What Can Digitization Do For Formulated Product Innovation and Development

James L. McDonagh[†*], William C. Swope[⋆], Richard L. Anderson[♮], Michael A. Johnson[‡], and David J. Bray[♮*]

[†]IBM Research U.K., Hartree Centre, Daresbury WA4 4AD, United Kingdom
[⋆]IBM Almaden Research Center, San Jose, California, United States
[‡]IBM Research Ireland, Dublin, Ireland
[♮]The Hartree Centre, STFC Daresbury Laboratory, Warrington, WA4 4AD, United Kingdom
[*]Corresponding authors

November 2019

1 Abstract

Digitization offers significant opportunities for the formulated product industry to transform the way it works and develop new methods of business. R&D is one area of operation that is challenging to take advantage of these technologies due to its high level of domain specialisation and creativity but the benefits could be significant. Recent developments of base level technologies such as artificial intelligence (AI)/machine learning (ML), robotics and high performance computing (HPC), to name a few, present disruptive and transformative technologies which could offer new insights, discovery methods and enhanced chemical control when combined in a digital ecosystem of connectivity, distributive services and decentralisation. At the fundamental level, research in these technologies has shown that new physical and chemical insights can be gained, which in turn can augment experimental R&D approaches through physics-based chemical simulation, data driven models and hybrid approaches. In all of these cases, high quality data is required to build and validate models in addition to the skills and expertise to exploit such methods. In this article we give an overview of some of the digital technology demonstrators we have developed for formulated product R&D. We discuss the challenges in building and deploying these demonstrators.

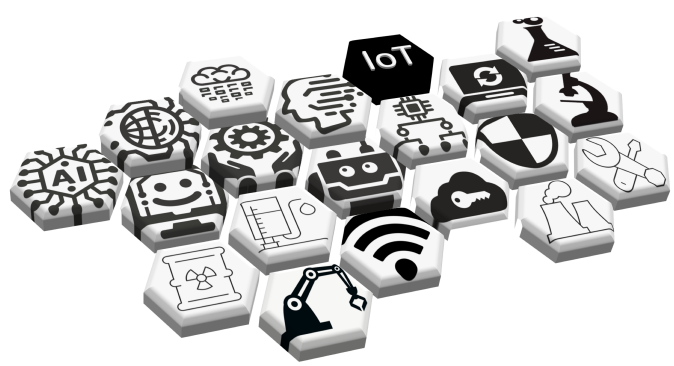

Figure 1: Idea for graphical abstract image

2 Introduction

Digitization is disrupting and transforming the way the commercial world works through new services and the automation of common tasks.[1, 2, 3] These digital services fall under the term *Industry 4.0* (I4.0). Such services are enabling democratization of knowledge, automation of common processes and remote monitoring, among many other applications. One advantage is that the required physical and computational infrastructure is distributed rather than co-located such that users can take advantage of multiple services, cloud storage and computing. This has obvious benefits for an organization as it enables decentralisation of staff and equipment,

facilitating wider use of expertise and up-skilling *via* the digitization of expertise into software. In practice this means that specialized skills can be shared throughout an organization without the need for all users to hold deep expertise.

Whilst I4.0 and other digital technologies have been pervasive in the formulated products industry[4] implementation and adoption of digital approaches is much more challenging in a product R&D setting.[5] This is in part due to high level of domain specialisation and creativity, which is not easy to capture in generic digital applications.[6] In addition R&D is often more poorly funded and of lower priority for an organisation than other departments (e.g., marketing). In some cases, R&D processes have remained largely unchanged for many years and have performed well, hence digital changes have not been forthcoming. Nevertheless, digital approaches to R&D are being adopted by the large players in the formulated products field[7] especially where product re-formulation occurs on a rapid timescale to respond to ever changing demands.

As changing regulatory frameworks [8] and societal [9] pressure move formulated products towards more sustainable and diverse feed stocks, materials and processes, existing formulation R&D methods will need to be augmented if organisations wish to remain ahead of the competition. A significant proportion of existing formulated products are based on technology developed over many decades. These new regulatory and societal pressures are likely to lead companies to explore novel and unfamiliar chemicals where long standing knowledge isn't available. For example, the growth in demand for novel battery technologies [10, 11, 12]. Against this back drop, we expect formulated product companies to engage with digital approaches to augment and accelerate their R&D programmes. In an era of Big Data, robotics, AI and High Performance Computing (HPC), those who fail to embrace and adopt new technology are likely to be left behind.

In the remainder of this article we discuss possible areas digital technologies can help in formulated product R&D processes before discussing examples of our research into the development of technology demonstrators. We discuss the insight gained by developing these approaches and the challenges involved.

3 What Can Digital Offer the Formulated Products Industry?

There are three areas where digital approaches can add value to formulated product R&D:

1. Increasing connectivity and transparency by providing: digital record keeping, such as digital laboratory notebooks; better communications through collaborative applications, social media and messaging applications; IoT sensing on lab equipment, providing real time recording.

2. Use of real time data analytics and extraction from online and internal knowledge databases for predicting properties calling on machine learning and natural language algorithms

3. Creating digital first based lab practices where scientists will trial formulations using accurate physical and chemical models before developing the most promising candidates in the real lab.

