5th Machine Learning and AI in Bio(Chemical) Engineering

Conference booklet

June 27-28 2022

Hosted by

UNIVERSITY OF CAMBRIDGE

Conference Sessions

27 June 2022

10:00 - 11:00	Welcome and refreshments
11:00 - 12:00	Keynote 1 – Connor Coley - *AI for chemical space navigation and synthesis*
12:00 - 12:25	Adarsh Arun – *Reaction impurity prediction using a data mining approach*
12:25 - 12:50	A Kondinski – *Automated rational design via knowledge engineering*
12:50 - 14:00	Lunch
14:00 - 14:35	Ruben Sanchez-Garcia - *Compound availability and the numbers we care about in computationally-driven drug discovery*
14:35 - 15:00	Harry Kay - *Developing a novel soft-sensing framework for industrial data analysis and batch process monitoring*
15:00 - 15:25	Calvin Tsay - *SnAKe: Bayesian Optimization with Pathwise Exploration*
15:25 - 15:35	Break
15:35 - 16:10	François-Xavier Felpin - *Autonomous Flow reactors Associating In-line/Online Analyses and Feedback Algorithms*
16:10 - 16:35	Ioana Gherman - *Accelerating whole cell modelling with machine learning*
16:35 - 17:00	Haiting Wang – *A Hybrid Modelling Framework for Bioprocess*
17:00 - 17:25	Pierre-Aurelien Gillot - *Systemic comparison of neural network architectures for protein expression prediction in bacteria*
17:25 - 19:15	Networking and dinner
19:15	Day 1 end

28 June 2022

09:00 - 09:15	Coffee reception
09:15 - 10:00	Workshop part 1 - *Scaling up & scaling out compound generation and simulation with Grid.ai*
10:0 - 10:15	Break
10:15 - 11:00	Workshop part 2
11:00 – 12:00	Poster Session
12:00 - 12:35	Carl Poelking - *Strategies for bias compensation and synthetic control in AI-driven structure-based drug design*
12:35 - 13:00	Miguel Angel de Carvalho Servia - *Automated Kinetic rate equation discovery- A methodological framework*
13:00 – 14:00	Lunch
14:00 - 15:00	Keynote 2 – Kerry Gilmore
15:00 - 15:25	Felix Strieth-Kalthoff - *Closing the Loop in Materials Discovery: The Quest for Organic Lasers*
15:25 – 15:45	Break
15:45 - 16:10	Venkat Kapil - *The first-principles diagram of monolayer nanoconfined water*
16:10 – 16:35	Abhishek Sharma - *AI-EDISON: Autonomous Intelligent Exploration, DIScovery and Optimisation of Nanomaterials*
16:35 - 17:00	Closing Remarks
17:00	End of day 2

All reported times in BST.

* Q&A sessions refer to pre-recorded talks

Keynote Speakers

Connor Coley

Connor W. Coley is an Assistant Professor at MIT in the Department of Chemical Engineering and the Department of Electrical Engineering and Computer Science. He received his B.S. and Ph.D. in Chemical Engineering from Caltech and MIT, respectively, and did his postdoctoral training at the Broad Institute. His research group at MIT develops new methods at the intersection of data science, chemistry, and laboratory automation to streamline discovery in the chemical sciences with an emphasis on therapeutic discovery. Key research areas in the group include the design of new neural models for representation learning on molecules, data-driven synthesis planning, in silico strategies for predicting the outcomes of organic reactions, model-guided Bayesian optimization, and de novo molecular generation. Connor is a recipient of C&EN's "Talented Twelve" award, Forbes Magazine's "30 Under 30" for Healthcare, the NSF CAREER award, and the Bayer Early Excellence in Science Award. Outside of MIT, Connor serves as an advisor to both early- and late-stage companies including Entos, Revela, Galixir, Kebotix, Anagenex, and Dow.

Kerry Gilmore

Kerry Gilmore is an assistant professor at the University of Connecticut. His group develops and applies new technologies and ideas within flow chemistry, computational chemistry and machine learning to solve problems in organic chemistry.

