

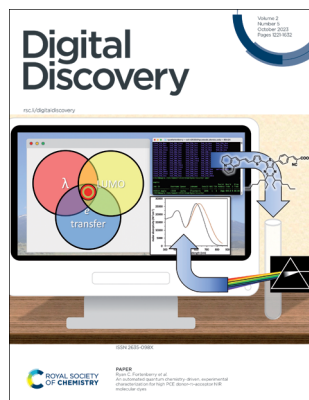
# Digital Discovery

rsc.li/digitaldiscovery

The Royal Society of Chemistry is the world's leading chemistry community. Through our high impact journals and publications we connect the world with the chemical sciences and invest the profits back into the chemistry community.

## IN THIS ISSUE

ISSN 2635-098X CODEN DDIIAI 2(5) 1221–1632 (2023)



### Cover

See Ryan C. Fortenberry et al., pp. 1269–1288.  
Image reproduced by permission of Ryan C. Fortenberry from *Digital Discovery*, 2023, 2, 1269.



### Inside cover

See Jean-Louis Reymond et al., pp. 1289–1296. Image reproduced by permission of Markus Orsi from *Digital Discovery*, 2023, 2, 1289.  
Background: David Teniers the Younger, "The Alchemist". Mauritshuis, The Hague.

## PERSPECTIVES

1233

### 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon

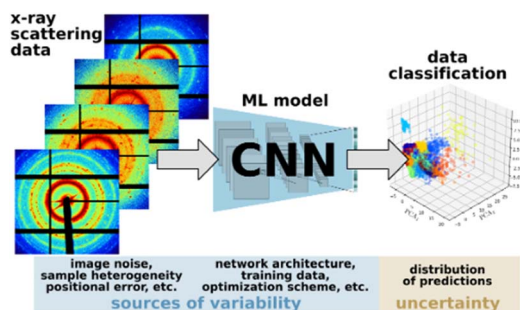
K. M. Jablonka,\* Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lala, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. Ramos, B. Ranković, S. G. Rodrigues, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik\*



1251

### A rigorous uncertainty-aware quantification framework is essential for reproducible and replicable machine learning workflows

Line Pouchard, Kristofer G. Reyes, Francis J. Alexander and Byung-Jun Yoon\*



## Editorial Staff

### Editor

Anna Rulka

### Deputy Editor

Audra Taylor

### Editorial Production Manager

Viktoria Titmus

### Assistant Editors

Shwetha Krishna, Angelica-Jane Onyekwere, Michael Whitelaw, Alexander Whiteside

### Editorial Assistant

Samantha Campos

### Publishing Assistant

Brittany Hanlon

### Publisher

Neil Hammond

For queries about submitted articles please contact Viktoria Titmus, Editorial Production Manager in the first instance. E-mail [digitaldiscovery@rsc.org](mailto:digitaldiscovery@rsc.org)

For pre-submission queries please contact Anna Rulka, Editor.

Email [digitaldiscovery-rsc@rsc.org](mailto:digitaldiscovery-rsc@rsc.org)

Digital Discovery (electronic: ISSN 2635-098X) is published 6 times a year by the Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, UK CB4 0WF.

Digital Discovery is a Gold Open Access journal and all articles are free to read. Please email [orders@rsc.org](mailto:orders@rsc.org) to register your interest or contact Royal Society of Chemistry Order Department, Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 0WF, UK Tel +44 (0)1223 432398; E-mail: [orders@rsc.org](mailto:orders@rsc.org)

Whilst this material has been produced with all due care, the Royal Society of Chemistry cannot be held responsible or liable for its accuracy and completeness, nor for any consequences arising from any errors or the use of the information contained in this publication. The publication of advertisements does not constitute any endorsement by the Royal Society of Chemistry or Authors of any products advertised. The views and opinions advanced by contributors do not necessarily reflect those of the Royal Society of Chemistry which shall not be liable for any resulting loss or damage arising as a result of reliance upon this material. The Royal Society of Chemistry is a charity, registered in England and Wales, Number 207890, and a company incorporated in England by Royal Charter (Registered No. RC000524), registered office: Burlington House, Piccadilly, London W1J 0BA, UK, Telephone: +44 (0) 207 4378 6556.