Significant progress has already been made by industry in adopting the first approach. For example, digital laboratory notebooks can provide enhanced and more reliable data collection using I4.0 type technology. Recent changes have seen digital laboratory notebooks becoming increasingly connected with common office software and inexpensive analytics. This connectivity can help provide a boost to productivity, but it is unlikely to transform how innovation occurs.[5]. The second and then third approaches are more difficult to implement and to adopt, however, will be increasingly important when trying more novel approaches or novel chemicals are required.

To change the way we innovate, digital applications which augment our current processes need to be implemented and coupled with more computationally intensive applications that can provide deep insights and predictions. The key goal for I4.0 in an R&D setting is to speed up the innovation and discovery process, reduce costs and minimize time to market through the use of digital technologies suited to organization's specific market area.

At a high level, I4.0 services generically comprise three layers. The foundation layer of such services are low level tasks and can range from something quite simple, such as updating a database, gathering data from sensor arrays or changing IoT device parameters to complex modelling such as resource forecasting, physical simulations or optimization.[6] The middle layer controls the connectivity, security and sharing of various software and hardware components to one another. In the simple picture we describe here this could be passing user inputs to the foundation application or sharing data from one component to another. The top layer provides the interface that the client can access and interact with such as a web portal. Figure 2 provides a pictorial explanation of these critical components, together with examples of the key skills and technologies required alongside a summary of the key opportunities and challenges.

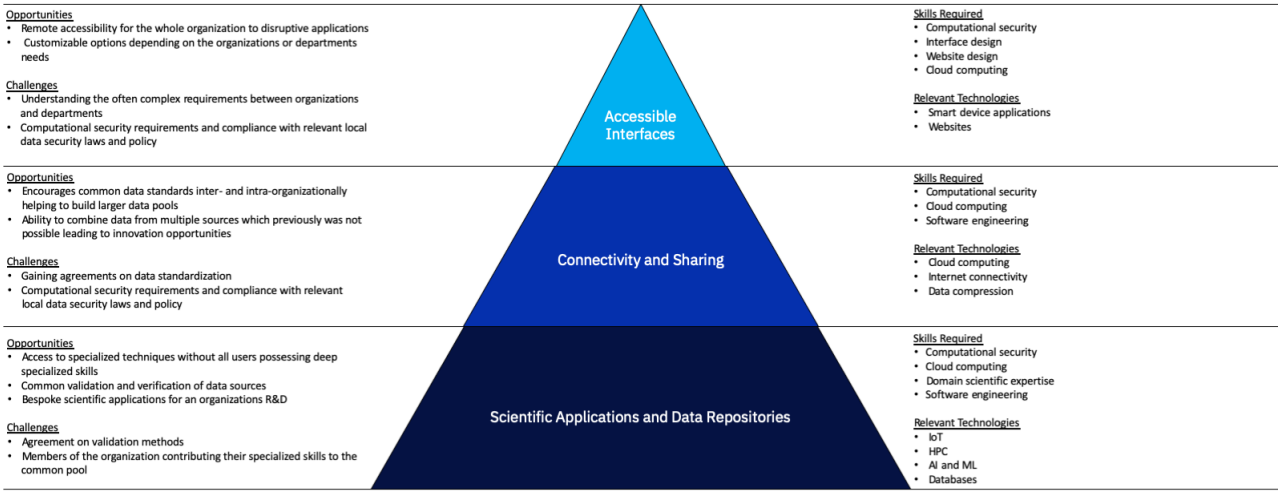

Figure 2: High level overview of industry 4.0, opportunities, challenges, skills and technologies.

Digitization of R&D requires a high degree of technical knowledge, both from a scientific domain and technology perspective [6, 2] which from our experience requires engagement from both the I4.0 vendor and industry user. Thus the major blockers for R&D therefore are not simply technological in nature.

4 Data Collection from the Physical Laboratory

Developing digital tools to augment the discovery process presents some not insignificant data challenges. The first hurdle is the capture of relevant data and meta-data in a suitable format for future exploitation. That is, data captured *via* either smart laboratory notebooks, data analysis carried out by domain scientist or automated measurement equipment, needs to be machine readable and contain enough information that the experiment's methodology, results and conditions is determinable later.

A set or sets of data standards, such as those suggested University of Liverpool's Manufacturing Innovation Factory [5], need to be agreed and adopted so that software can be built for data extraction as a first step in development of these digital assets. Meta-data should include as a minimum the thermodynamic conditions, experimental method, apparatus and software used. This will help ensure a fair comparison, of compatible data is used in data analysis and model development. In laboratories which already employ automation, through robotics for example, the automatic storage of data may even be able to consider employing inexpensive IoT sensors such as temperature and humidity to gather additional information.

