Data-driven Discovery in the Chemical Sciences

Trinity College, Oxford,
United Kingdom and online

10–12 September 2024

FARADAY DISCUSSIONS

Volume 256, 2025

ROYAL SOCIETY OF CHEMISTRY

This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Open Access Article. Published on 16 January 2025. Downloaded on 1/11/2026 10:52:41 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

The **Faraday Community for Physical Chemistry** of the Royal Society of Chemistry, previously the Faraday Society, was founded in 1903 to promote the study of sciences lying between chemistry, physics and biology.

Editorial Staff

Executive Editor
Michael A. Rowan

Deputy Editor
Edward Gardner

Development Editors
Bee Hockin, Andrea Carolina Ojeda-Porras

Editorial Manager
Gisela Scott

Associate Editorial Manager
Chris Goodall

Publishing Coordinator
Konoya Das

Publishing Editors
Emma Gorrell and Lauren Yarrow-Wright

Editorial Assistant
Daphne Houston

Publishing Assistants
Lee Colwill and Robert Griffiths

Publisher
Sam Keltie

Faraday Discussions (Print ISSN 1359-6640, Electronic ISSN 1364-5498) is published 8 times a year by the Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, UK CB4 0WF.

Volume 256 ISBN 978-1-83767-442-8

2025 annual subscription price: print+electronic £1342 US $2363; electronic only £1279, US $2250.
Customers in Canada will be subject to a surcharge to cover GST. Customers in the EU subscribing to the electronic version only will be charged VAT.

All orders, with cheques made payable to the Royal Society of Chemistry, should be sent to the 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**

If you take an institutional subscription to any Royal Society of Chemistry journal you are entitled to free, site-wide web access to that journal. You can arrange access via Internet Protocol (IP) address at **www.rsc.org/ip**

Customers should make payments by cheque in sterling payable on a UK clearing bank or in US dollars payable on a US clearing bank.

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.

Printed in the UK

MIX
Paper from responsible sources
FSC® C013604
FSC www.fsc.org

ROYAL SOCIETY OF CHEMISTRY

Faraday Discussions

Faraday Discussions are unique international discussion meetings that focus on rapidly developing areas of chemistry and its interfaces with other scientific disciplines.

Scientific Committee volume 256

Co-Chairs
Volker Deringer, University of Oxford, UK
Fernanda Duarte, University of Oxford, UK

Committee
Graeme Day, University of Southampton, UK
Janine George, Federal Institute for Materials Research and Testing (BAM), Germany
Nadine Schneider, Novartis, Switzerland
Philippe Schwaller, École Polytechnique Fédérale de Lausanne, Switzerland

Faraday Standing Committee on Conferences

Chair
Susan Perkin, University of Oxford, UK

Secretary
Susan Weatherby, Royal Society of Chemistry, UK

George Booth, King's College London, UK
Rachel Evans, University of Cambridge, UK
David Fermin, University of Bristol, UK
Julia Lehman, University of Birmingham, UK
David Lennon, University of Glasgow, UK
Andrew Mount, University of Edinburgh, UK
Julia Weinstein, University of Sheffield, UK

Advisory Board

Vic Arcus, The University of Waikato, New Zealand
Timothy Easun, Cardiff University, UK
Dirk Guldi, University of Erlangen-Nuremberg, Germany
Marina Kuimova, Imperial College London, UK
Luis Liz-Marzán, CIC biomaGUNE, Spain
Andrew Mount, University of Edinburgh, UK
Frank Neese, Max Planck Institute for Chemical Energy Conversion, Germany
Michel Orrit, Leiden University, The Netherlands
Zhong-Qun Tian, Xiamen University, China
Siva Umapathy, Indian Institute of Science, Bangalore, India
Bert Weckhuysen, Utrecht University, The Netherlands
Julia Weinstein, University of Sheffield, UK
Sihai Yang, University of Manchester, UK

Information for Authors

This journal is © the Royal Society of Chemistry 2025. 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.

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48–1992 (Permanence of Paper).

Registered charity number: 207890

Open Access Article. Published on 16 January 2025. Downloaded on 1/11/2026 10:52:41 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Data-driven Discovery in the Chemical Sciences

Faraday Discussions

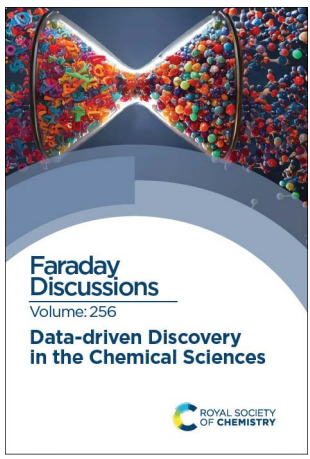
www.rsc.org/faraday_d

A General Discussion on Data-driven Discovery in the Chemical Sciences was held in Oxford, UK and online on the 10th, 11th and 12th of September 2024.

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.

CONTENTS

ISSN 1359-6640; ISBN 978-1-83767-442-8



Cover
See Brett Savoie *et al.*, *Faraday Discuss.*, 2025, **256**, 104–119.

Large property models convert design constraints (depicted by letters- left) directly into molecules (emerging from the model - right).

Image reproduced with permission of Brett Savoie from B. Savoie *et al.*, *Faraday Discuss.*, 2025, **256**, 104–119.