### Advertisement sales:

Tel +44 (0) 1223 432246; Fax +44 (0) 1223 426017;

E-mail [advertising@rsc.org](mailto:advertising@rsc.org)

For marketing opportunities relating to this journal, contact [marketing@rsc.org](mailto:marketing@rsc.org)

# Digital Discovery

[rsc.li/digitaldiscovery](http://rsc.li/digitaldiscovery)

*Digital Discovery* is a gold open access journal publishing top research at the intersection of chemistry, materials science and biotechnology. Blurring the barriers between computation and experimentation, we focus on the integration of digital and automation tools with science, putting data first to ensure reproducibility and faster progress.

## Editorial Board

### Editor in Chief

Alán Aspuru-Guzik, University of Toronto, Canada

### Associate Editors

Jason E. Hein, University of British Columbia, Canada  
Linda Hung, Toyota Research Institute, USA  
Joshua Schrier, Fordham University, USA  
Kedar Hippalgaonkar, Nanyang Technological University, Singapore  
Cesar de la Fuente, University of Pennsylvania, USA

### Members

Yousung Jung, KAIST, South Korea  
Anat Milo, Ben-Gurion University of the Negev, Israel  
Lilo D. Pozzo, University of Washington, USA  
Ekaterina Skorb, ITMO University, Russia

## Advisory Board

Juan Alegre, Colorado State University, USA  
Silvana Botti, Friedrich Schiller University Jena, Germany  
Pablo Carbonell, University of Valencia, Spain  
Cecilia Clementi, Freie Universität Berlin, Germany  
Conor Coley, MIT, USA

Abigail Doyle, University of California Los Angeles, USA  
Ola Engkvist, AstraZeneca and Chalmers University of Technology, Sweden  
Ian Foster, University of Chicago, USA  
Jan Jensen, University of Copenhagen, Denmark  
Heather Kulik, MIT, USA

Shuye Ping Ong, University of California San Diego, USA  
Marwin Segler, Microsoft, Germany  
Berend Smit, EPFL, Switzerland  
Isao Tanaka, Kyoto University, Japan  
Alexandre Tkatchenko, University of Luxembourg, Luxembourg  
Koji Tsuda, The University of Tokyo, Japan

## Information for Authors

Full details on how to submit material for publication in Digital Discovery are given in the Instructions for Authors (available from <http://www.rsc.org/authors>). Submissions should be made via the journal's homepage: [rsc.li/digitaldiscovery](http://rsc.li/digitaldiscovery)

Authors may reproduce/republish portions of their published contribution without seeking permission from the Royal Society of Chemistry, provided that any such republication is accompanied by an acknowledgement in the form: (Original Citation)–Reproduced by permission of the Royal Society of Chemistry.

This journal is © The Royal Society of Chemistry 2023.

Apart from fair dealing for the purposes of research or private study for non-commercial purposes, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988 and the Copyright and Related Rights Regulation 2003, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the Publishers or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency in the UK. US copyright law is applicable to users in the USA.

Registered charity number: 207890

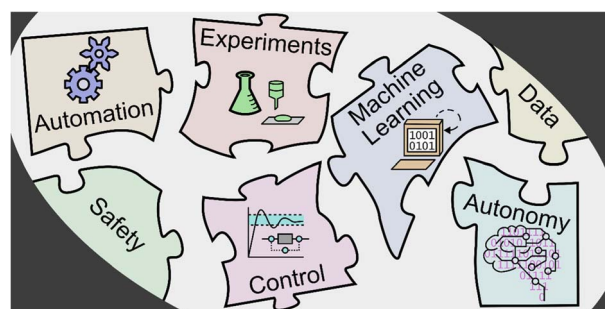


## PERSPECTIVES

1259

## Integrating autonomy into automated research platforms

Richard B. Canty, Brent A. Koscher, Matthew A. McDonald and Klavs F. Jensen\*

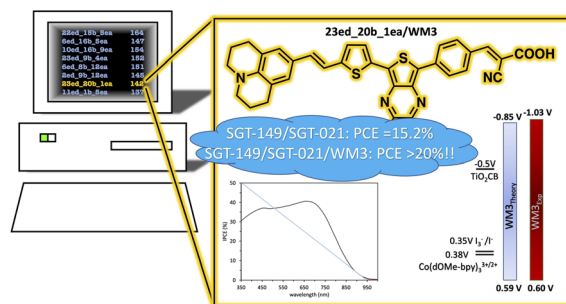


## PAPERS

1269

An automated quantum chemistry-driven, experimental characterization for high PCE donor- $\pi$ -acceptor NIR molecular dyes