Invited Speakers

François-Xavier Felpin

Professor François-Xavier Felpin leads the Sustainable Chemistry and New Technologies (SCNT) research group at Université de Nantes. The group works at the interfaces of catalysis, (macro)molecular chemistry and chemical engineering, in areas as diverse as homogeneous and heterogeneous palladium catalysis, cellulose modifications and smart flow chemistry.

Carl Poelking

Dr Carl Poelking is a Post-doctoral research associate at the University of Cambridge. His research focusses on data-driven modelling approaches for molecular simulation and materials discovery.

Ruben Sanchez-Garcia

Ruben Sanchez-Garcia is a postdoctoral fellow in the Department of Statistics at the University of Oxford. His work focusses on the application of artificial intelligence and deep learning for drug discovery.

Oral talks

Reaction impurity prediction using a data mining approach

Adarsh Arun[1,3], Zhen Guo[2,3] , Simon Sung[3], Alexei Lapkin[1,2,3]*
* Corresponding author

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, UK
[2] Chemical Data Intelligence (CDI) Pte Ltd, Robinson Road, #02-00, 068898 Singapore
[3] Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore

Keywords: Impurity prediction, data mining

Abstract (<400 words):

Automated prediction of reaction impurities can be useful in facilitating rapid early-stage reaction development, synthesis planning and optimization. Existing reaction predictors are catered towards main product prediction, (1,2) and are often black-box, making it difficult to troubleshoot erroneous outcomes. This work aims to present an automated impurity prediction workflow that is interpretable and transparent, as it is based on data mining large chemical reaction databases. A 14-step workflow was implemented in Python and RDKit using Reaxys data.(3) Evaluation of potential chemical reactions between functional groups (4) present in the same reaction environment in the user-supplied query species can be accurately performed by directly mining the Reaxys database for similar or 'analogue' reactions involving these functional groups. Reaction templates can then be extracted from analogue reactions and applied to the relevant species in the original query to return impurities and transformations of interest. Three proof-of-concept case studies based on active pharmaceutical ingredients (paracetamol, agomelatine and lersivirine) were conducted, with the workflow able to suggest the correct impurities within the top two outcomes. At all stages, suggested impurities can be traced back to the originating template and analogue reaction in the literature, allowing for closer inspection and user validation. Ultimately, this work could be useful as a benchmark for more sophisticated algorithms or models since it is interpretable, as opposed to purely black-box solutions, and illustrates the potential of chemical data in impurity prediction.

References

1. Johansson S, Thakkar A, Kogej T, Bjerrum E, Genheden S, Bastys T, et al. AI-

assisted synthesis prediction. Drug Discov Today Technol. 2020;32–33:65–72.

2. Thakkar A, Johansson S, Jorner K, Buttar D, Reymond JL, Engkvist O. Artificial intelligence and automation in computer aided synthesis planning. React Chem Eng. 2021;6(1):27–51.
3. Reaxys - An expert-curated chemistry database [Internet]. [cited 2021 Sep 30]. Available from: https://www.elsevier.com/solutions/reaxys
4. Ertl P. An algorithm to identify functional groups in organic molecules. J Cheminform. 2017;9(1):1–7.

Automated Rational Design via Knowledge Engineering

A. Kondinski[1], A. Menon[1], D. Nurkowski[2], F. Farazi[1], S. Mosbach[1], J. Akroyd[1], M. Kraft*[1,2,3,4,5]

* Corresponding author

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK
[2] CMCL Innovations, Cambridge, UK
[3] CARES Cambridge Centre, Singapore
[4] School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore
[5] The Alan Turing Institute, London, UK

Keywords: Rational Design, Knowledge Engineering, Metal-Organic Polyhedra.

Abstract:

Rational design (RD) is a form of design thinking that enables reliable molecular engineering of complex materials with desired properties. Emulation of RD with the help of artificial intelligence is currently highly desirable, as it enables the facile development of autonomous systems for chemical discovery. However, as RD is a cognitively complex process its direct emulation remains challenging. Inspired by our previous work in the development of didactical tools and digitally interoperable knowledge-based systems, we initially hypothesized that knowledge engineering (KE) may be best suited for the emulation of RD. In this talk we first showcase the different components of a KE system, that is an ontology, instances, and a software agent and how they can be tailored to address advanced molecular systems such as the hybrid organic-inorganic molecules called metal-organic polyhedra (MOPs). The KE approach is then successfully used for RD of new MOP instances in an evidence-based manner, which essentially demonstrates its utility for the automation of RD processes.[1]

References

1. A. Kondinski, A. Menon, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd, M. Kraft Technical Report 292, 2022, c4e-Preprint Series, Cambridge.