Chemical identifiers permit rapid data base searching and curation. For formulations this could be achieved by using a unique number for each formulation in a similar manner to CAS numbers.[13] However, this would lead to a vast number of identifiers often referencing minor variations, hence may not be the most suitable choice. Alternatively, a more complex identifier developed from extensions of the standard single molecule string identifiers such as SMILES [14] and InChI [15] could provide a more informative identifier, for example, a formulation version of SMILES defining the concentration of each molecule followed by it sting identifier $30.0 < O > 50.0 < CO > 20.0 < Cc1ccccc1 >$. These chemical string identifiers convey a chemical structure to varying degrees of complexity, which standard chemical software is able to parse and produce the chemical structures and links to their physical properties. The degree of flexibility and variation in formulation science however, may require new thinking, considering methods of dimensionality reduction for example to generate a reversible compressions of a formulation string identifier. Such a method can enable a multipurpose formulation representation that can be used for rapid data base searching in the compressed form and uncompressed to provide scientists with some limited information on the formulation.

In Section 5.3 we will return to the knock-on impact on exploitation by digital applications when data is not of good quality and well documented.

5 The Digital First Formulation Lab

Formulated products such as lubricants, fuel additives, paints and shampoos rely for their performance on complex physico-chemical processes such as adsorption, aggregation and micelle formation. These processes are becoming increasingly well modelled using computer simulation techniques.[16, 17, 18, 19, 20] However, the ability for this type of approach to be taken up in industry is limited as many of the underpinning skills

are simply not present in organizations developing formulated products. Specialists in chemical modelling, computer scientists and high performance computing administrators traditionally need to come together to make simulation work.

We have developed a technology demonstrator in the form of an iPad-based digital platform that could form part of the I4.0 solution transforming R&D laboratory practices to a digital first based approach, i.e. model before make mentality. This application enables the user to develop and model chemical formulations virtually, to understand and predict their physical properties and performance. The results can be used to inform an experimental scientist of the most promising formulations or even to directly instruct a formulation robot. This way of working reduces wasted materials and in some cases, given sufficient computer resources, will out pace physical testing.[6, 21]

5.1 Demonstration of Technology

Our prototype digital application is able to ingest user input in which chemicals and concentrations are specified and launch complex physical simulations and analytics, returning to the user a simulated predicted outcome (see Fig. 2)).[21] This prototype utilizes a smart mobile application as a user interface that communicates with a cloud gateway to submit computer intensive simulation and analytics to an HPC server.

Figure 3 gives a schematic overview of the core pieces of the prototype framework. The only visible component to an end user is the interface application. As a result the design and construction of this is critical enable the flexibility for R&D concurrently with ease of use. If these objectives are not achieved, it can limit the utility of the application and dissuade users from attempting to augment their practices with the new technology.

The user interface may be the start and end point from a user perspective, but the underlying layers must be constructed in order to enable the application to provide a utility. Connections from the interface go out to a cloud gateway. This can be thought of as a management service, orchestrating the connections and data transfer from the user to the HPC application. For instance, the molecules and concentrations to include in a simulation. This gateway also offers the principle security barrier to prevent miscellaneous access.

Here, the cloud gateway coordinates with the available HPC or cloud resources to allocate the hardware required for the requested simulations and analytics. The state-of-the-art simulation and analysis process is encoded into software as a workflow. An automated scheduler interprets these steps and schedules the relevant calculations.

Results are returned to the user interface via the cloud gateway in a format analogous to experimentally determined results together with estimates of uncertainty. The raw data underpinning this is also made available via the cloud gate way in a computationally convenient format, such that more advanced users can make additional enquiries from that data.

This use case illustrates the concept of digitizing specialist knowledge of simulation techniques and making them accessible to anyone able to use a smart phone application. For an organization this represents a dramatic lowering of the skill barriers for using HPC to perform physical modelling and analytics relevant in day to day R&D.

Figure 3: Image of the prototype iPad application which was constructed demonstrating its ability to launch and provide a predicted ternary phase diagram.

Some of the challenges in introducing these applications such as the system stability, reliability and deployment environment, are solvable and the same as seen in other areas of I4.0. Such issues can manifest in variations in end users data security requirements, preferred methods of data sharing and restrictions on the location of software and hardware. More specific challenges of adoption, come from the specific tailoring of the device or services to fit within an organisation existing structures, such as matching the nomenclature used by their experts, working practices and the general willingness of staff to adopt new technologies and digitally augmented practices. Some of this will be due to unfamiliarity with digital approaches compared to lab based approached and how they relate to one another. Such barriers can be overcome by building strong relationships, developing trust and providing knowledge transfer between the different stakeholders. A critical challenge for devices aimed at R&D is knowing when a result is accurate and precise enough to be useful. Quality and precision concerns can be mitigated with suitable visual representations and metrics of describing the uncertainty in simulation, data and computation, so that confidence can be easily interpreted by a formulation domain scientist. Accuracy is due to either the method of measurement or the underlining model. Here it may be necessary to determine the accuracy of the underlying models and the data upon which such models were based.

5.2 Public Sources of Data

Chemical data providers, describing molecular properties have incorporated database methods and informatics for many years. This has produced easily accessible data sources, which we take for granted when considering chemical purchasing, risk labels and emergency advice. Larger online chemical databases such as ChemSpider[22], PubChem[23] and CCDC [24] provide computational interfaces and API's to easily access the data within.