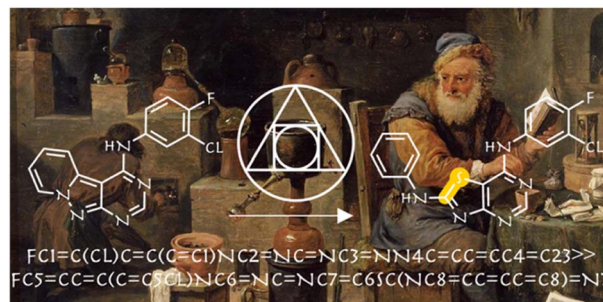
Taylor J. Santaloci, William E. Meador, Austin M. Wallace, E. Michael Valencia, Blake N. Rogers, Jared H. Delcamp and Ryan C. Fortenberry\*



1289

## Alchemical analysis of FDA approved drugs

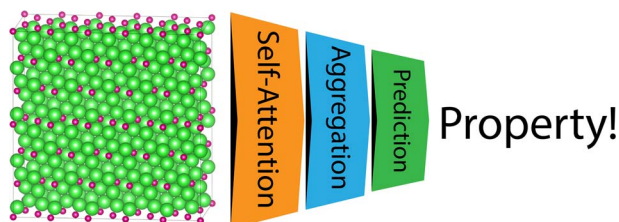
Markus Orsi, Daniel Probst, Philippe Schwaller and Jean-Louis Reymond\*



1297

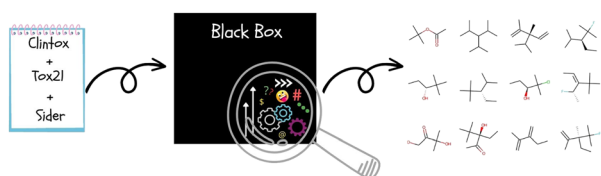
## Site-Net: using global self-attention and real-space supercells to capture long-range interactions in crystal structures

Michael Moran, Michael W. Gaultois,\* Vladimir V. Gusev and Matthew J. Rosseinsky



## PAPERS

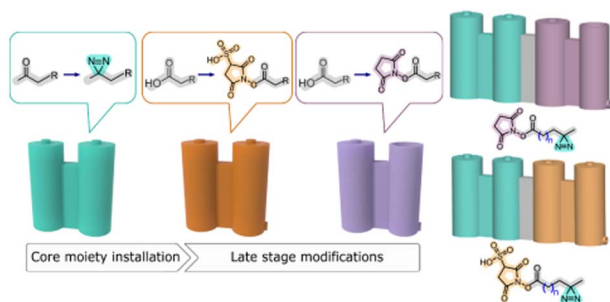
1311



### Generating structural alerts from toxicology datasets using the local interpretable model-agnostic explanations method

Cayque Monteiro Castro Nascimento, Paloma Guimarães Moura and Andre Silva Pimentel\*

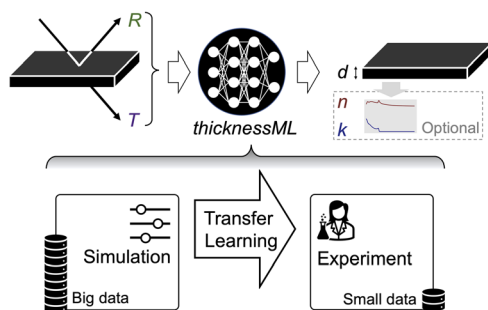
1326



### Digital design and 3D printing of reactionware for on demand synthesis of high value probes

Przemyslaw Frei, Philip J. Kitson, Alexander X. Jones and Leroy Cronin\*

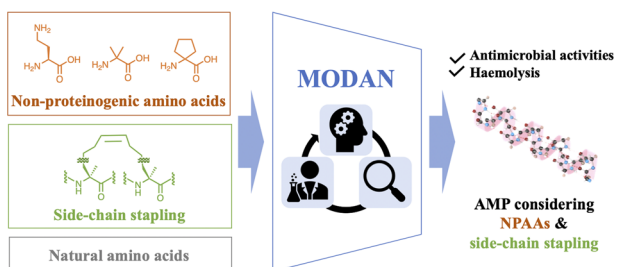
1334



### Tackling data scarcity with transfer learning: a case study of thickness characterization from optical spectra of perovskite thin films