INTRODUCTORY LECTURE

10 **Spiers Memorial Lecture: How to do impactful research in artificial intelligence for chemistry and materials science**
Austin H. Cheng, Cher Tian Ser, Marta Skreta, Andrés Guzmán-Cordero, Luca Thiede, Andreas Burger, Abdulrahman Aldossary, Shi Xuan Leong, Sergio Pablo-García, Felix Strieth-Kalthoff and Alán Aspuru-Guzik

PAPERS AND DISCUSSIONS

61 **Beyond theory-driven discovery: introducing hot random search and datum-derived structures**
Chris J. Pickard

85 **Integration of generative machine learning with the heuristic crystal structure prediction code FUSE**
Christopher M. Collins, Hasan M. Sayeed, George R. Darling, John B. Claridge, Taylor D. Sparks and Matthew J. Rosseinsky

Open Access Article. Published on 16 January 2025. Downloaded on 1/11/2026 10:52:41 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Industrial Chemistry & Materials

Chemical Science

MSDE

Reaction Chemistry & Engineering

PCCP

Open Access Article. Published on 16 January 2025. Downloaded on 1/11/2026 10:52:41 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Digital Discovery

Open Access Article. Published on 16 January 2025. Downloaded on 1/11/2026 10:52:41 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

104 **Large property models: a new generative machine-learning formulation for molecules**
Tianfan Jin, Veerupaksh Singla, Hsuan-Hao Hsu and Brett M. Savoie

120 **Data-efficient fine-tuning of foundational models for first-principles quality sublimation enthalpies**
Harveen Kaur, Flaviano Della Pia, Ilyes Batatia, Xavier R. Advincula, Benjamin X. Shi, Jinggang Lan, Gábor Csányi, Angelos Michaelides and Venkat Kapil

139 **Knowledge distillation of neural network potential for molecular crystals**
Takuya Taniguchi

156 **Modelling ligand exchange in metal complexes with machine learning potentials**
Veronika Juraskova, Gers Tusha, Hanwen Zhang, Lars V. Schäfer and Fernanda Duarte

177 **Discovering chemical structure: general discussion**

221 **Web-BO: towards increased accessibility of Bayesian optimisation (BO) for chemistry**
Austin M. Mroz, Piotr N. Toka, Ehecatl Antonio del Río Chanona and Kim E. Jelfs

235 **Sequence determinants of protein phase separation and recognition by protein phase-separated condensates through molecular dynamics and active learning**
Arya Changiarath, Aayush Arya, Vasileios A. Xenidis, Jan Padeken and Lukas S. Stelzl

255 **Discovery of highly anisotropic dielectric crystals with equivariant graph neural networks**
Yuchen Lou and Alex M. Ganose

275 **Leveraging natural language processing to curate the tmCAT, tmPHOTO, tmBIO, and tmSCO datasets of functional transition metal complexes**
Ilia Kevlishvili, Roland G. St. Michel, Aaron G. Garrison, Jacob W. Toney, Husain Adamji, Haojun Jia, Yuriy Román-Leshkov and Heather J. Kulik

304 **Are we fitting data or noise? Analysing the predictive power of commonly used datasets in drug-, materials-, and molecular-discovery**
Daniel Crusius, Flaviu Cipcigan and Philip C. Biggin

322 **Prediction rigidities for data-driven chemistry**
Sanggyu Chong, Filippo Bigi, Federico Grasselli, Philip Loche, Matthias Kellner and Michele Ceriotti

345 **Accurate and reliable thermochemistry by data analysis of complex thermochemical networks using Active Thermochemical Tables: the case of glycine thermochemistry**
Branko Ruscic and David H. Bross

373 **Discovering structure–property correlations: general discussion**

413 **Specialising and analysing instruction-tuned and byte-level language models for organic reaction prediction**
Jiayun Pang and Ivan Vulić

434 **Predictive crystallography at scale: mapping, validating, and learning from 1000 crystal energy landscapes**
Christopher R. Taylor, Patrick W. V. Butler and Graeme M. Day

 This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 16 January 2025. Downloaded on 1/11/2026 10:52:41 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

459 **Optical materials discovery and design with federated databases and machine learning**
Victor Trinquet, Matthew L. Evans, Cameron J. Hargreaves, Pierre-Paul De Breuck and Gian-Marco Rignanese

483 **How big is big data?**
Daniel Speckhard, Tim Bechtel, Luca M. Ghiringhelli, Martin Kuban, Santiago Rigamonti and Claudia Draxl

503 **Making the InChI FAIR and sustainable while moving to inorganics**
Gerd Blanke, Jan Brammer, Djordje Baljozovic, Nauman Ullah Khan, Frank Lange, Felix Bänsch, Clare A. Tovee, Ulrich Schatzschneider, Richard M. Hartshorn and Sonja Herres-Pawlis

520 **Discovering trends in big data: general discussion**

551 **Analysis of uncertainty of neural fingerprint-based models**
Christian W. Feldmann, Jochen Sieg and Miriam Mathea

568 **Re-evaluating retrosynthesis algorithms with Syntheseus**
Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl and Marwin H. S. Segler

587 **Embedding human knowledge in material screening pipeline as filters to identify novel synthesizable inorganic materials**
Basita Das, Kangyu Ji, Fang Sheng, Kyle M. McCall and Tonio Buonassisi

601 **Mapping inorganic crystal chemical space**
Hyunsoo Park, Anthony Onwuli, Keith T. Butler and Aron Walsh

614 **A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes**
Wenhao Sun and Nicholas David

639 **Discovering synthesis targets: general discussion**

CONCLUDING REMARKS

664 **Concluding remarks: *Faraday Discussion* on data-driven discovery in the chemical sciences**
Andrew I. Cooper

ADDITIONAL INFORMATION

691 **Poster titles**

696 **List of participants**