Open data is now a commonly used term, meaning data which is freely available, shareable and utilizable by anyone, anywhere. Of the open chemical data sets which have been published, there are many which are relevant for formulation R&D. For example, NIST data sets, [25] which have been well curated over years, provide promising data sets for formulation.

Smaller more focused databases have been a catalyst for innovative approaches in targeted communities. For example publicly available standardized datasets, such as provided by blind challenge competitions like the solubility challenge, [26, 27], industrial fluid properties simulation challenge [28] or the crystal structure prediction challenges[29], have delivered new generations of algorithms and enabled comparative testing of different predictive models, which has led to rapid growth in these areas. These data sets have proven valuable in the modelling of small organic molecules relevant to the pharmaceutical industry [30, 27, 26, 31, 32, 33, 34] and in some cases formed *de facto* standards for modelling certain properties.

5.3 Obtaining Reliable Physical Models

Computer based models of chemical systems come in two forms, data-driven models (see section 5.4) and physical models. Physical models seek to describe the microscopic interactions between molecular entities.[35, 36, 37, 38] They usually consist of a set of mathematical functions that encode the possible interactions between the entities, with the functions parameters varying depending on which entities are interacting. Physical models can be used to simulate chemical systems and to make measurements analogous to wet-lab experiments. As a result digital applications based on physical models can be thought of as "virtual experiments".

The model behind the virtual experiments offered by our digital first formulation laboratory is called Dissipative Particle Dynamics (DPD) and has been used successfully for modelling many processes and chemical mixtures in a chemical formulation context. [39] There are two features of this modeling approach that are worth noting. First is that is is a coarse grained approach whereby groups of atoms are treated with a single interaction site. This allows simulation of significantly larger systems than more common atomistic methods, such as molecular dynamics. Second is that the interactions between sites are simplified. This translates to fewer and simpler mathematical functions which are faster to compute. Taken together these features enable DPD to simulate the length and time scales required to capture many physical properties of liquid-based formulations.

The success of any virtual experiment depends on the accuracy of the physical model. It needs to be capable of providing useful predictions of the quantities of interest, for example liquid densities, octanol-water partition coefficients ($log_{10}P$), or in the case of surfactants, Critical Micelle Concentration (CMC) and micelle mean aggregation number (N_{agg}). [38, 18, 40, 19, 41] Assuming the physical model is appropriate for the system under consideration, the main factor affecting accuracy is the quality of the models parameters. Improving these parameters requires acquisition of suitable experimental data for model tuning and validation, a process termed *parameterisation*.

Since novel chemistry is a common circumstance in industrial R&D, realising the potential of computational techniques in I4.0 requires not only accurate models but also a rapid parameterisation process. A model that is very accurate, but takes 10 years produce, will likely have little impact. Unfortunately, to date, and from our own experience, it has been common for DPD parameterisation to be a multi-year effort requiring significant human capital. Hence, there has been renewed interest in automatic parameterisation methods. These methods can potentially ingest experimental data on the novel chemistry of interest and produce a tuned physical model with minimal human involvement. A number of automated parameterisation techniques are being developed, from gradient-free global optimisation methods, to local-gradient methods leveraging statistical mechanics to obtain the gradient of macroscopic properties with-respect-to force-field parameters. Whichever method is used the promise is a rigorous automated protocol to find the optimal parameters for the model.

All of these techniques require as a base a consistent set of experimental measurements on a diverse set of molecules. In the absence of organized and curated data sources, one is tempted to obtain the necessary formulation data from the experimental literature and, in fact, there is a wide variety of data available. However, one soon discovers not only that the coverage is sparse, but, when data can be found, the inter-lab or inter-method variability is very high, rendering the use of this data for model development and validation challenging.

During the development of a data-set for parameterisation of a model for micelles we encountered numerous issues which can be divided into two types: problems with the reporting and provenance of the experimental data being collected, in this case CMCs and mean-aggregation numbers; and problems related to a lack of sufficient meta-data, which reports the experimental protocols, methods, instrumentation, and data processing assumptions utilized by the measurement.[42]

Here we will briefly outline some examples of these issues. Interested readers can find a detailed account in Swope *et al.* and its supplementary information [40]. Falling into the first category a problem we found frequently in the literature was tabulating data as comparable when it is not, by failure to recognize the difference between weight and number averaging in determining N_{agg} number in micelles. Similarly we found that values in the literature were reported without including the temperature and/or the concentration at which the measurements were done, presumably because the authors did not understand at the time that micelle size has both thermal and concentration dependence.

When parameterising a model, a key step is measuring properties in as similar a way as possible, in both the experiment and simulation. Since the experimental method is usually set a deep understanding of what exactly it measures is required to replicate it in the simulation. Issues falling in the second category, lack of metadata, primarily impact this process. For example, on observing the raw distributions obtained from dynamic light scattering (DLS) experiments we observed that the signal from small aggregates is weak or missing due to it being obscured by the signal from the larger aggregates. This meant that when measuring the aggregation number in simulation we must similarly discount small aggregates. Since this number varies with surfactant access to the raw data is required to determine where the cutoff is, or to re-process the experimental data by beginning averaging from a size where one believes the signal has reached full strength.