Siyu Isaac Parker Tian, Zekun Ren, Selvaraj Venkataraj, Yuanhang Cheng, Daniil Bash, Felipe Oviedo, J. Senthilnath, Vijila Chellappan, Yee-Fun Lim, Armin G. Aberle, Benjamin P. MacLeod, Fraser G. L. Parlane, Curtis P. Berlinguette, Qianxiao Li, Tonio Buonassisi\* and Zhe Liu

1347



### Design of antimicrobial peptides containing non-proteinogenic amino acids using multi-objective Bayesian optimisation

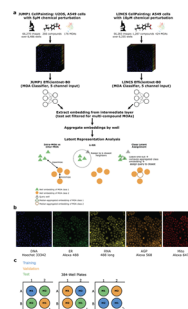
Yuki Murakami, Shoichi Ishida, Yosuke Demizu and Kei Terayama\*



## PAPERS

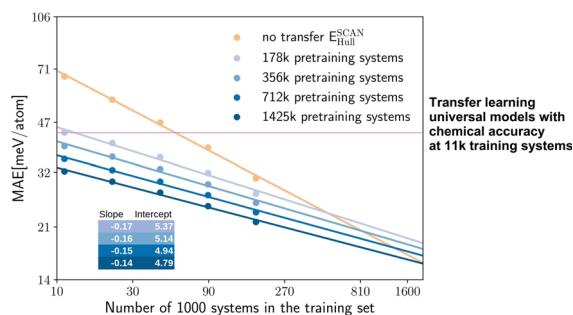
1354

## Deep representation learning determines drug mechanism of action from cell painting images

Daniel R. Wong,<sup>\*</sup> David J. Logan, Santosh Hariharan, Robert Stanton, Djork-Arné Clevert and Andrew Kiruluta

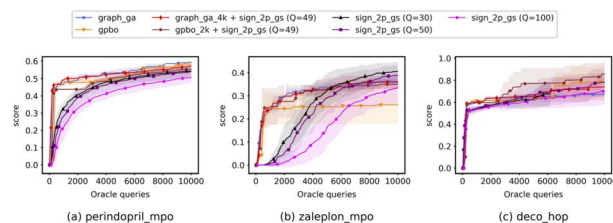
1368

## Transfer learning on large datasets for the accurate prediction of material properties

Noah Hoffmann, Jonathan Schmidt, Silvana Botti and Miguel A. L. Marques<sup>\*</sup>

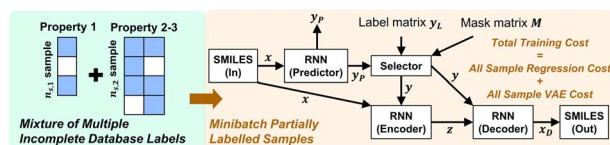
1380

## Understanding and improving zeroth-order optimization methods on AI-driven molecule optimization

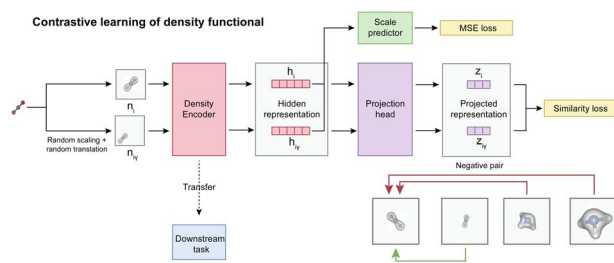
Elvin Lo and Pin-Yu Chen<sup>\*</sup>

1390

## Multi-constraint molecular generation using sparsely labelled training data for localized high-concentration electrolyte diluent screening

Jonathan P. Mailoa,<sup>\*</sup> Xin Li, Jiezhong Qiu and Shengyu Zhang<sup>\*</sup>

