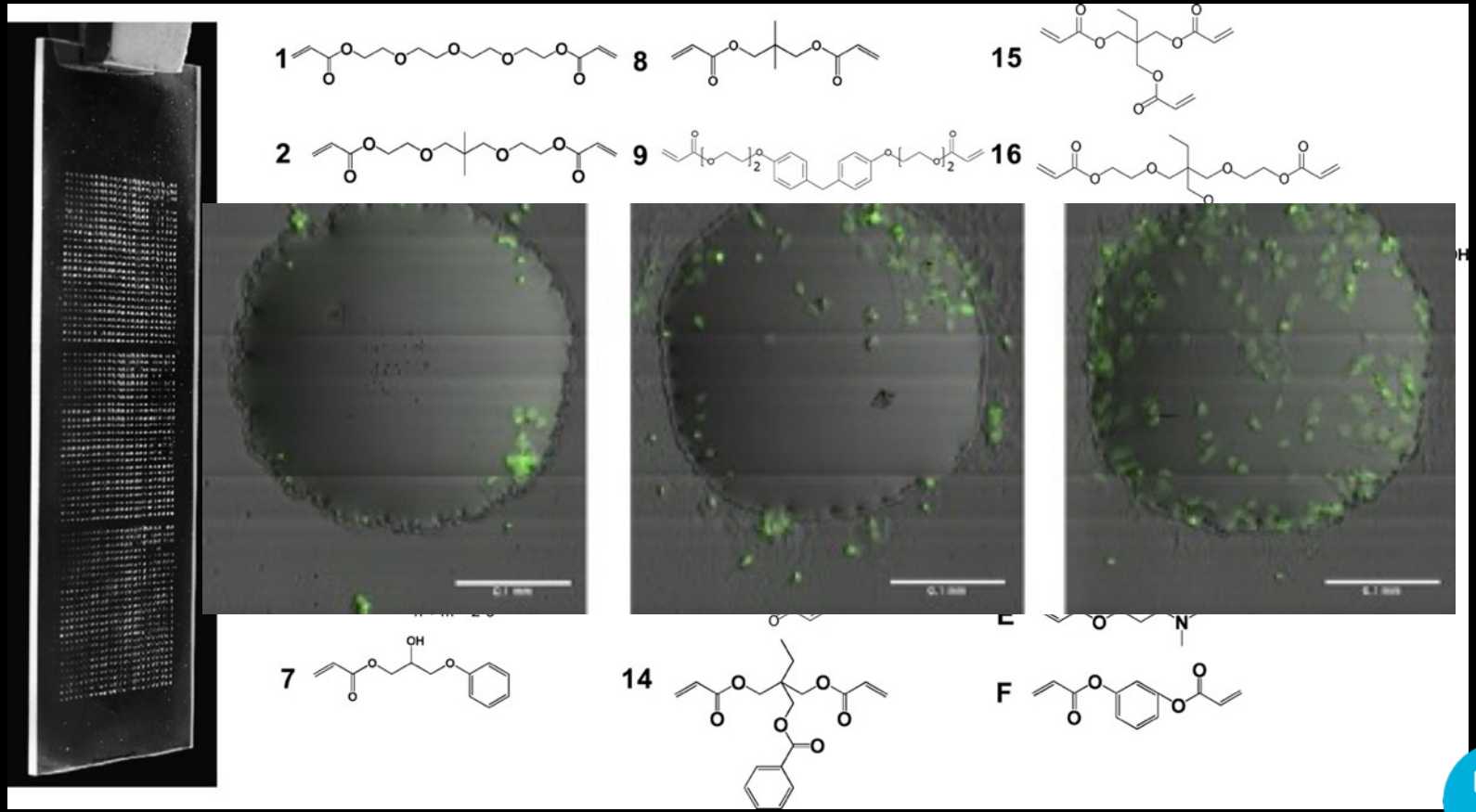


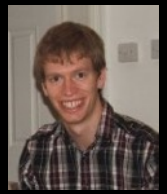


Polymer microarrays incubated with a suspension of *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and uropathogenic *Escherichia coli* (tagged with GFP or mCherry) and the fluorescent intensity measured.

The University of Nottingham

Polymer library (576 members)



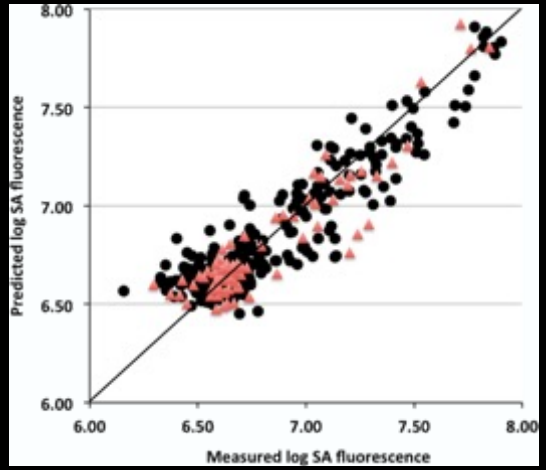


Single pathogen attachment models

LA TROBE INSTITUTE FOR MOLECULAR SCIENCE

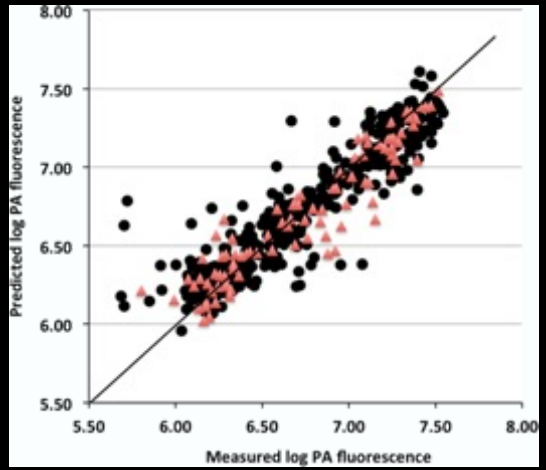
- ● Training set
- ▲ Test set

S. aureus



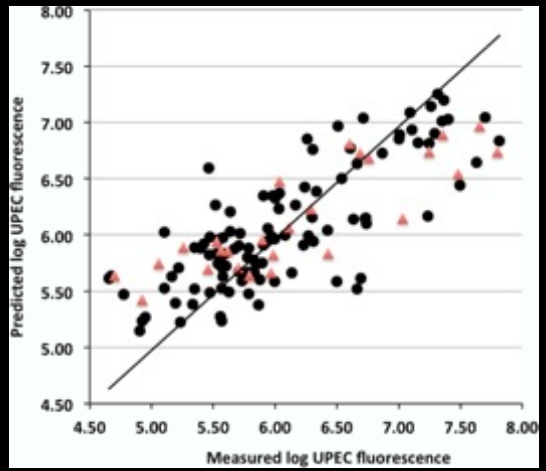
458 data points

P. aeruginosa



464 data points

Uropathogenic E. coli

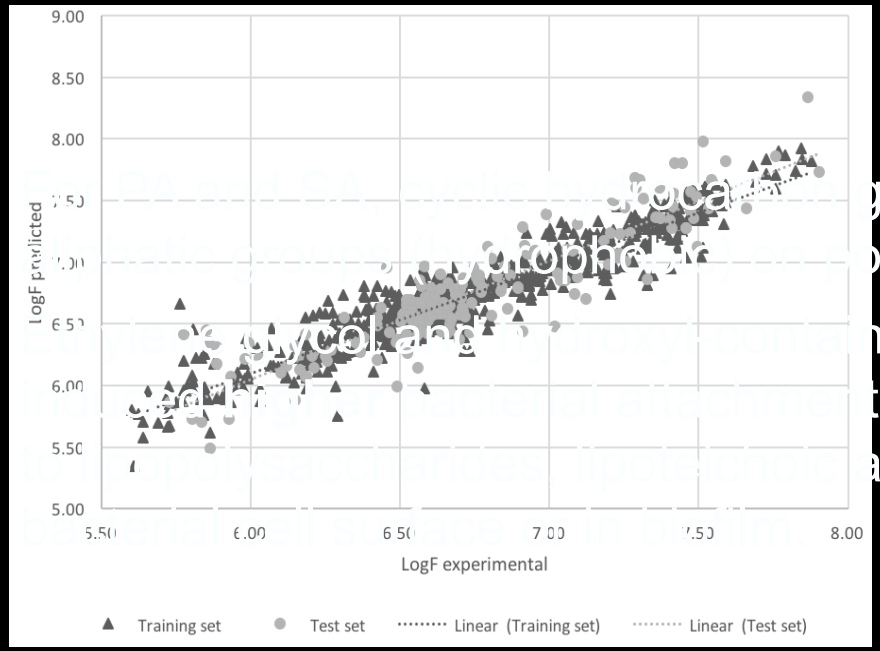


133 data points

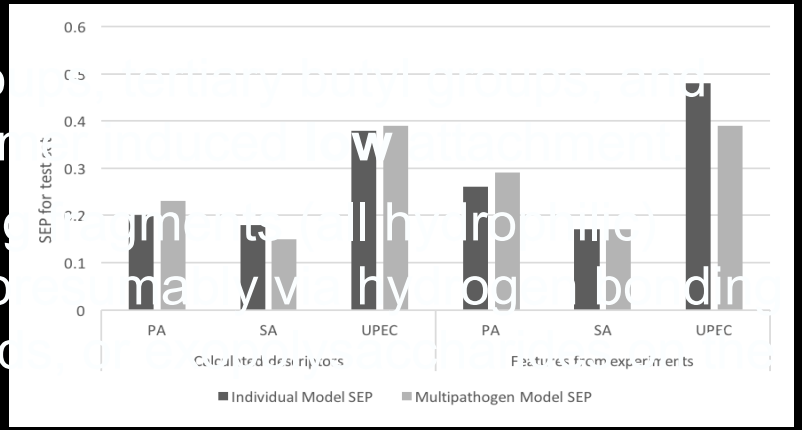
Epa et al. *Adv. Funct. Mater*. 2014; **24**: 2085