Developing a novel soft-sensing framework for industrial data analysis and batch process monitoring

Sam Kay‡1, Harry Kay‡1, Max Mowbray1, Amanda Lane2, Philip Martin1, Dongda Zhang*1

[1] Department of Chemical Engineering and Analytical Science, University of Manchester, Oxford Road, Manchester, M1 3AL, UK.
[2] Unilever Research Port Sunlight, Quarry Rd East, Bebington, C63 3JW, UK.
‡*: These authors contributed equally to this work.*
*:dongda.zhang@manchester.ac.uk

Keywords: Machine learning, data analytics, dimensionality reduction, quality control, uncertainty estimation.

Abstract:

Viscosity represents a key indicator of formulated product quality but has traditionally been difficult to measure in-process in real-time. This is particularly true if the process involves complex mixing and reaction phenomena operated at dynamic conditions. To address this challenge, a promising solution to monitoring product viscosity is to design soft-sensors which correlate viscosity with easily measured process variables. In this study, we developed an innovative machine learning based soft-sensor construction framework by integrating different types of advanced artificial neural networks. The framework first employs a dimensionality reduction technique to generate information-rich statistic latent variables by compressing high-dimensional industrial data, and then adopts a novel heteroscedastic noise neural network (HNN) to simultaneously predict product viscosity and its associated uncertainty based on the extracted latent features. Specifically, to guarantee extraction of key process information, this study investigated two dimensionality reduction techniques, namely partial least squares (PLS) and a deep learning autoencoder, succeeded by a fully comprehensive analysis and comparison of the performance of the respective soft-sensors. To evaluate the accuracy and robustness, the data-driven soft-sensors were used to predict product viscosity for a number of industrial batches operated over different seasons and product variants. It is found that the soft-sensors constructed using both dimensionality reduction techniques have both high accuracy (prediction error <12%) and high reliability (predicted uncertainty similar to the measurement uncertainty within the factory) in most cases, indicating their great potential for industrial batch process monitoring and quality control.

A Hybrid Modelling Framework for Bioprocess

Haiting Wang [1], Cleo Kontoravdi [1], Ehecatl Antonio del Rio Chanona *[1]

[1] Imperial College London, Exhibition Rd, South Kensington, London, SW7 2BX, United Kingdom.
{haiting.wang19, cleo.kontoravdi98, a.del-rio-chanona}@imperial.ac.uk

Keywords: Hybrid modelling, Bioprocess modelling, Model selection.

Abstract (<400 words):

Industrial bioprocesses have experienced a fast development in many fields such as renewable energy generation, wastewater treatment and pharmaceutical manufacturing. However, due to the limited knowledge of the biological mechanisms of living cells, most bioprocesses are still limited to lab-scale, and experiments are expensive and time-consuming. Alternatively, mathematical modelling of bioprocesses is an effective method to computationally simulate the bioprocess behaviours. Mechanistic models are commonly designed based on prior knowledge of the biosystem. However, expressing all potential mechanisms mathematically can be heavy work and simulating unclear mechanisms is tough. Instead, we can use data-driven models based on machine learning without mechanistic knowledge. However, machine learning models do not extrapolate, and this is particularly damaging when we wish to predict the behaviour in new areas, to optimise the biosystems, and particularly in low data regimes. Hybrid models can be considered an effective modelling alternative, and "the best of the two worlds". A well-built hybrid model can provide fast and reliable simulation performance with good extrapolation capabilities [1]. The problem is, however, how can we build a robust and reliable hybrid model?