1404

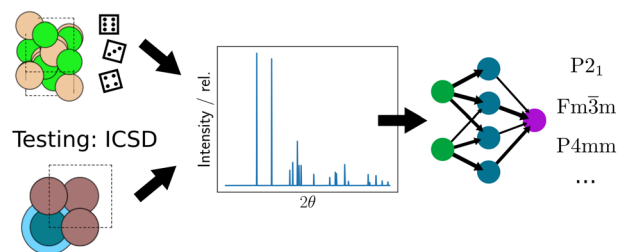


### Incorporation of density scaling constraint in density functional design via contrastive representation learning

Weiye Gong, Tao Sun, Hexin Bai, Shah Tanvir ur Rahman Chowdhury, Peng Chu, Anoj Aryal, Jie Yu, Haibin Ling,\* John P. Perdew\* and Qimin Yan\*

1414

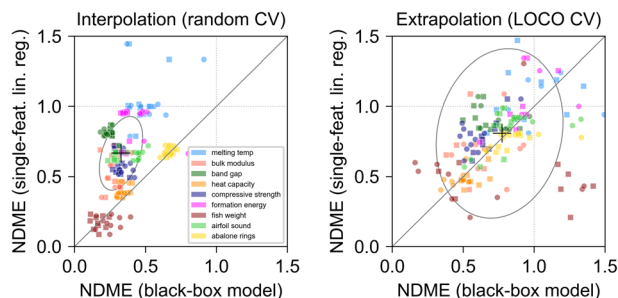
Training: Synthetic



### Neural networks trained on synthetically generated crystals can extract structural information from ICSD powder X-ray diffractograms

Henrik Schopmans, Patrick Reiser and Pascal Friederich\*

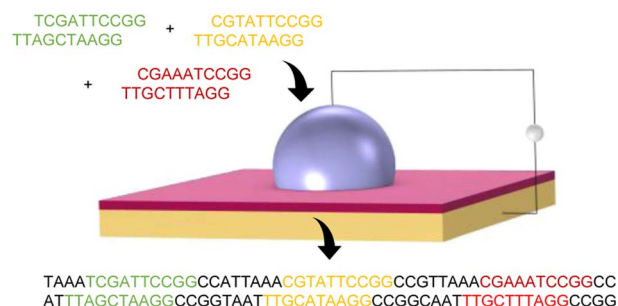
1425



### Interpretable models for extrapolation in scientific machine learning

Eric S. Muckley, James E. Saal,\* Bryce Meredig, Christopher S. Roper and John H. Martin

1436



### Automated routing of droplets for DNA storage on a digital microfluidics platform

Ajay Manicka, Andrew Stephan, Sriram Chari, Gemma Mendonsa, Peyton Okubo, John Stolzberg-Schray, Anil Reddy and Marc Riedel

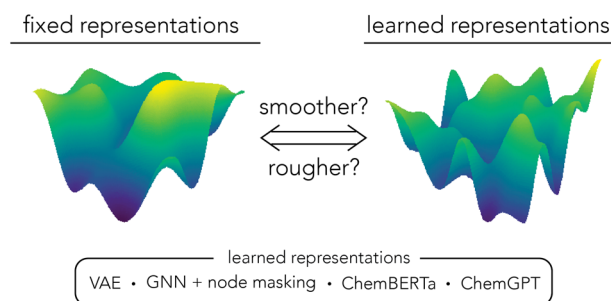


## PAPERS

1452

## Evaluating the roughness of structure–property relationships using pretrained molecular representations

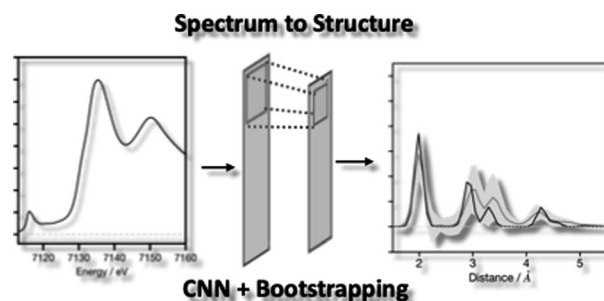
David E. Graff, Edward O. Pyzer-Knapp, Kirk E. Jordan, Eugene I. Shakhnovich and Connor W. Coley