The University of Nottingham

Multiple pathogen attachment models



1055 data points



Nonlinear BRANN model

Mikulskis et al. *ACS Appl. Mater. Interf.* 2018; 10:139

New applications: –

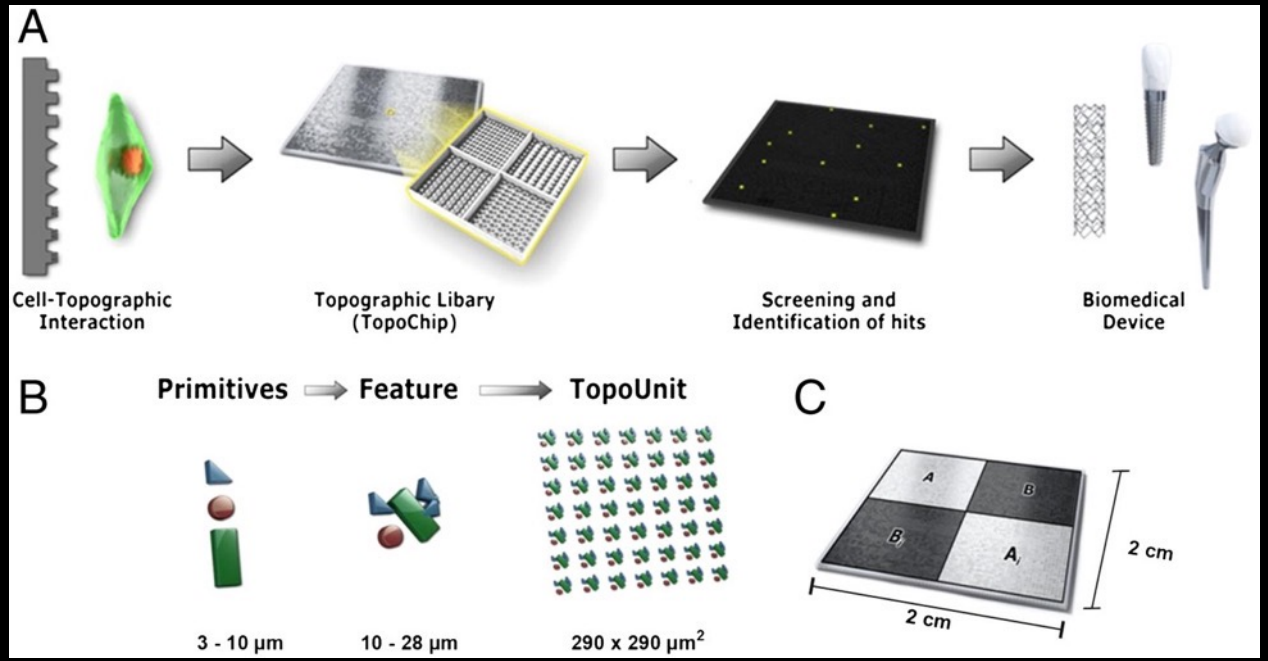
- stem cell bioreactors
- biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology

EXIT

Topographical biomaterials

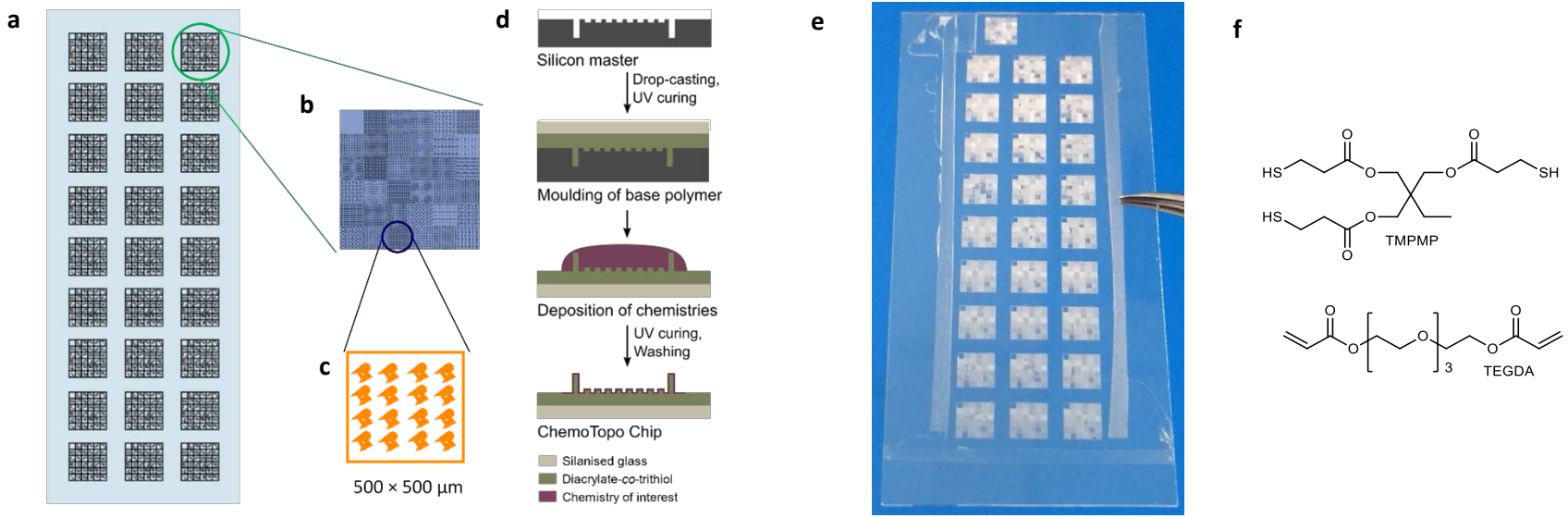
Jan de Boer

Control of cell fates achieved by surface microtopography, chemistry, or both



Unadkat et al. PNAS 2011, 108 , 16565–16570

ChemoTopoChip

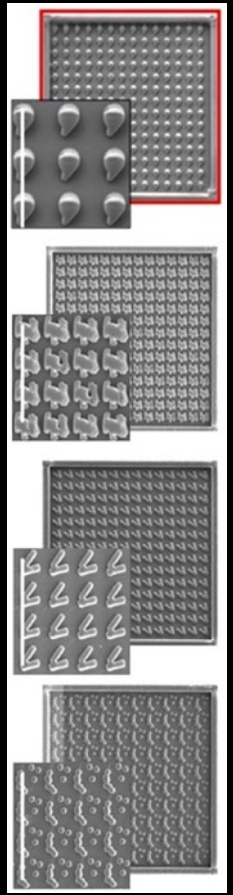


a) Schematic showing ChemoTopoChip layout (walls of 30 µm height are used to separate each Topo unit); b) ChemoTopoChip ChemoTopo unit containing 35 topographies + flat area; and 28 polymer chemistries c) Example Topo unit; d) ChemoTopoChip production process; e) Photographed ChemoTopoChip; f) TMPMP and TEGDA, used to mould ChemoTopoChip features.

Burroughs, et al. BioRxiv 2020.04.29.067421

Why do we need topographical biomaterials?

- Surface topography alone can evoke cellular responses
- Synthetic biomaterials with controlled microtopographies (TopoChips) will have instructive properties similar to growth factors
- Materials libraries that vary surface chemistry and micro/nano topographies (2D ChemoTopoChip and 3D ChemoArchiChip) have wider scope for bespoke control of cell fate
- We use libraries of 2176 different topographies



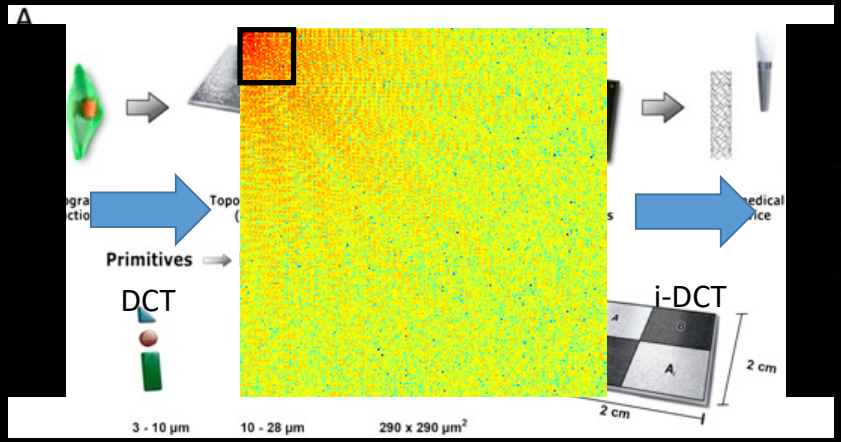
Vassey, et al. Matter, 2023, 6(3), 887-90.

Rostam, et al. Matter (Cell), 2020, 2, 1-18.

How do describe nanotopography mathematically?

Discrete cosine transform (DCT)

Figueredo, et al. Effective descriptors for machine learning models of properties of topographical biomaterials, 2023, in preparation



Topographical biomaterials

How do describe nanotopography mathematically?