Motivated by these issues and the need for a consistent data set of micelle sizes for model development, we recently experimentally determined a small consistent dataset for non-ionic surfactants which can provide a

useful starting point for parameterising surfactant models. [40]. In particular this data set contains abundant information on the experimental conditions and the data processing pipeline applied, information which is crucial for parmaeterisation effort. Clearly such a approach, whereby scientists must return to the lab to produce data of the necessary quality for parameterisation, is un-scalable and unsustainable long term. We recommend that support needs to be provided to encourage the development of new data sets oriented to computational model development. Such activities can be encouraged by funding bodies and learned societies, offering financial backing and expertise. Recently, several funding bodies have released challenges in this area.[43, 44]

5.4 Data Driven Formulation Property Modelling *via* Machine learning

In the pharmaceutical industry there are active efforts to use open data and to explore data driven computational platforms.[45, 46, 47, 48] These efforts are not without difficulties, and are often attempted by single organizations, hence, may lack the inter-organizational standards needed completely open data. However, some benefits are being seen in this area in terms of insights from calculations and high throughput screening of lead molecule structures.[49, 45] The opportunities specific to this area have been discussed in a number of other articles, a compact summary is given in the following communication.[50]

Similar efforts in other formulation heavy industrial sectors such as personal care, automotive and coatings are less commonly discussed but could benefit from similar initiatives, technology and methods. Organizations in these sectors may in fact be more able to adopt such methods owing to less stringent regulation compared to food, drink and pharmaceutical organizations. Potentially, these less discussed sectors, are able to adopt I4.0 applications more rapidly and widely.

In this section, we provide an example of a data driven model utilizing machine learning for CMC prediction. There are numerous examples of data driven modelling for chemical properties predictions relevant to formulation.[51, 52, 53] However, in many cases, the focus is on properties related to pharmaceuticals, for example, solubility, toxicity and organic-aqueous partition coefficients.[31, 33, 32, 54] Some work has attempted to consider modelling of formulated products *via* ML methods more generally. [55, 56, 57, 58] Methods such as neural networks have been applied to optimize the processes in formulation [58] and suggested as useful tools in discovery and chemical product optimization.[59, 60, 61, 51] In this brief example, we demonstrate the application of a ML method called Random Forest (RF) to the prediction of CMC values. We discuss briefly the data collected, modelling method and information which can be extracted. The dataset used here is provided as supporting information taking account of the recommendations given in this manuscript. Details of the model are given in appendix A.

We chose to apply the popular RF model to predicting the CMC of electrically neutral surfactant molecules because of its simplicity and success elsewhere. To do this we built a data set from the open literature. The collection and curation of this data set was a laborious task, requiring approximately five days of a researchers time to search and curate the data set. This compares to about a day to build the model. Whilst it is common for data pre-processing to take the largest portion of the time, the data we are working with in this example is small (87 molecules). The reasons for this are due to the data being stored in formats which are not machine processable and a lack of contextual meta-data. For example, the largest single source of data was extracted from "Critical Micelle Concentrations of Aqueous Surfactant Systems" [25]. This document was created as a pdf of scanned images, meaning all information had to be translated by hand. This text was curated well before many of standards for molecular structure representation were commonly employed, hence these had to be located or generated. The second major contributor was from a previous similar study [51]. This source contained all information within the a table of the main article which was easier to extract, but still sub-optimal. The major issue with this source was the lack of standardized chemical names and molecular representations such as SMILES which were present at the time the article was written. Locating or generating this missing data is the major reason for the five day data collection and curation.

Databases or electronically processible formats can dramatically speed up this process. For instance workflow tools such as Taverna[62] have been used to automatically extract molecular representations from online gateways for solubility prediction, making the process faster and more importantly repeatable.[33]

We include the 87 molecule data set in the supporting information in a csv format together with meta-data such as IUPAC names, SMILES, InChI, temperature and experimental method where possible. Whilst the format we present is far from the data base model, it is a step towards this and enables our work to be easily repeated and built upon.

A test set of 12 molecules were chosen from the 87 molecules and is comprised of previously validated data by the authors.[40] The model was generated by training on 75 molecules and predicting the remaining 12 molecules as an unseen test set. The results of the test set predictions are presented in table 1 and figure 4.

Table 1: Results for predicting the CMC in molar units using the RF model for 12 test set molecules. R^2 is the correlation coefficient, RMSE is the root mean squared error, MAE is the mean absolute error and MSE is the mean signed error.