1461

## Towards the automated extraction of structural information from X-ray absorption spectra

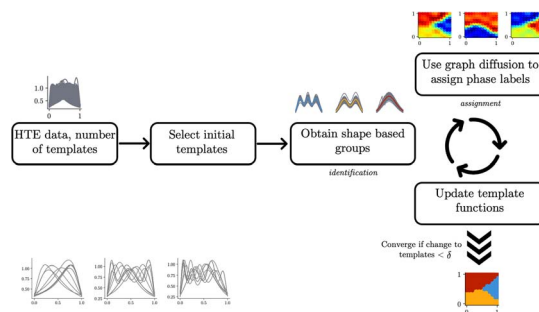
Tudur David,\* Nik Khadijah Nik Aznan, Kathryn Garside and Thomas Penfold



1471

## Metric geometry tools for automatic structure phase map generation

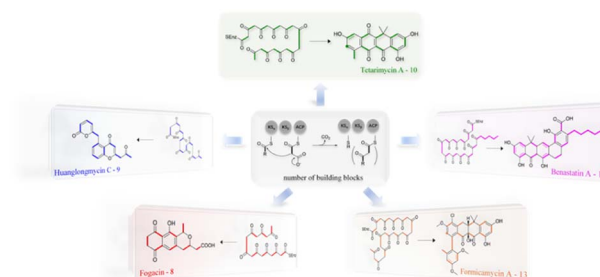
Kiran Vaddi,\* Karen Li and Lilo D. Pozzo



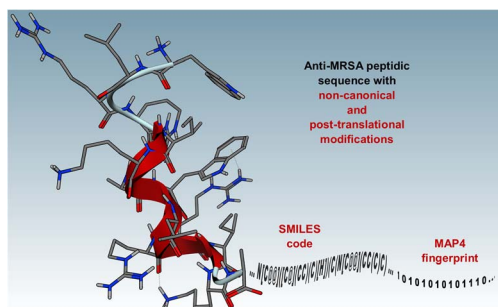
1484

## A deep learning model for type II polyketide natural product prediction without sequence alignment

Jiaquan Huang, Qiandi Gao, Ying Tang, Yaxin Wu, Heqian Zhang\* and Zhiwei Qin\*



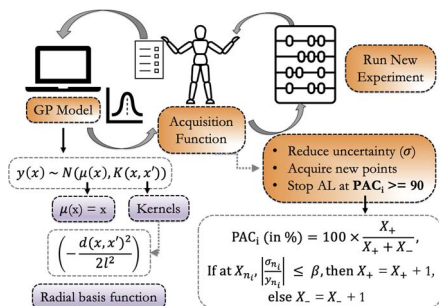
1494



## Mapping the structure–activity landscape of non-canonical peptides with MAP4 fingerprinting

Edgar López-López,<sup>\*</sup> Oscar Robles, Fabien Plisson and José L. Medina-Franco<sup>\*</sup>

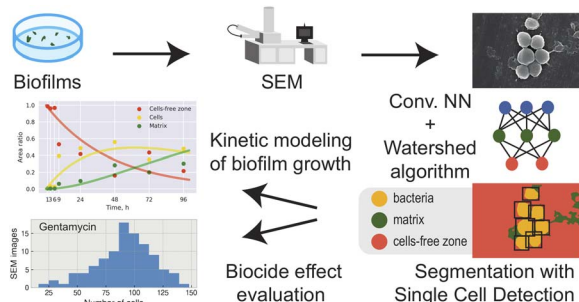
1506



## Active learning for efficient navigation of multi-component gas adsorption landscapes in a MOF

Krishnendu Mukherjee, Etinosa Osaro and Yamil J. Colón<sup>\*</sup>

1522



## Digital biology approach for macroscale studies of biofilm growth and biocide effects with electron microscopy

Konstantin S. Kozlov, Daniil A. Boiko, Elena V. Detusheva, Konstantin V. Detushev, Evgeniy O. Pentsak, Anatoly N. Vereshchagin and Valentine P. Ananikov<sup>\*</sup>

1540



## Go with the flow: deep learning methods for autonomous viscosity estimations

Michael Walker, Gabriella Pizzuto, Hatem Fakhruldeen and Andrew I. Cooper<sup>\*</sup>



## Using GPT-4 in parameter selection of polymer informatics: improving predictive accuracy amidst data scarcity and 'Ugly Duckling' dilemma

The diagram consists of three concentric ellipses. The outermost ellipse is light blue and labeled 'All descriptors'. The middle ellipse is light orange and labeled 'GPT-4 oriented'. The innermost ellipse is light green and labeled 'Data-oriented'. To the right of these ellipses is a vertical arrow pointing downwards, labeled 'Selection'.