Discrete cosine transform (DCT) descriptors

S aureus attachment

P aeruginosa attachment

Descriptors	P. Aeruginosa			
	R^2	RMSE	Original Size	Size post LASSO
DCT 5 frequencies	0.78	0.04	25	13
DCT 10 Frequencies	0.77	0.04	100	32
DCT 15 Frequencies	0.81	0.04	225	83
DCT 20 Frequencies	0.82	0.04	400	71
DCT 25 Frequencies	0.84	0.09	625	110
DCT 50 Frequencies	0.85	0.10	2500	269

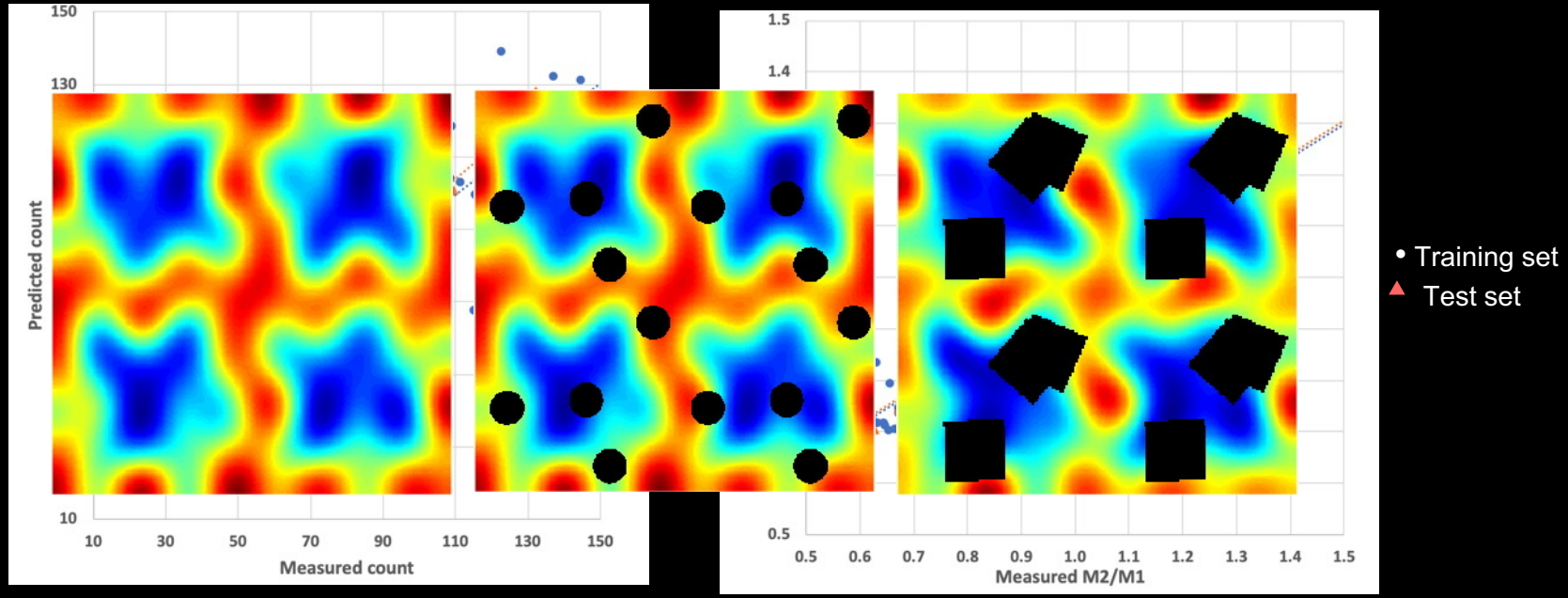
Test set predictions

Vallieres, et al., **Sci. Adv**. 2020
Vassey et al. **Adv. Sci**, 2020

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

The University of Nottingham

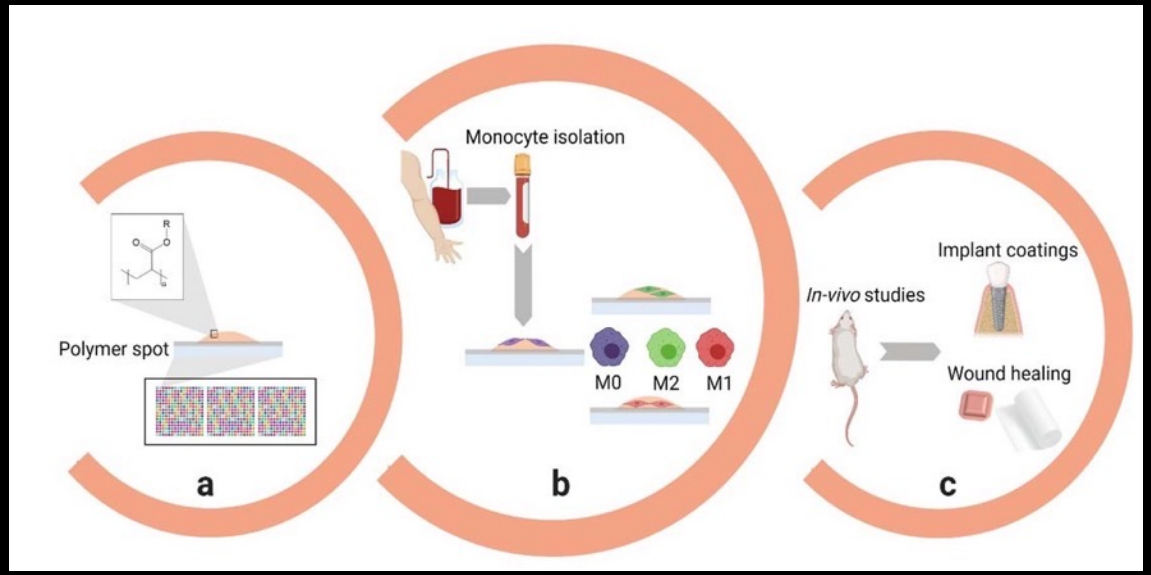
Interpreting M2 polarizing chemistries and topographies



Regression model of macrophage attachment and M2/M1 phenotype. Bayesian neural net ML method and 1-hot descriptors for the chemistries and topographies

Burroughs, et al. Biomater., 2021, 271, 120740

Control of macrophage polarization by topography

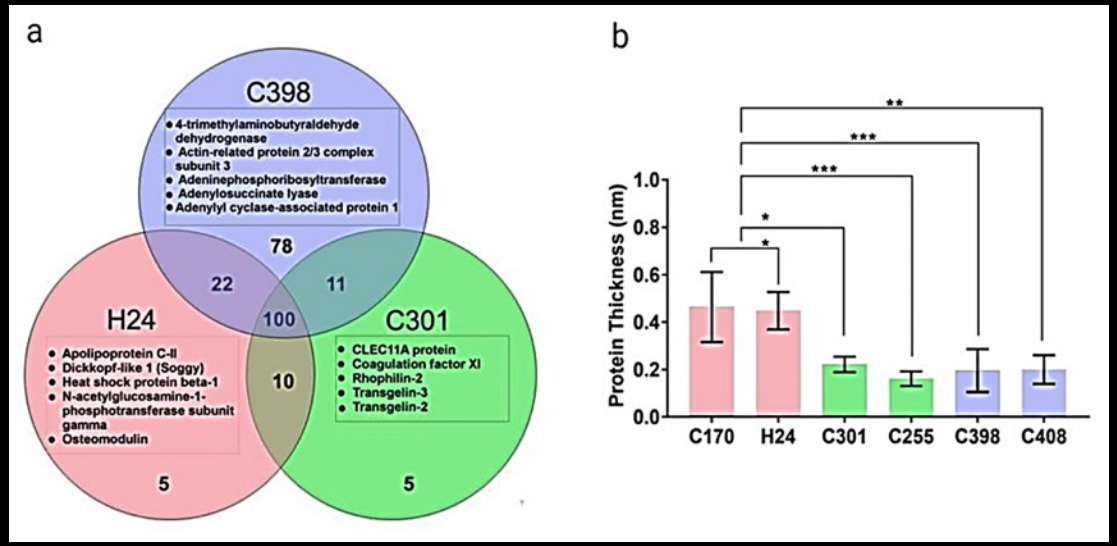


HTS approach to identify polymers that drive macrophage phenotype towards pro- or anti-inflammatory status, in-vitro and in-vivo.

Useful to encourage healing in dental and wound applications.

(a) High throughput printing of polymer arrays with different surface chemistries, (b) monocyte isolation from human buffy coats and seeding onto polymer arrays for 6 days, followed by macrophage phenotype assignment to pro-inflammatory (M1, red calprotectin fluorescent marker) and anti-inflammatory (M2, green mannose receptor fluorescent marker) phenotypes. (c) Polymers with high macrophage attachment and polarization in-vitro coated onto catheter segments, inserted subcutaneously into a mouse model, and assessed for their foreign body response.

Proteomic analysis of hit polymers



Venn diagram for number of adsorbed proteins on 3 different polymer surfaces. Overnight incubation with RPMI-1640 medium supplemented with 10% FBS, 1% L- glutamine, 1% penicillin and streptomycin

Quantification of protein adsorbate thickness on polymer spots by XPS.

Rostam et al. *Immune-Instructive Polymers Control Macrophage Phenotype and Modulate the Foreign Body Response In Vivo*, **Matter (Cell)**., 2020, 2, 1; 25; Vassey, et al. *Immune modulation by design: using topography to control human monocyte attachment and macrophage differentiation*, **Adv. Sci**, 2020, 1903392

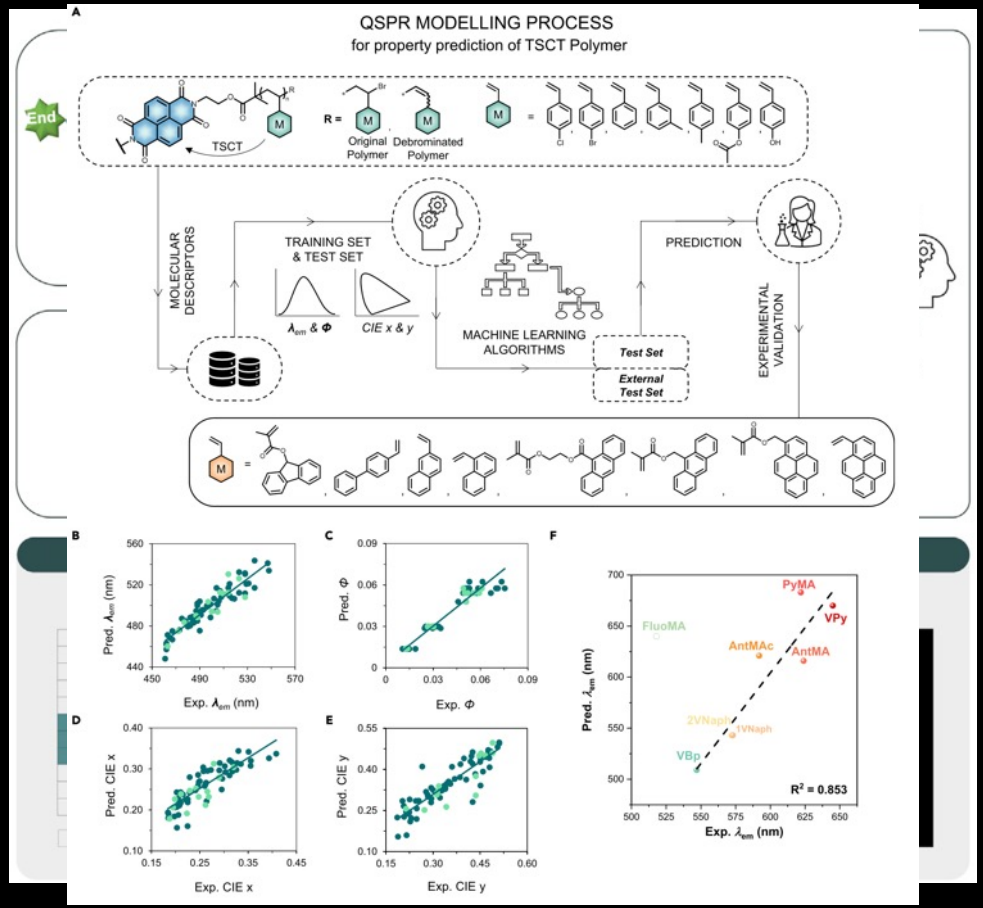
New applications: –

- stem cell bioreactors
- biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology

EXIT

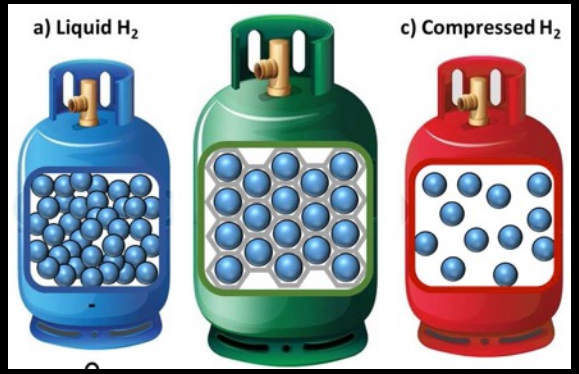
Polymer with charge transfer-dependent full-color emission

Ye, Chrisofferson et al., Machine learning-assisted exploration of a versatile polymer platform with charge transfer-dependent full-color emission, Chem (2022), 9(4), 924-947



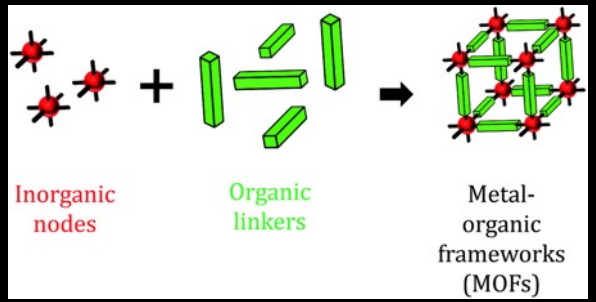
New applications: –

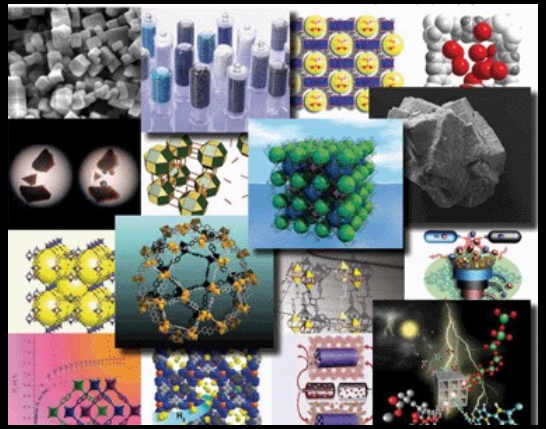
- stem cell bioreactors
- biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology



EXIT

Porous materials for hydrogen storage





Energy & Environmental Science

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

CrossMark click for updates

Cite this: *Energy Environ. Sci.*, 2015, 8, 1190

The materials genome in action: identifying the performance limits for methane storage†

Cory M. Simon,[a] Jihan Kim,[b] Diego A. Gomez-Gualdron,[c] Jeffrey S. Camp,[d] Yongchul G. Chung,[c] Richard L. Martin,[e] Rocio Mercado,[f] Michael W. Deem,[g] Dan Gunter,[e] Maciej Haranczyk,[e] David S. Sholl,[d] Randall Q. Snurr*[c] and Berend Smit*[afh]

Analogous to the way the Human Genome Project advanced an array of biological sciences by mapping the human genome, the Materials Genome Initiative aims to enhance our understanding of the fundamentals of materials science by providing the information we need to accelerate the development of new materials. This approach is particularly applicable to recently developed classes of nanoporous materials, such as metal–organic frameworks (MOFs), which are synthesized from a limited set of molecular building blocks that can be combined to generate a very large number of different structures. In this Perspective, we illustrate how a materials genome approach can be used to search for high-performance adsorbent materials to store natural gas in a vehicular fuel tank. Drawing upon recent reports of large databases of existing and predicted nanoporous materials generated *in silico*, we have collected and compared on a consistent basis the methane uptake in over 650 000 materials based on the results of molecular simulation. The data that we have collected provide candidate structures for synthesis, reveal relationships between structural characteristics and performance, and suggest that it may be difficult to reach the current Advanced Research Project Agency-Energy (ARPA-E) target for natural gas storage.

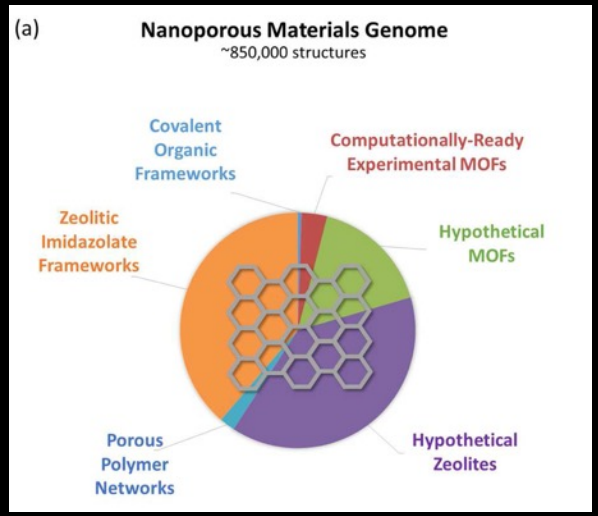
Received 6th November 2014
Accepted 12th January 2015

DOI: 10.1039/c4ee03515a

www.rsc.org/ees

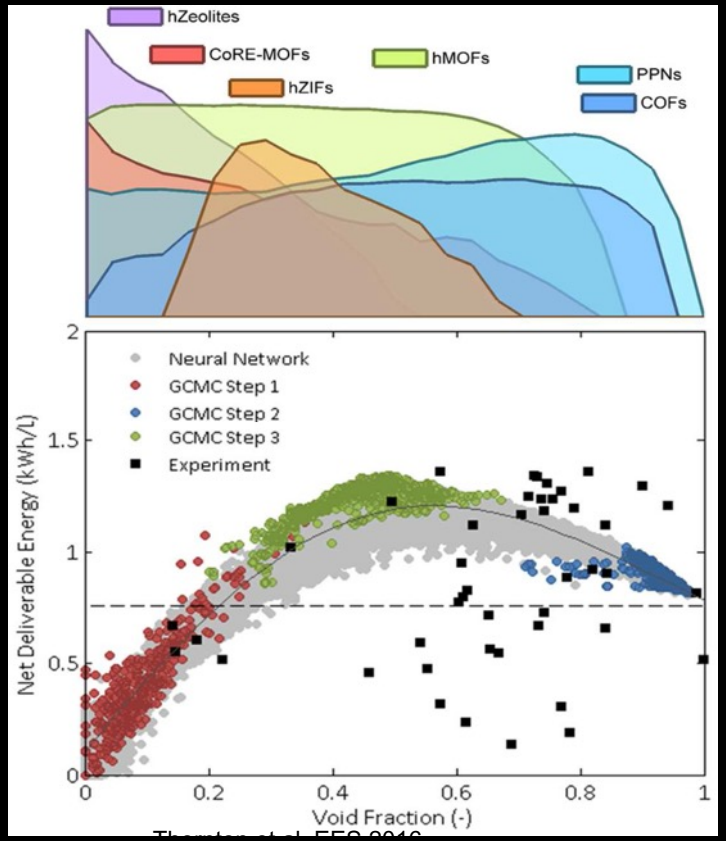
Porous materials for hydrogen storage

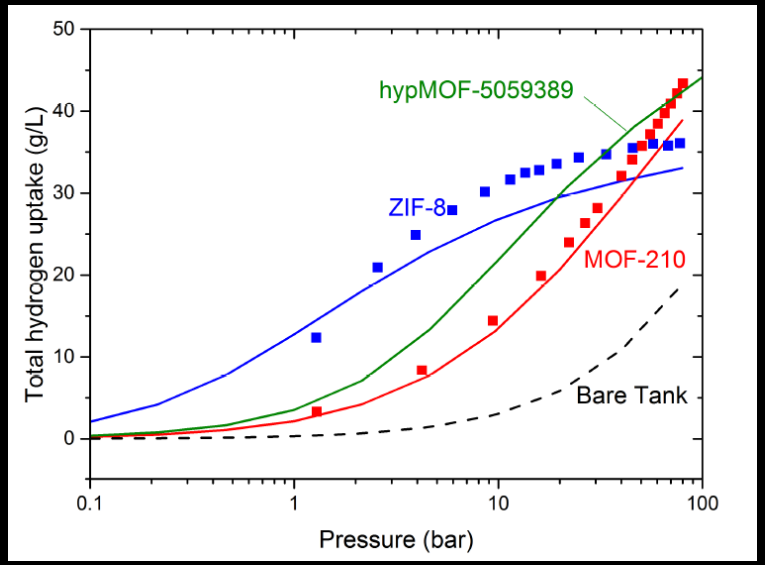
Thornton et al. EES 2016



Machine learning and in silico evolutionary methods have been useful in MOF discovery, helping explore **massive chemical spaces**

Porous materials for hydrogen storage





A combination of Grand Canonical Monte Carlo calculations and machine learning as fitness functions located the hypothetic MOFs with optimum operating properties

Porous materials for CO_2 capture and conversion

Nanoporous materials
very promising candid
for CO_2 capture and
reduction.
Materials for CO_2 red
need to adsorb H_2 an
near catalytically acti
sites.