In this work, we proposed a hybrid model construction methodology as shown in Figure 1. The mechanistic backbone is designed based on prior knowledge of biological mechanisms and process engineering. At the same time, data-driven methods can be incorporated to represent the unknown biomechanisms efficiently. This methodology is based on asymptotically complete global optimisation algorithms for the parameter estimation step, and branch and bound and genetic algorithms for the model complexity selection, depending on the complexity of the models. To avoid overfitting problems, different statistically-based model selection methods are used to choose the best model structure to balance the fitting and extrapolation performance of the hybrid model.

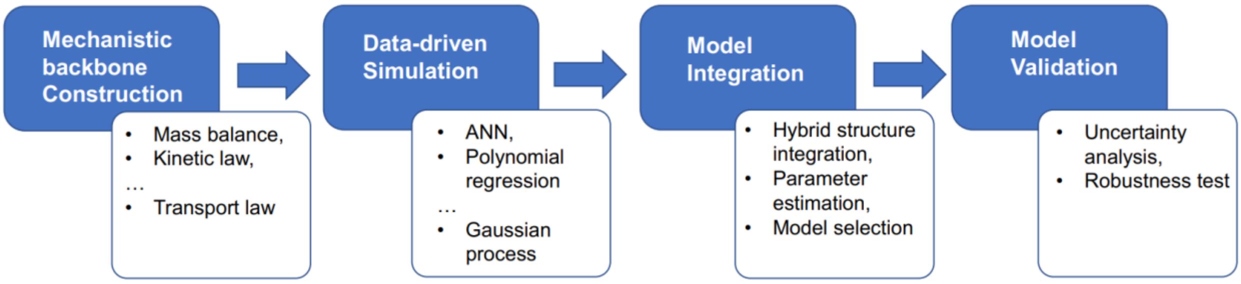

Figure 1 Illustration of the proposed hybrid model construction methodology.

Then the proposed methodology can be tested in a case study for the simulation of the microalgae cultivation process. The mechanistic structure of the model is designed based on the mass balance of state variables. The influence of light attenuation, substrate and nitrate consumption on the microalgae growth rate is simulated through data-driven methods including polynomial regression and Artificial Neural Networks. The data-driven model structure is selected through statistic methods such as Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and Hannah Quinn Criterion (HQC). The model is further validated to provide robust extrapolation ability.

References

1. Von Stosch M, Oliveira R, Peres J, de Azevedo SF. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Computers & Chemical Engineering. 2014;60:86-101.

Systematic comparison of neural network architectures for protein expression prediction in bacteria

Pierre-Aurelien Gilliot[1],* and Thomas E. Gorochowski[1]
[1]School of Biological Sciences, University of Bristol, Bristol, UK

Keywords: gene expression, neural networks, sequence design.

Abstract:

Protein expression in bacteria is a tightly regulated phenomenon whose activity can span up to five orders of magnitude across the genome. Achieving precise control over protein levels is important for many bioengineering applications, for example, allowing us to maximize the synthesis of a high value product or to control fluxes in a metabolic pathway. Recent studies have investigated the determinants of protein expression and have shown translation initiation to often be the rate-limiting step. During translation initiation, the ribosome interacts with the non-coding start of the messenger RNA within the 5' untranslated region (UTR). Mutations in this region can have a large impact on protein expression, but finding the appropriate mutations for a particular application usually involves lengthy experiments. To circumvent these experiments, predictive models of translation initiation have been developed, but their performance is generally poor. Recent high-throughput experiments linking 5'-UTR mutations to protein expression activities now provide researchers with a trove of data to improve on existing models and pave the way for statistical learning. Here we show the superior performance of neural networks over traditional biophysical models to predict protein expression from sequence alone. By systematically comparing different neural networks operating on various representations of the 5'-UTR sequence, we show a 30% improvement in accuracy over state-of-the-art methods. We demonstrate how the best architecture can also be fine-tuned to a different sequence context using only a few additional experimental measurements. We anticipate our algorithm, which is publicly available (https://gitlab.com/Pierre-Aurelien/rebeca), will accelerate the genetic sequence design workflow and help deepen our understanding of how gene expression is regulated.