R^2	RMSE	MAE	MSE
0.96	0.31	0.01	-0.01

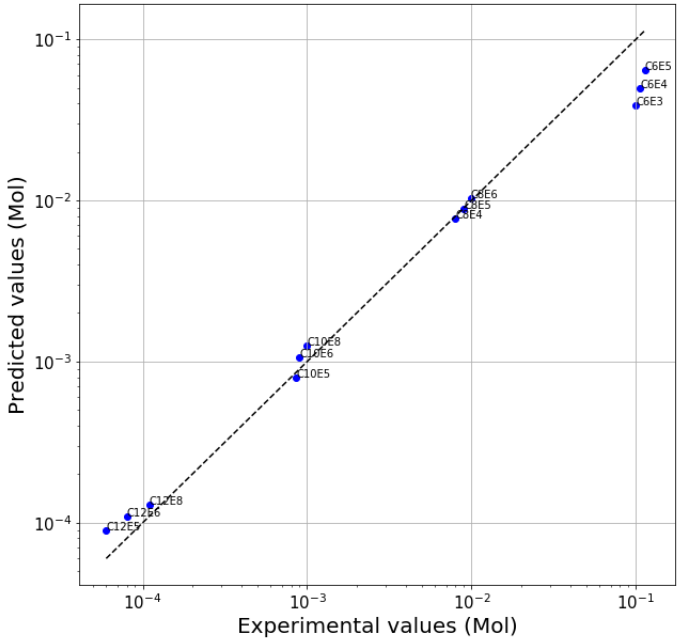

Figure 4: Predictions for 12 test set molecules. Black dashed line shows the ideal model. Note this is plotted on a log scale to make the relatively small concentration changes visually clear.

These results showing a reasonable predictive model for CMC over this test set. The model is able to explain the vast majority of the data variance as evidenced by the R^2 value of 0.96 and low Root Mean Square Error (RMSE).

In both the simulation and data cases one should be aware that both models are typically bounded in terms of the chemistry they can interpret and predict. This bound comes from the data which was used to construct the model. A well constructed model should be able to generalize to a reasonable extent and thus have utility when considering related chemistry.

ML models in general do not provide physical understanding of a chemical system and are data intensive, requiring much larger consistent data sets than are required for physical simulation counter parts. However, Such data driven models are typically orders of magnitude faster to evaluate and make predictions than their physical simulation counter parts once trained.[63] These models are therefore compute intensive to train, taking many hours on current computer hardware, but able to be utilized many times afterwards for rapid predictions. This kind of modelling is particularly useful for high through put predictive screening methodologies. Where organization have large data repositories or access to suitable open data, such models can make highly effective screening or ranking methods to rapidly guide design of experiments approaches in the lab, whether robotically or in collaboration with a research scientist. The organizational level benefit is to enable rapid screening of innovative designs and solution to R&D challenges. Ultimately such as changes aims to minimize the time to solution when one is considering reformulation for a regulatory change for example.

6 Discussion and Conclusions

In this article, we have identified several technical issues which are tractable, but currently preventing industrial formulation R&D from taking full advantage of I4.0 methods. The challenges fall into two categories:

- Development of novel digital applications that capture the relevant science for formulation R&D

- Long term accessible storage for well validated and documented experimental data

To the first point, we have noted that the technology needed to access, deploy and utilize such methods is already in existence. We can see much of this technology utilized for other business related I4.0 applications. The challenge here, is the development of suitable foundation I4.0 applications for innovation in formulation R&D. The development of digital applications requires a mixture of expertise and experience often not found within a single organization. To gain traction, it is critical that users and developers of such digital services, are willing and able to work collaboratively between organizations to construct applications suitable as augmentation to existing working practices.

On the second point, there are a number of smaller challenges and potential research projects which are required to solve this challenge in the longer term. The major requirement in moving towards a solution is the hosting of trusted and accessible formulation databases. Efforts have been made in the past few years in this direction from the Chemical Abstracts Service (CAS) who have generated a formulation database product.[64] However, an openly accessible formulation database is presently not available. Such a database, could draw in broad attention across academia and industry, leading to a community wide asset. Such a community asset can assist in building bridging skill sets, with students learning to apply computational techniques on formulation data whilst at university.

For these digital services to become widely adopted the technical challenges have to be addressed and the solutions must provide accurate and precise results. These services will only be adopted if formulation organizations can rely on the results that are produced. From a technology point of view we must insure that the devices produce an accurate and precise result, as outlined previously, which is why it is essential that the underlying experimental data that goes into the these applications is correct. Additionally, organizations need to consider the human psychological aspects involved when introducing new ways of working. Times of change within an organization can be difficult, with individual willingness to embrace new technology varying. In order to successfully adopt new methods technology advocates must establish working partnerships between organizations demonstrating the value of these new methods within their own organizations. When adopting the new methods, it is also important that the end users are upskilled, to enable them to take advantage of the new opportunities and decide for themselves where best such methods can augment their existing working practices.[65] Such psychological factors usually require strong relationships and trust to be built between the different domain experts in order to understand the concerns from both sides and find a suitable resolution where difficulties occur.

Digitization and I4.0 applications are already driving organizational changes and augmenting working practices. Formulation science R&D remains a largely laboratory driven enterprise. As societal and regulatory pressures grow and shift more rapidly with increasing environmental challenges formulators will need to be able to respond at greater pace, which will most likely out pace the laboratory focused processes. I4.0 should not be viewed as replacement to laboratory processes but as an augmentation to help guide experimental resources in the most promising directions first, in order to minimize the time to solution and provide efficiency savings.

Presently formulation product R&D faces challenges in adopting I4.0 opportunities due to the need for specialized applications and a lack of accessible and curated data. These challenges can be minimized in the short term with minor modifications for working and publication practices. We would recommend that journals in these fields make reasonable efforts to ensure that authors do provide such data, in appropriate formats, which will also assist research reproducibility.