(a) **Nanoporous Materials Genome**
~850,000 structures

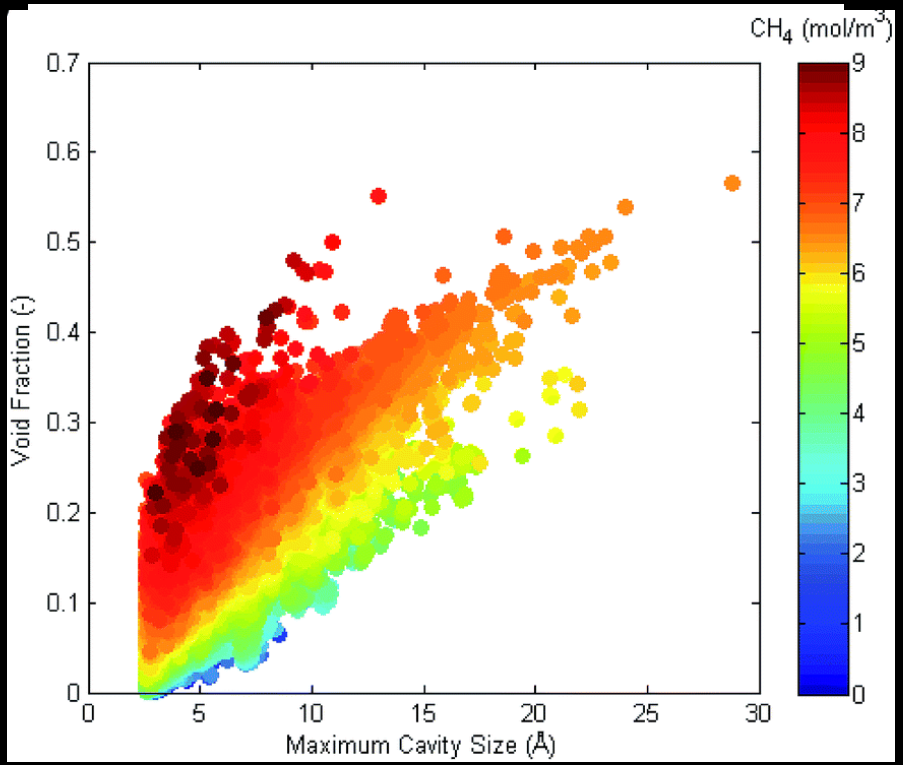
Reaction	Gas phase ΔH_{gas} (kJ mol^{-1})	Gas phase ΔS_{gas} (J mol^{-1})
Formic acid: $CO_2 + H_2 \rightarrow HCOOH$	15	−87
Formaldehyde: $CO_2 + 2H_2 \rightarrow HCOH + H_2O$	36	−64
Methanol: $CO_2 + 3H_2 \rightarrow H_3COH + H_2O$	−53	−161
Methane: $CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$	−165	−337

Top Candidates

Thornton, Winkler et al.
RSC Adv. 2015; 5, 44361

Porous materials for CO_2 and H_2 storage and reaction

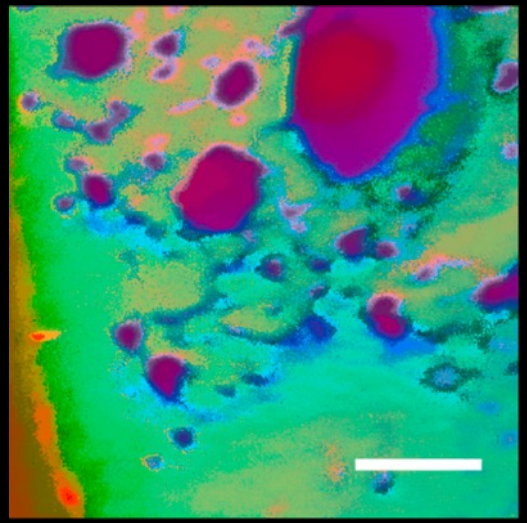
Neural network model predictions for CO_2 and H_2 uptake from 3 (10) thousand pure materials.



Thornton, Winkler et al. RSC Adv. 2015; 5, 44361

New applications: –

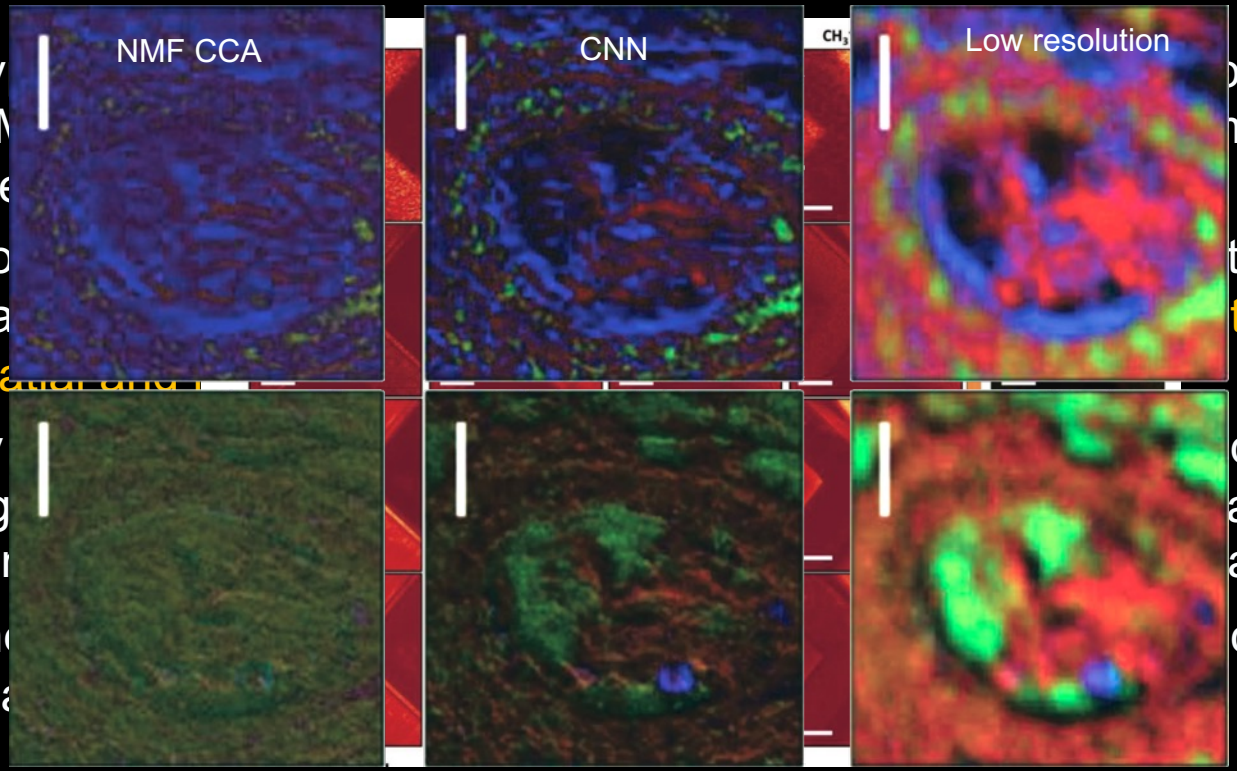
- stem cell bioreactors
- biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology



EXIT

Spatial-Spectral Resolution Enhancement using a Convolutional Neural Network

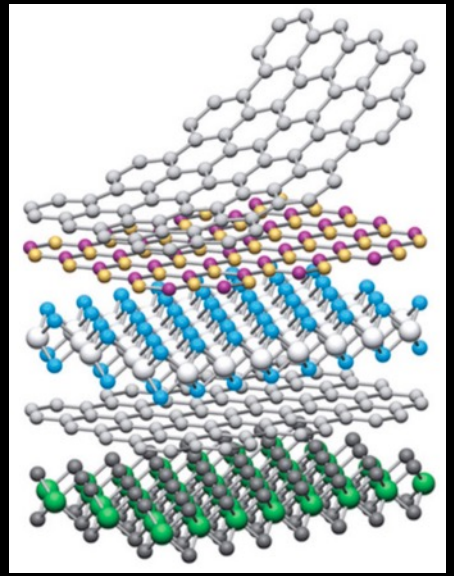
- Hy … pectrometry (ToF-SIM … mic and molecular spe …
- Mo … tion relative to most ma … trade-off between spatial and …
- By … d to fuse correlated hig … al data sets we can ger … and mass resolution.
- Th … cal correlation ana …



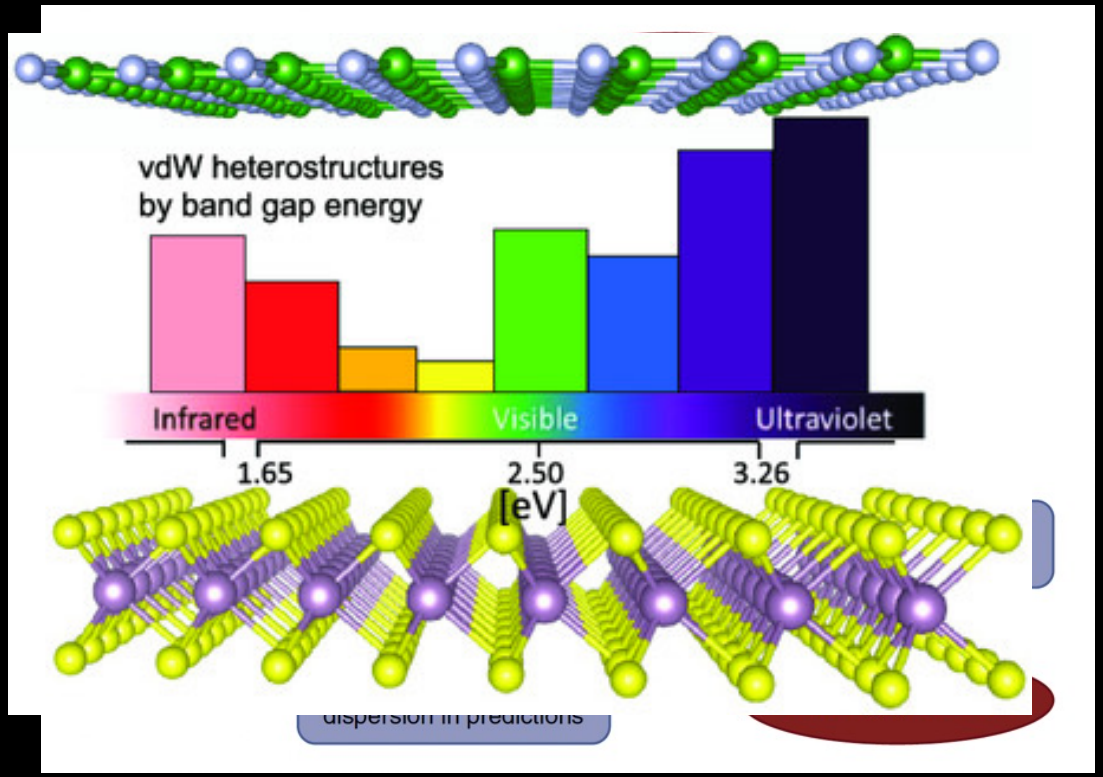
Gardner et al. Adv. Mater. Interfaces **2022**, 2201464