Accelerating whole cell modelling with machine learning

Ioana Gherman*[1], Zahraa Abdallah[1], Wei Pang[3], Thomas E. Gorochowski[2], Claire S. Grierson[2], Lucia Marucci[1]

* Corresponding author

[1] Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

[2] School of Biological Science, University of Bristol, Bristol, United Kingdom

[3] School of Mathematical and Computer Science, Heriot Watt University, Edinburgh, United Kingdom

Keywords: whole cell modeling, surrogate models, machine learning, bioengineering

Abstract:

Whole cell models are mathematical models designed to capture the function of all genes and core processes within a cell. Developing whole cell models is seen as a grand challenge of the 21st century [1] and although explored for over a decade, only two partially complete models have been published to date, for bacteria *Mycoplasma genitalium* [2] and *Escherichia coli* [3]. The interest in whole cell models stems from their ability to provide an integrated picture of diverse processes within a cell, uncover novel cellular phenotypes, and understand the behaviour of engineered cells (e.g. containing new metabolic pathways or having genes knocked out) for biotechnology purposes (e.g. bioproduction) [4,5,6]. Despite their value, whole cell models also bring some challenges, with the most pressing of these being the huge computational demand of the simulations. To simulate the life cycle of a single *Mycoplasma genitalium* cell (one of the simplest organisms we know of), using the most comprehensive model to date, takes up to 24 hours on a typical desktop computer. This makes it difficult to run the tens of thousands of simulations required for understanding the effect of changes to a cell and to engineer applications like genome design, where we attempt to augment or alter core functionalities of the cell. Here, we aim to address this challenge by building a 'surrogate' of a whole cell model that uses machine learning algorithms to accelerate the speed of simulations. Surrogates also referred to as emulators or metamodels, represent an alternative representation of a full mathematical model. They are usually trained/fitted using simulation data from the full system model, and once the surrogate's performance closely matches the original model, it can be used in its place to accelerate future simulations [7]. We explore the applicability and usage of machine learning surrogates in the context of whole cell models and demonstrate how they can both speed-up simulations in specific circumstances and be used to uncover interesting dynamics of cellular phenotypes, that would be nearly impossible to assess with current experimental methods. Surrogate models may hold the key to making whole cell modelling practical for studying cellular biology and bioengineering on a typical desktop computer and help improve accessibility of this powerful modelling technique.

References

1. Tomita M. Whole-cell simulation: a grand challenge of the 21st century. TRENDS in Biotechnology. 2001 Jun 1;19(6):205-10.

2. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, Assad-Garcia N, Glass JI, Covert MW. A whole-cell computational model predicts phenotype from genotype. Cell. 2012 Jul 20;150(2):389-401.

3. Macklin DN, Ahn-Horst TA, Choi H, Ruggero NA, Carrera J, Mason JC, Sun G, Agmon E, DeFelice MM, Maayan I, Lane K. Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. Science. 2020 Jul 24;369(6502):eaav3751.
4. Carrera J, Covert MW. Why build whole-cell models?. Trends in cell biology. 2015 Dec 1;25(12):719-22.
5. Landon S, Chalkley O, Breese G, Grierson C, Marucci L. Understanding Metabolic Flux Behaviour in Whole-Cell Model Output. Frontiers in molecular biosciences. 2021;8.
6. Rees-Garbutt J, Chalkley O, Landon S, Purcell O, Marucci L, Grierson C. Designing minimal genomes using whole-cell models. Nature communications. 2020 Feb 11;11(1):1-2.
7. Asher MJ, Croke BF, Jakeman AJ, Peeters LJ. A review of surrogate models and their application to groundwater modeling. Water Resources Research. 2015 Aug;51(8):5957-73.

Automated Kinetic Rate Equation Discovery – A Methodological Framework

Miguel Ángel de Carvalho Servia[1], Dongda Zhang[2], Klaus Hellgardt[1], King Kuok (Mimi) Hii[1], Ehecatl Antonio del Rio Chanona[1, *]

[*] Corresponding author

[1] Imperial College London, United Kingdom.

[2] The University of Manchester, United Kingdom.