In the longer term the formulation community can, in collaboration with domain experts in computational science, define innovative models and data bases which will over come these challenges and enable formulation science to take advantage of I4.0 opportunities. Such collaboration and skill sharing across organizations will be critical to enabling this change and upskilling work forces. Here learned societies and funding bodies should assist as trusted neutral partners to provide experience, resources and connections. Such institutions can also help to span domain areas, creating open forums for the fair discussion and evaluation of such techniques

Looking forward these opportunities can undoubtedly provide advantage in formulation science R&D, disrupting the current standards to increase flexibility, knowledge and innovation. From this perspective, the longer term (5 - 10 years) could see exciting opportunities for formulation organizations to develop innovative and novel approaches to product modification, development and optimization.

7 Acknowledgements

This work was supported by the STFC Hartree Centre's *Innovation: Return on Research programme*, funded by the Department for Business, Energy & Industrial Strategy.

References

[1] Lasi H, Fettke P, Kemper HG, Feld T, Hoffmann M. Industry 4.0. Business & information systems engineering. 2014;6(4):239–242.

[2] Stock T, Seliger G. Opportunities of sustainable manufacturing in industry 4.0. Procedia Cirp. 2016;40:536–541.

[3] Gentner S. Industry 4.0: reality, future or just science fiction? How to convince today's management to invest in tomorrow's future! Successful strategies for industry 4.0 and manufacturing IT. CHIMIA International Journal for Chemistry. 2016;70(9):628–633.

[4] Thienen SV, Clinton A, Mahto M, Sniderman B. Industry 4.0 and the chemicals industry. Catalyzing transformation through operations improvement and business growth. Deloitte Insights; 2016. Accessed: 27-01-2020. https://www2.deloitte.com/us/en/insights/focus/industry-4-0/chemicals-industry-value-chain.html.

[5] Innovation 4.0: A Digital Revolution for R&D;. 17-1-2020. https://www.newstatesman.com/spotlight/manufacturing/2019/09/innovation-40-digital-revolution-rd.

[6] Rodič B. Industry 4.0 and the new simulation modelling paradigm. Organizacija. 2017;50(3):193–207.

[7] University of Manchester, University and Unilever launch collaborative project; 2017. Accessed: 27-01-2020. https://www.staffnet.manchester.ac.uk/news/display/?id=19007.

[8] BEIS. Industrial Strategy: building a Britain fit for the future. 2017;.

[9] Constable DJ, Dunn PJ, Hayler JD, Humphrey GR, Leazer Jr JL, Linderman RJ, et al. Key green chemistry research areas—a perspective from pharmaceutical manufacturers. Green Chemistry. 2007;9(5):411–420.

[10] Forsyth M, Porcarelli L, Wang X, Goujon N, Mecerreyes D. Innovative electrolytes based on ionic liquids and polymers for next-generation solid-state batteries. Accounts of chemical research. 2019;52(3):686–694.

[11] Huang Y, Zhao L, Li L, Xie M, Wu F, Chen R. Electrolytes and Electrolyte/Electrode Interfaces in Sodium-Ion Batteries: From Scientific Research to Practical Application. Advanced materials. 2019;31(21):1808393.

[12] Lanlan F, Nanping D, Jing Y, Zhenhuan L, Weimin K, Bowen C. The recent research status quo and the prospect of electrolytes for lithium sulfur batteries. Chemical Engineering Journal. 2019;.

[13] CAS Registry System. Journal of Chemical Information and Computer Sciences. 1978;18(1):58–58. Available from: https://pubs.acs.org/doi/abs/10.1021/ci60013a609.

[14] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences. 1988;28(1):31–36.

[15] Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI-the worldwide chemical structure identifier standard. Journal of cheminformatics. 2013;5(1):7.

[16] Seaton MA, Anderson RL, Metz S, Smith W. DLMESO: Highly Scalable Mesoscale Simulations. Mol Sim. 2013;39(10):796–821.

[17] Lee MT, Vishnyakov A, Neimark AV. Calculations of Critical Micelle Concentration by Dissipative Particle Dynamics Simulations: The Role of Chain Rigidity. J Phys Chem B. 2013;117(35):10304–10310.

[18] Johnston MA, Swope WC, Jordan KE, Warren PB, Noro MG, Bray DJ, et al. Toward a standard protocol for micelle simulation. The Journal of Physical Chemistry B. 2016;120(26):6337–6351.

[19] Anderson RL, Bray DJ, Del Regno A, Seaton MA, Ferrante AS, Warren PB. Micelle Formation in Alkyl Sulfate Surfactants Using Dissipative Particle Dynamics. J Chem Theory Comput. 2018;14(5):2633–2643.

[20] Panoukidou M, Wand CR, Del Regno A, Anderson RL, Carbone P. Constructing the phase diagram of sodium laurylethoxysulfate using dissipative particle dynamics. Journal of Colloid and Interface Science. 2019;557:34–44.