New applications: –

- stem cell bioreactors
- biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology



Active Learning for Bandgap Predictions of Novel 2D Heterostructures



Active learning is a special case of machine learning in which a learning algorithm can interactively query a user (or some other information source) to label new data points with the desired outputs. In statistics literature, it is sometimes called optimal experimental design.

Fronzi et al. Adv. Intell. Sys. 2021, 3, 2100080.

Australian Government
Australian Research Council

Active Learning for Bandgap Predictions of Novel 2D Heterostructures

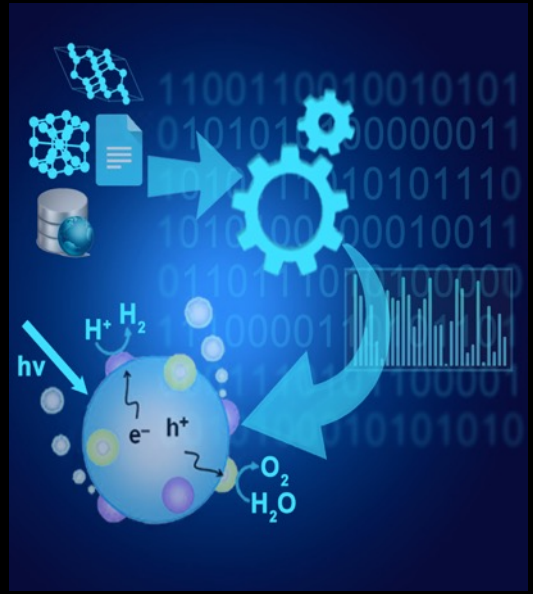
Model results are labeled progressively by four steps where each step adds additional data point sets (XAL1…XAL3) to the initial 109 bilayers.

The fifth run was carried out using additional 52 bilayers (XAL4) to test the convergence of the parameters.

Set	R^2	RMSE [eV]	MAE [eV]	MAPE [%]
First run (X_L)				
BNN-test	0.37	0.92	0.66	0.6
BNN-train	0.75	0.51	0.34	0.4
Second run (X_{AL1})				
BNN-test	0.51	0.75	0.60	0.4
BNN-train	0.77	0.51	0.35	0.3
Third run (X_{AL2})				
BNN-test	0.71	0.74	0.65	0.3
BNN-train	0.82	0.53	0.41	0.2
Fourth run (X_{AL3})				
BNN-test	0.81	0.45	0.31	0.2
BNN-train	0.92	0.41	0.28	0.1
Fifth run (X_{AL4})				
BNN-test	0.80	0.44	0.30	0.2
BNN-train	0.93	0.40	0.28	0.1

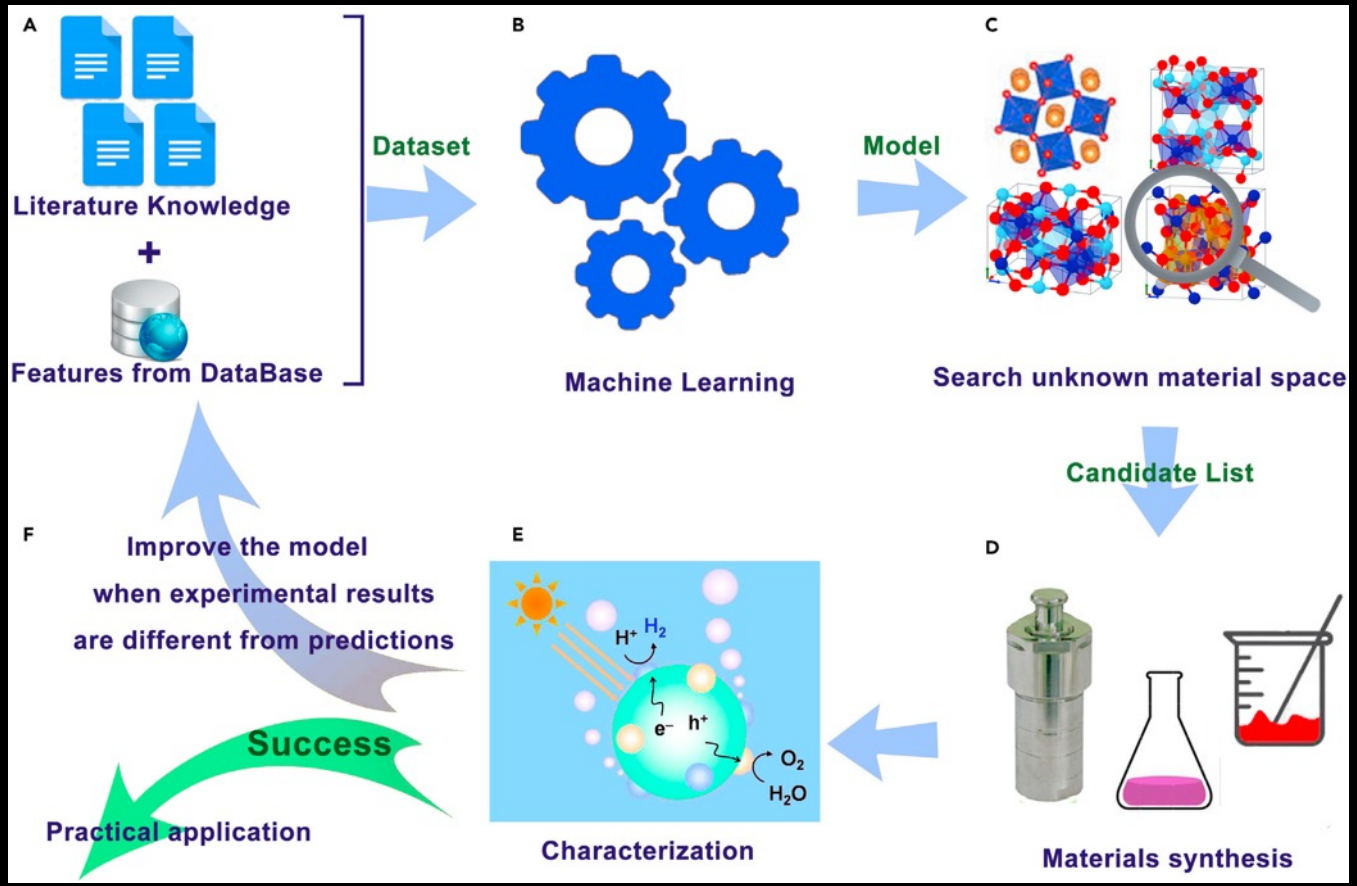
New applications: –

- stem cell bioreactors
- biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology



EXIT

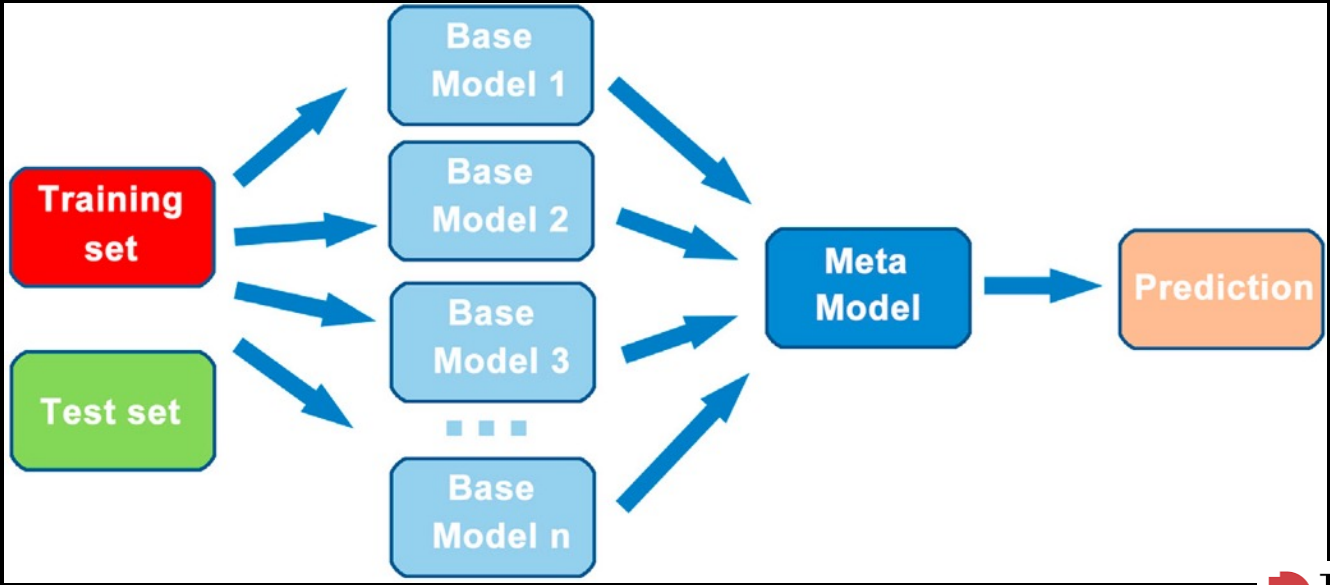
Metamodels for narrow bandgap oxide photocatalyst discovery



Mai et al. Use of Meta Models for Rapid Discovery of Narrow Bandgap Oxide Photocatalysts, iScience (Cell), 2021 24 (9), 103068.