Keywords: kinetic rate discovery, symbolic regression, model selection criteria

Abstract:

Accurate, predictive, and interpretable mathematical models are important from a theoretical and practical point of view. Theoretically, these models allow engineers to gain fundamental understanding of physical phenomena. Practically, these models allow engineers to optimize, control and even develop novel processes. Nevertheless, the automated discovery of true kinetic rate models remains an open challenge within the chemical engineering community. This challenge has many ramifications within the industrial world, ranging from sub-optimal control to difficulties within the design and upscaling of chemical processes.

Different modelling techniques have been proposed and explored in the literature: white-box modelling, grey-box modelling, and black-box modelling. The grey-box modelling technique exploits the advantages of white-box modelling, namely its predictive ability, and the advantages of black-box modelling, namely its ease of construction. However, most hybrid models presented in the literature make undetermined assumptions about the chemical system investigated (e.g.: assuming kinetic formalisms) and do not include a rigorous model selection method. While these assumptions hinder the accuracy and predictability of the model proposed, the absence of a rigorous model selection method limits the capabilities of finding the underlying ground truth of the system.

Due these limitations within the model building framework, we have developed a method that tackles them. Our method uses minimal – but important and physically-driven – prior knowledge to guide a symbolic regression algorithm to propose competing kinetic rate equations for a given chemical system. Then, using carefully analyzed model selection criteria and model-based design of experiments, we can robustly identify and choose the model that accurately describes the system's kinetics while providing limited, but highly informative data.

To strengthen our approach, we benchmarked a plethora of model selection criteria on different case studies, whilst varying the quantity and information content of the data provided. We also assessed the level of noise that each criterion was able to withstand until it started selecting wrong models. Our objective was to discover which criterion, if any, was better suited for the kinetic rate discovery task, and investigate which criterion was the most robust. Our study demonstrated that, from the criteria examined, the Hannan and Quinn criterion is the most robust and well-suited for the problem class at hand. In conclusion, our meticulous choice of model selection criterion integrated within our proposed methodological framework

maximizes the probability of the true kinetic rate model being retrieved from the data used, proving the essential role of a rigorous model selection method.

Automated Kinetic Rate Equation Discovery – A Methodological Framework

[1]Felix Strieth-Kalthoff

[1] University of Toronto

Abstract:

Augmenting automated experimentation with artificial intelligence has emerged as the next-generation paradigm to streamline and accelerate materials discovery workflows. In this "self-driving laboratory" (SDL) framework, the processes of materials design, compound preparation, and property optimization are automated in order to close the traditional "design–make–test" loop and enable autonomous experimentation.

In this talk, I will outline our recent efforts towards such a SDL for discovering novel gain materials for organic solid-state lasers (OSL). Given the omnipresence of lasers and their numerous technological applications, OSLs have attracted significant attention owing to distinct advantages regarding cost, color tunability or device fabrication. At the same time, the discovery of gain materials for OSLs – usually highly conjugated organic molecules – has been hampered by a range of decomposition and excited-state deactivation processes.

Our SDL attempts to streamline this discovery process by "closing the loop" of designing suitable candidates for OSL gain materials, synthesizing them, and measuring their lasing properties: In a cloud-centered platform for data storage and experiment design, a Bayesian optimization algorithm suggests suitable molecular structures using a fragment-based approach. The AI-proposed molecules are then "downloaded" and synthesized in the laboratory, assembling the molecular building blocks using an iterative cross-coupling strategy. Automated reaction analysis, coupled to in-line characterization of optical materials properties, allows for experimental execution in a fully integrated end-to-end workflow. Feeding the obtained data back to the cloud, the Bayesian optimizer is refined to propose next-generation candidate molecules, enabling autonomous iterative optimization.

Throughout the talk, I will discuss the implementation of this SDL, along with its evolution into a platform for asynchronous, delocalized optimization campaigns, where experimentation is distributed over multiple sites and instruments. Eventually, I will outline how these collaborative efforts have enabled the discovery of early-generation leads – on the way to better OSL gain materials.