## Element similarity in high-dimensional materials representations

## Chemical Elements as Vectors

One hot encoding	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \dots \end{bmatrix}$ H
Distributed embedding	$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \dots \end{bmatrix}$ From machine learning

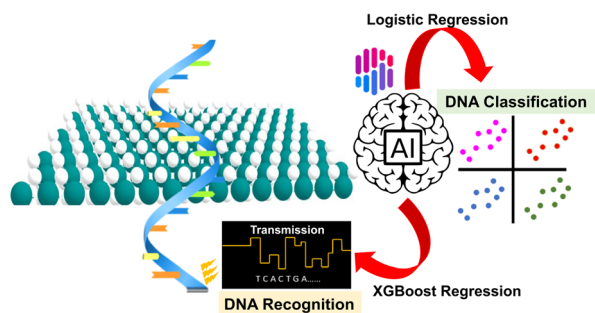
# Density functional theory and machine learning for electrochemical square-scheme prediction: an application to quinone-type molecules relevant to redox flow batteries

# Combined data-driven and mechanism-based approaches for human-intestinal-absorption prediction in the early drug-discovery stage

The diagram illustrates two contrasting approaches to drug discovery, centered around a brain icon on a desk with a laptop and books.

- Data-driven Approach (Left):** This approach is represented by a large stack of papers. Two callouts show chemical structures with their respective frequencies: "Fa: 5%" and "Fa: 99%". Arrows point from these structures to the stack of papers, indicating a data-centric process.
- Mechanism-based Approach (Right):** This approach is represented by a glass of water labeled "Do" (Drug) and "On" (Organism). A red, multi-colored structure labeled "Pn" (Protein) is shown interacting with the water. Arrows point from the glass to the protein, indicating a process based on understanding the underlying mechanism.

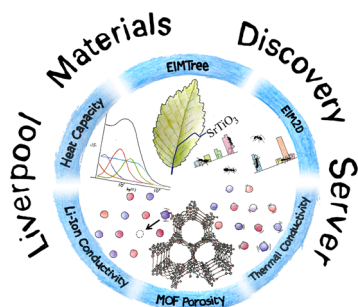
1589



### Artificial intelligence aided recognition and classification of DNA nucleotides using MoS<sub>2</sub> nanochannels

Sneha Mittal, Souvik Manna, Milan Kumar Jena and Biswarup Pathak\*

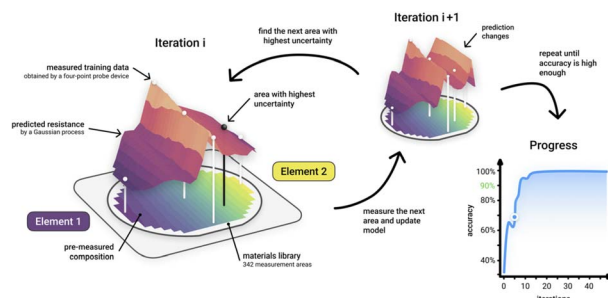
1601



### The Liverpool materials discovery server: a suite of computational tools for the collaborative discovery of materials

Samantha Durdy, Cameron J. Hargreaves, Mark Dennison, Benjamin Wagg, Michael Moran, Jon A. Newnham, Michael W. Gaultois, Matthew J. Rosseinsky and Matthew S. Dyer\*

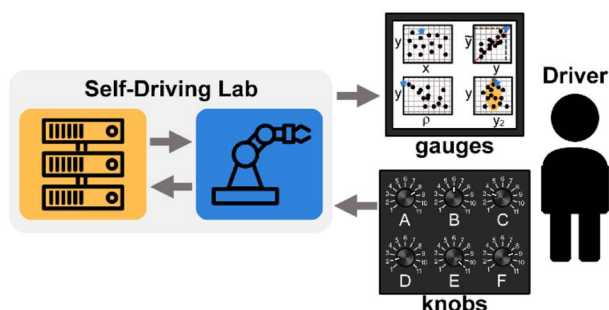
1612



### Speeding up high-throughput characterization of materials libraries by active learning: autonomous electrical resistance measurements

Felix Thelen, Lars Banko, Rico Zehl, Sabrina Baha and Alfred Ludwig\*

1620



### Driving school for self-driving labs

Kelsey L. Snapp and Keith A. Brown\*