Metamodels for narrow bandgap oxide photocatalyst discovery

Meta-learning in machine learning refers to learning algorithms that learn from other learning algorithms.



Mai et al. iScience (Cell), 2021 24 (9), 103068.

Metamodels for narrow bandgap oxide photocatalyst discovery

Table 1. Performance of various models on the bandgap predictions on the BG dataset

Models	R^2	RMSE [eV)	MAE [eV]	t-value/p value
SVR(rbf)	0.94/0.81	0.26 ± 0.00/0.47 ± 0.00	0.19 ± 0.00/0.32 ± 0.00	3.72/0.021
SVR(poly)	0.61/0.58	0.61 ± 0.00/0.63 ± 0.00	0.51 ± 0.00/0.60 ± 0.00	4.91/0.008
SVR(linear)	0.46/0.38	0.73 ± 0.00/0.74 ± 0.00	0.62 ± 0.00/0.72 ± 0.00	5.77/0.004
LASSO	0.51/0.46	0.69 ± 0.00/0.71 ± 0.00	0.66 ± 0.00/0.72 ± 0.00	5.47/0.005
Ridge	0.51/0.45	0.69 ± 0.00/0.71 ± 0.00	0.67 ± 0.00/0.72 ± 0.00	5.57/0.005
KRR	0.92/0.82	0.46 ± 0.00/0.71 ± 0.00	0.23 ± 0.00/0.35 ± 0.00	5.05/0.007
RF	0.94/0.87	0.30 ± 0.02/0.37 ± 0.05	0.18 ± 0.01/0.28 ± 0.02	11.5/0.0003
EXT	0.94/0.88	0.28 ± 0.02/0.38 ± 0.05	0.20 ± 0.01/0.30 ± 0.02	9.55/0.0007
GBR	0.99/0.87	0.10 ± 0.02/0.35 ± 0.05	0.06 ± 0.01/0.30 ± 0.02	9.30/0.0007

Results are reported as training set/test set, RMSE and MAE are acquired from the average of 100 times training/testing. Paired sample t test is carried out between one base-model and STRBG model on RMSE, degrees of freedom is 5 and the pre-selected level of significance is 0.05.

Metamodels for narrow bandgap oxide photocatalyst discovery

Table 4. Performance of the six base- and stacking meta-models on the H_2 activity classification with bandgap descriptor (results are reported as training set/test set)

Models	AUC	Accuracy	F1 score
RF	1.00/0.88	0.96/0.83	0.98/0.85
GBT	1.00/0.85	0.95/0.84	0.97/0.86
EXT	0.96/0.84	0.92/0.84	0.96/0.85
Bagging (SVC-poly)	1.00/0.85	0.99/0.82	0.92/0.83
Bagging (SVC-rbf)	1.00/0.87	0.97/0.87	0.98/0.86
Bagging (SVC-linear)	0.91/0.84	0.98/0.82	0.99/0.84
STC_{H2} II	0.99/0.97	0.98/0.96	0.98/0.96

Mai et al. Use of Meta Models for Rapid Discovery of Narrow Bandgap Oxide Photocatalysts, iScience (Cell), 2021 24 (9), 103068.

Metamodels for narrow bandgap oxide photocatalyst discovery

Table 2. Predictions of the bandgap (eV) of the 10 unknown samples from base and metamodels

Material	STR_{BG}	RF	GBR	EXT	KRR	SVR(rbf)	SVR (poly)	Reported bandgap
$(Ba_{0.5}Ni_{0.5})Bi_2NbTaO_9$	2.72	3.16	3.14	2.92	2.98	2.69	2.94	2.55
$Bi_2Ti_4O_{11}$	2.80	2.90	2.99	2.87	2.56	1.97	2.51	3.10
$Bi_5Ti_3FeO_{15}$	2.25	2.43	2.15	2.17	2.07	1.67	1.98	2.03
$Bi_6Ti_3Fe_2O_{18}$	3.37	2.42	2.15	2.17	2.57	3.06	1.55	3.72
$Ca_2Fe_2O_5$	2.22	2.43	2.00	2.14	2.22	2.06	2.39	2.10
$LiVO_3$	3.34	3.15	3.41	2.92	3.38	2.88	3.56	3.30
$KBiFe_2O_5$	1.88	3.34	1.70	2.60	3.48	3.18	3.31	1.68
$SrBi_2Nb_2O_9$	2.66	3.33	3.25	3.44	2.69	2.64	2.47	2.70
$Sr_{0.99}Bi_{2.01}Nb_{1.99}Ni_{0.01}O_{8.99}$	2.50	3.34	3.24	3.35	2.98	2.71	2.32	2.45
$Sr_{0.91}Bi_{2.09}Nb_{1.91}Ni_{0.09}O_{8.91}$	2.48	3.34	3.24	3.39	2.99	2.69	2.30	2.25
Predictions within 10%	10	1	4	3	3	3	1	

Mai et al. Use of Meta Models for Rapid Discovery of Narrow Bandgap Oxide Photocatalysts, iScience (Cell), 2021 24 (9), 103068.

New applications: –

- stem cell bioreactors
- next generation biomaterials
- topographical biomaterials
- fluorescent polymers
- porous materials for energy and environment
- surface chemistry analysis
- 2D materials
- photovoltaics
- catalysts
- corrosion & battery technology

EXIT

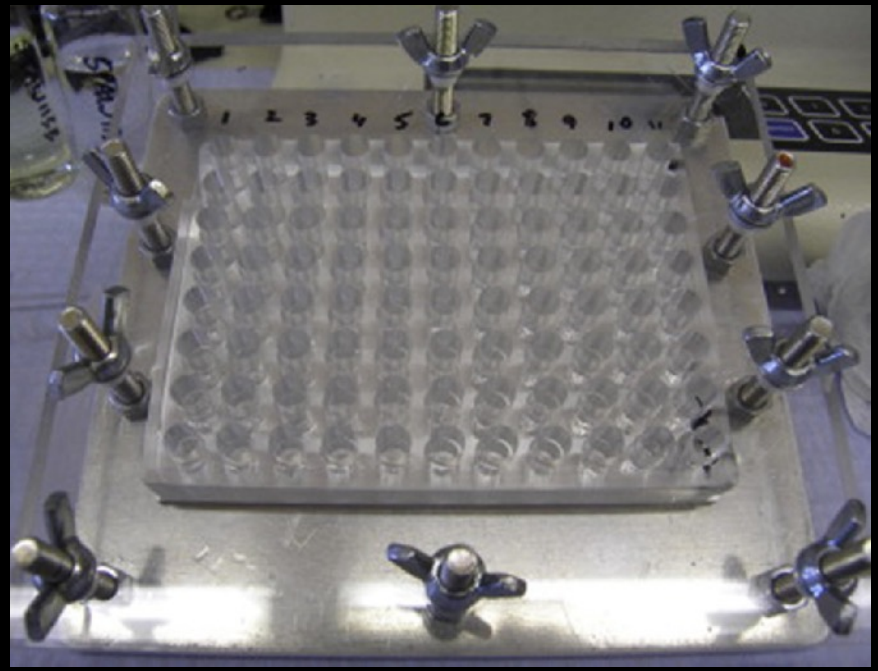
Machine learning models of White 100 inhibitor set

Inhibition measured optically after 24 hour immersion in 0.1 M NaCl

Aerospace alloy compositions : –

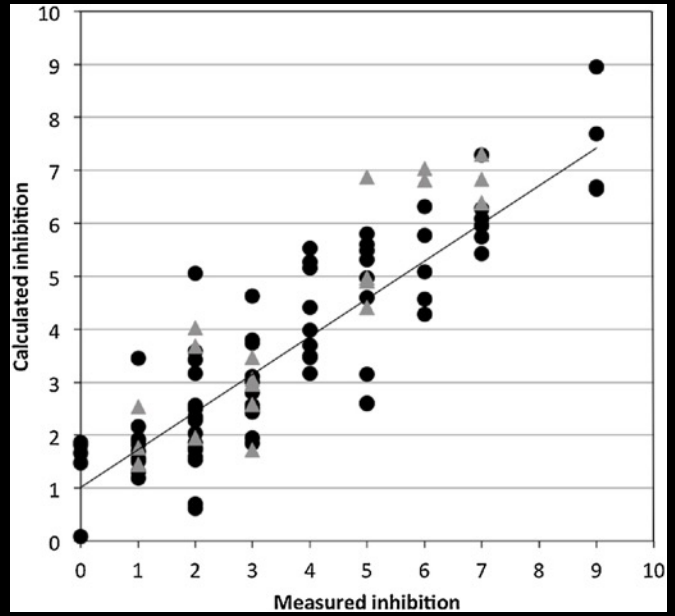
- AA2024-T3 (Cu 5.3%, Mg 1.6%, Mn 0.6%, Fe 0.2%, Zn <0.1%)
- AA7075-T6 (Cu 1.4%, Mg 2.4%, Mn <0.1%, Fe 0.2%, Zn 5.4%)

100 chemically diverse inhibitors at several initial pH values.