AI-EDISON: Autonomous Intelligent Exploration, DIScovery, and Optimisation of Nanomaterials

Abhishek Sharma[1], Yibin Jiang[1], Daniel Salley[1], Leroy Cronin[1]*

*Corresponding author email: Lee.Cronin@glasgow.ac.uk

[1] School of Chemistry, University of Glasgow, University Avenue, Glasgow G128QQ, UK

Keywords: nanomaterials, chemical automation, closed-loop, morphological control, exploration-optimization

Abstract:

The design and development of nanomaterials have been extensively explored due to their unique properties and wide range of applications in medical science, sensing, electronic devices, catalysis, and energy storage. Their unique properties can be controlled by fine-tuning the morphological features; however, the synthetic procedures often suffer from lower yields and irreproducibility. These problems often emerge due to extreme sensitivity to high-dimensional experimental conditions such as concentrations, mixing rates, temperature etc. To overcome these problems, here we present an AI-enabled closed-loop nanomaterial synthesis platform (AI-EDISON) which employs automated synthesis, real-time characterization, and theory within the feedback loop driven by machine learning algorithms [1]. We designed and built a modular robotic architecture that performs multi-step automated seed-mediated synthesis of gold nanoparticles (AuNPs) together with inline characterization using UV-Vis spectroscopy.

Based on the hypothesis that exploring diversity in UV-Vis spectra could lead to the exploration of nanostructures with unique morphological properties, we performed the exploration of multiple synthetic spaces of AuNPs using the MAP-Elites algorithm. In the closed-loop approach, the observational space was discretized into finite intervals where sampling points with the highest fitness (elites) were used to create a new set of experiments via mutation, crossover, and random sampling. We were able to discover distinct AuNPs with various morphologies such as spheres, rods, polyhedral, bicones, and stars with a higher yield. After exploration, AI-EDISON was used to optimize the synthetic conditions towards a pre-defined target generated by using a fast GPU-accelerated scattering simulation engine on the nanoparticle shapes generated from the electron micrographs. Due to the non-uniqueness of UV-Vis towards a specific structure, optimization was performed utilizing a global search algorithm with local sparseness to find multiple synthetic conditions towards the target spectra achieving significant improvement in yield and monodispersity.

Additionally, we have demonstrated the platform's capability to perform parallel and reproducible synthesis, by setting pre-defined multistep synthetic targets utilizing a Chemical Description Language (χDL)[2]. We believe AI-EDISON's closed-loop methodology to perform experiments offers a viable way toward efficient nanomaterial discovery, optimization, and reproducible synthesis.

References

1. Jiang Y, Salley D, Sharma A, Keenan G, Mullin M, Cronin L. An Artificial Intelligence Enabled Chemical Synthesis Robot for Exploration and Optimisation of Nanomaterials. 2022 (under review)

2. Mehr SHM, Craven M, Leonov AI, Keenan G, Cronin L. A universal system for digitization and automatic execution of the chemical synthesis literature. Science (1979). 2020 Oct 2;370(6512):101–8.

The first-principles phase diagram of monolayer nanoconfined water

Venkat Kapil*[1], Christoph Schran[1], Andrea Zen[2], Ji Chen[3], Chris Pickard[1], Angelos Michaelides[1]

* Corresponding author
[1] University of Cambridge, UK
[2] Universitá di Napoli Federico II, Italy
[3] Peking University, Beijing, China

Keywords: Machine Learning, Statistical Mechanics, Electronic Structure, Nanoconfinement, Phase diagrams

Abstract:

Water in nanoscale cavities is ubiquitous and of central importance to everyday phenomena ingeology and biology, and at the heart of current and future technologies in nanoscience. A molecular-level picture of the structure and dynamics of nanoconfined water is a prerequisite to understanding and controlling the behavior of water under confinement. Here we explore a monolayer of water confined within a graphene-like channel using a framework that combines developments in high-level electronic structure theory, machine learning, and statistical sampling. This approach enables a treatment of nanoconfined water at unprecedented accuracy. We find that monolayer water exhibits surprisingly rich and diverse phase behaviour that is highly sensitive to temperature and the van der Waals pressure acting within the nanochannel. Monolayer water exhibits numerous molecular ice phases with melting temperatures that vary by over 400 degrees in a non-monotonic manner with pressure. In addition, we predict two unexpected phases: a ”hexatic-like” phase, which is an intermediate between a solid and a liquid, and a superionic phase with a high electrical conductivity exceeding that of battery materials. Our work suggests that nanoconfinement could be a promising route towards superionic behavior at easily accessible conditions.