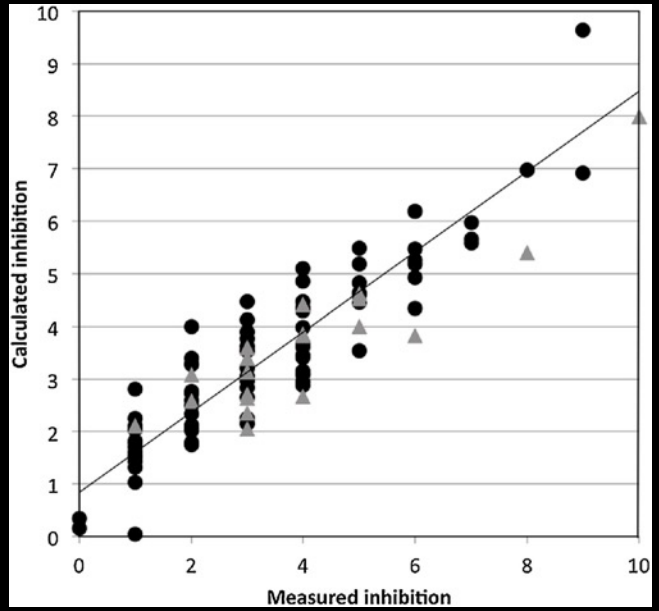


Winkler, et al. Corros. Sci. 2016, 106, 229-235

High throughput corrosion inhibition testing



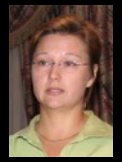
AA 7075 initial pH 4

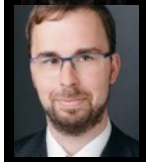


AA 2024 initial pH 4

Winkler, et al. Corros. Sci. 2016, 106, 229-235

Green organic corrosion inhibitors





- Effective [illegible] are toxic and [illegible] being banned, more benign organic [illegible]
- The structure [illegible] similarities between the [illegible] modulated [illegible] to high [illegible] SOAP kernel [illegible] with a [illegible] strategy [illegible]
- SOAP kernel compa[illegible] logical [illegible] compounds.
- The [illegible] Match kernel condenses [illegible] structures into a glob[illegible] [illegible] is defined [illegible] and is built by a [illegible]

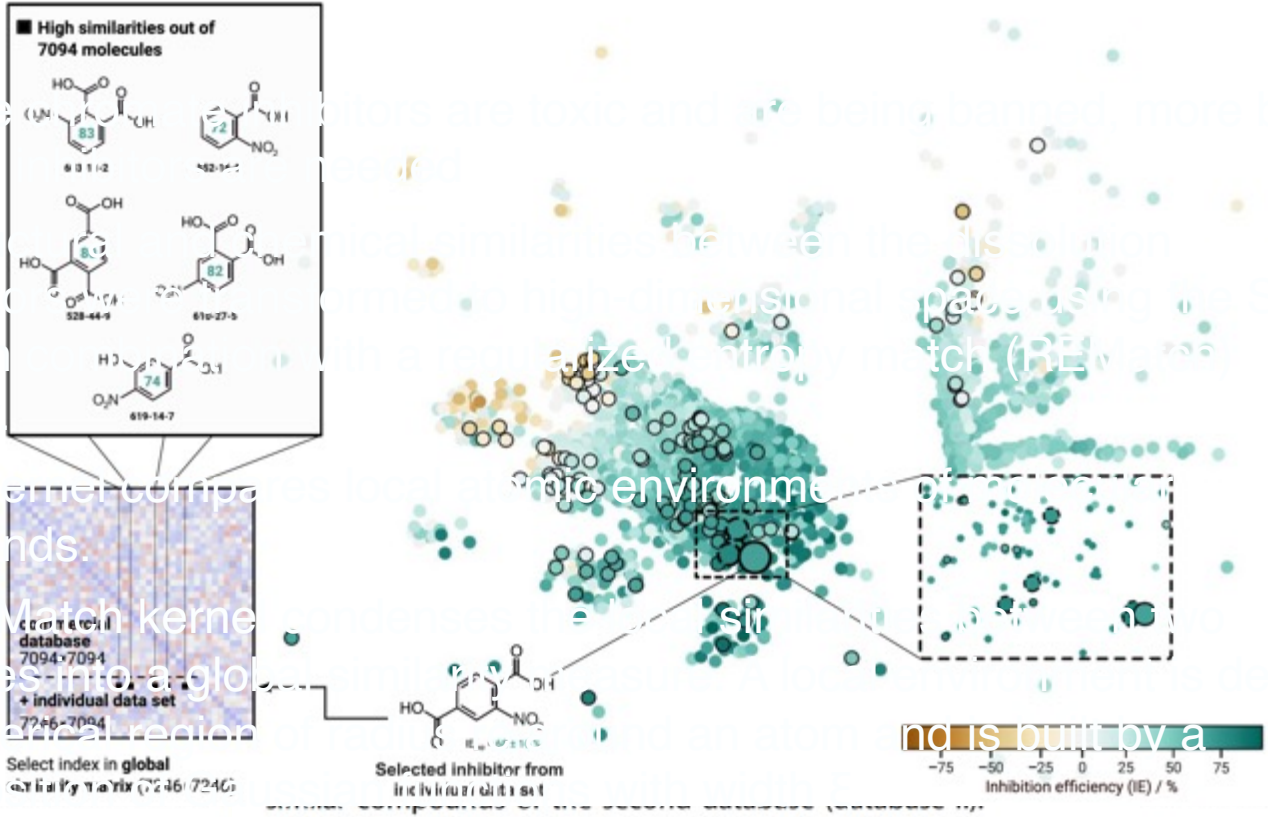
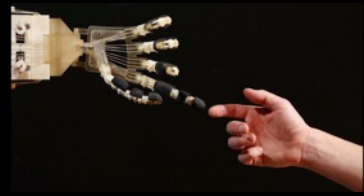


Fig. 3 Sketch-r[illegible] [illegible]ted (colored according to IE) [illegible] in the lower left corner.

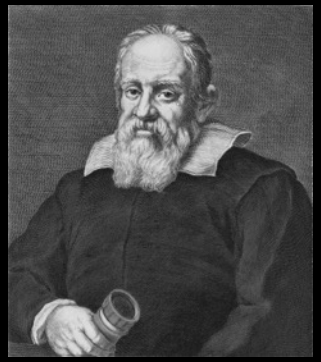
Würger et al, npj Mater. Degrad. (2021) 5:2

Remaining QSAR/machine learning issues



- Overcoming insufficient, noisy, or low diversity training data
- Generating more efficient and interpretable mathematical features/descriptors
- Selecting effective, context-aware features to reduce overfitting and aid interpretation
- Which machine learning algorithm is best?
- How to best validate model robustness, predictivity, and domain of applicability
- Understanding feature importances and how they affect the modelled properties
- Deployment of models, prediction of new data, ‘inverting’ the model to generate better molecules
- Applying the power of QSAR/machine learning to new areas
- Taking advantage of new deep learning algorithms and large language models

“*In order to understand the universe, you must know the language in which it is written. And that language is mathematics*.”
— Galileo



“We are perhaps not far removed from the time when we shall be able to submit the bulk of chemical phenomena to calculation”. Joseph Louie Gay-Lussac (1888)

Acknowledgements

- EU COST office (COST Action MODENA)
- EU H2020 grants SABYDOMA, Nano, INSIGHT, AIDD grants
- ARC Discovery Projects, Linkage, Center of Excellence, and Office of National Intelligence major grants
- UK EPSRC major grants
- Newton Turner Fellowship for exceptuional senior scientists


EVERY DAY
DISCOVERING
SOMETHING
MARCH FOR SCIENCE
LONDON
#sciencemarchldn

Thank you –

d.winkler@Latrobe.edu.au; david.winkler@monash.edu

HORIZON 2020

MARIE CURIE ACTIONS

Selected career highlights

Rebuilt QSAR using modern mathematical methods to overcome shortcomings

Mechanisms of self-organization, self-assembly, and emergent properties of complex systems

Tripeptide motifs and application to drug design

Application of AI to materials science, nanoscience, regenerative medicine

Mechanism of strontium-directed differentiation of MSCs to bone

Biomarkers for symmetric versus asymmetric stem cell division

Design of peptidomimetics, 4 clinical trials candidates, IPO

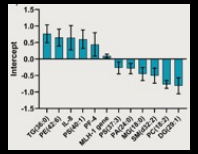
Potent thrombopoietin agonists and antagonists, first-in-class lead for myelofibrosis

Machine learning for £10M EPSRC next generation biomaterials (Nottingham)

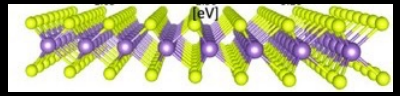
Identified outlier contributed to commercial scale amine solvent for CO_2 capture

Collaborators

Cancer biomarkers: Sanduru Thamarai Krishnan, Darren Creek, Dovile Anderson, David Rudd, Nico Voelcker (Monash) Ehud Hauben, Chandra Kirana, Guy Maddern, Kevin Fenix (Adelaide and Basil Hetzel Institute)

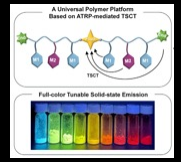


2D materials: Olexander Isayev (Carnegie Mellon), Joe Shapter (UQ), Amanda Ellis, Peter Sherrell, Nick Shepelin, Alexander Corletto (Melbourne), Marco Fronzi, Mike Ford (UTS)



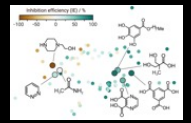
OPVs and Photocatalysts: Haoxin Mai, Tu Le, Dehong Chen, Rachel Caruso (RMIT), Takashi Hisatomi (Shinshu University) , Kazunari Domen (Tokyo)

Biomaterials: Manuel Romero, Jeni Luckett, Grazziela Figueredo, Alessandro Carabelli, David Scurr, Andrew Hook, Jean-Frédéric Dubern, Amir Ghaemmmaghami, Morgan Alexander, Paul Williams (Nottingham), Aliaksei Vasilevich, Steven Vermeulen, Jan de Boer (Eindhoven) Aurélie Carlier (Maastricht), Dan Anderson, Bob Langer (MIT)



Fluorescent polymers and perovskite solar cells: Nas Meftahi, Andrew Christofferson, Salvi Russo and colleagues (RMIT)

Batteries and green corrosion inhibitors: Mikhail Zheludkevich, Christian Feiler, Sviatlana Lamaka, Tim Würger, Rolf Meißner (Helmholtz-Zentrum hereon), Tony Hughes (CSIRO)



Surface methods: Paul Pigram, Wil Gardner, Sarah Bamford, Robert Maddiona and team (La Trobe), Ben Muir (CSIRO), Davide Ballabio (Milan)