References

1. Venkat Kapil, Christoph Schran, Andrea Zen, Ji Chen, Chris J. Pickard, Angelos Michaelides. The first-principles phase diagram of monolayer nanoconfined water. Nature. 2022 (accepted) arXiv:2110.14569

FLab – a Fast, Flexible and Fun coding platform for democratizing artificial intelligence-accelerated laboratories

Nicholas A. Jose[1-3] *, Alexei A. Lapkin[1-4]

* Corresponding author

[1] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Dr, Cambridge UK CB3 0AS

[2] Cambridge Centre for Advanced Research and Education in Singapore, 1 CREATE Way, Singapore 138602

[3] Accelerated Materials Ltd. 71-75 Shelton St, Covent Garden, London, UK WC2H 9JQ

[4] Innovation Centre in Digital Molecular Technologies, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK CB2 1EW

Keywords: Automation, AI, IOT, Laboratory, Education

Abstract:

Applying artifical intelligence to create self driving chemical laboratories may deliver orders of magnitude enhancements in research productivity. However, implementation of artificial intelligence into realistic use cases often encounters a number of severe hurdles: chemical scientists need to obtain advanced skills in automation/coding, commercial platforms are costly, and suppliers of automated lab devices often "lock" a users into supplier automation workflows. These hurdles limit laboratories across the world from reaching the critical mass required for widespread adoption of "smart" laboratories.

In the pursuit of solving these challenges, and democratising smart laboratories, we have developed FLab,[1] an open-source Python coding framework for linking automation, IOT technologies, and artificial intelligence in chemical laboratories. FLab utilises an intuitive, modular, object-oriented architecture to streamline manipulation of shared devices, tasks, AI bots and user interfaces. In this presentation, we describe the inner workings of Flab, and its differentiation from existing frameworks like Labview, ROS and Matlab. We then illustrate its uses in real laboratory experiments, which include a bespoke automated flow chemistry scale-up rig, a high-throughput batch formulation robot, and a system for measuring hydrodynamic parameters in a gas/liquid flow reactor. Finally, we discuss critical aspects for improvement within FLab and within self-driving labs, pointing to new avenues of further research and collaboration in this growing field.

References

1. https://pypi.org/project/flab/

Poster Session

Presenter	Poster title
Cai Y Ma	Machine Learning, Imaging and Image Processing for 3D Crystal Shape Characterisation
Dongda Zhang, Alexander Rogers	Comparing different hybrid modelling approaches for bioprocess predictive modelling and uncertainty propagation
Andrea Friso	Optimal design of experiment for model structure identification coupling MBDoE techniques and RL methodologies
Kobi Felton	
Dongda Zhang	Safe Chance Constrained Reinforcement Learning for Batch Process Control
Tania Mahmood	Using Molecular dynamics simulations to unpin interactions occurring under high concentration mAb formulations
Charles Gong	Evaluating and interpreting uncertainty in QSAR models
Marcus Wang	Using automated machine learning for the prediction of developmental and reproductive toxicity
Srijit Seal	Biological Interpretation of Cell Painting and Gene Expression Features for Mitochondrial Toxicity Prediction
Elena Gelzinyte	ML force fields for open- and closed-shell organic molecules
Yuhan Wang	Using molecular dynamics simulation to predict the aggregation propensity of monoclonal antibodies formulations & accelerate development
Austin Tripp	Meta-learning Adaptive Deep Kernel Gaussian Processes for Molecular Property Prediction
Ryan-Rhys Griffiths	GAUCHE: A Library for Gaussian Processes and Bayesian Optimisation in Chemistry
Niccolo Veanzi	Predicting protein properties using molecular dynamics and machine learning
Abbey/Michael	Fluidic Neural Networks for Silver Nanoparticle Synthesis in Helical Microreactors

Sponsors

Reaction Chemistry
& Engineering

Student Poster Session Sponsor

Digital
Discovery

Early Career Research Talks Sponsor