[21] AbdelBaky M, Diaz-Montes J, Johnston M, Sachdeva V, Anderson RL, Jordan KE, et al. Exploring HPC-based scientific software as a service using CometCloud. In: 10th IEEE international conference on collaborative computing: networking, applications and worksharing. IEEE; 2014. p. 35–44.

[22] Pence HE, Williams A. ChemSpider: an online chemical information resource. ACS Publications; 2010.

[23] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic acids research. 2015;44(D1):D1202–D1213.

[24] Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallographica Section B: Structural Science. 2002;58(3):380–388.

[25] Mukerjee P, Mysels KJ. Critical micelle concentrations of aqueous surfactant systems. National Standard reference data system; 1971.

[26] Hewitt M, Cronin MT, Enoch SJ, Madden JC, Roberts DW, Dearden JC. In silico prediction of aqueous solubility: the solubility challenge. Journal of chemical information and modeling. 2009;49(11):2572–2587.

[27] Llinas A, Glen RC, Goodman JM. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? Journal of chemical information and modeling. 2008;48(7):1289–1303.

[28] Moore JD, Mountain RD, Ross RB, Shen VK, Siderius DW, Smith KD. The ninth industrial fluid properties simulation challenge. Fluid phase equilibria. 2018;476:1–5.

[29] Day G, Motherwell W, Ammon H, Boerrigter S, Della Valle R, Venuti E, et al. A third blind test of crystal structure prediction. Acta Crystallographica Section B: Structural Science. 2005;61(5):511–527.

[30] Hopfinger AJ, Esposito EX, Llinas A, Glen RC, Goodman JM. Findings of the challenge to predict aqueous solubility. Journal of chemical information and modeling. 2008;49(1):1–5.

[31] Hughes LD, Palmer DS, Nigsch F, Mitchell JB. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. Journal of chemical information and modeling. 2008;48(1):220–232.

[32] McDonagh JL, van Mourik T, Mitchell JBO. Predicting melting points of organic molecules: applications to aqueous solubility prediction using the general solubility equation. Molecular informatics. 2015;34(11-12):715–724.

[33] McDonagh JL, Nath N, De Ferrai L, van Mourik T, Mitchell JB. Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. Journal of chemical information and modeling. 2014;54(3):844–856.

[34] Skyner RE, Mitchell JB, Groom C. Probing the average distribution of water in organic hydrate crystal structures with radial distribution functions (RDFs). CrystEngComm. 2017;19(4):641–652.

[35] Jorgensen WL, Maxwell DS, Tirado-Rives J. Development And Testing Of The Opls All-Atom Force Field On Conformational Energetics And Properties Of Organic Liquids. J Am Chem Soc. 1996;118(45):11225–11236.

[36] Wang LP, Martinez TJ, Pande VS. Building Force Fields: An Automatic, Systematic, And Reproducible Approach. J Phys Chem Letters. 2014;5(11):1885–1891.

[37] McDonagh JL, Silva AF, Vincent MA, Popelier PL. Machine learning of dynamic electron correlation energies from topological atoms. Journal of chemical theory and computation. 2017;14(1):216–224.

[38] McDonagh JL, Shkurti A, Bray DJ, Anderson RL, Pyzer-Knapp EO. Utilizing Machine Learning for Efficient Parameterization of Coarse Grained Molecular Force Fields. Journal of chemical information and modeling. 2019;59(10):4278–4288.

[39] Español P, Warren PB. Perspective: Dissipative particle dynamics. J Chem Phys. 2017;146(15):150901. Available from: http://dx.doi.org/10.1063/1.4979514.

[40] Swope WC, Johnston MA, Duff AI, McDonagh JL, Anderson RL, Alva G, et al. Challenge to Reconcile Experimental Micellar Properties of the CnEm Nonionic Surfactant Family. The Journal of Physical Chemistry B. 2019;123(7):1696–1707.

[41] Anderson RL, Bray DJ, Ferrante AS, Noro MG, Stott IP, Warren PB. Dissipative particle dynamics: Systematic parametrization using water-octanol partition coefficients. The Journal of chemical physics. 2017;147(9):094503.

[42] Zimmerman AS. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology, & Human Values. 2008;33(5):631–652.

[43] Department of Defense - Air Force Research Lab, Materials Science and Engineering Data Challenge; 2015. Accessed: 21-01-2020. https://www.challenge.gov/challenge/materials-science-and-engineering-data-challenge/.

[44] UKRI. New £81 million Materials Innovation Factory opens; 2018. Accessed: 21-01-2020. https://www.ukri.org/news/new-81-million-materials-innovation-factory-opens/.

[45] Green DV, Pickett S, Luscombe C, Senger S, Marcus D, Meslamani J, et al. BRADSHAW: a system for automated molecular design. Journal of computer-aided molecular design. 2019;p. 1–19.

[46] Kogej T, Blomberg N, Greasley PJ, Mundt S, Vainio MJ, Schamberger J, et al. Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer Pharma AG case. Drug discovery today. 2013;18(19-20):1014–1024.

[47] Data Library;. Accessed: 17-01-2020. https://openinnovation.astrazeneca.com/data-library.html.

[48] Skyner RE, McDonagh JL, Groom CR, Van Mourik T, Mitchell JBO. A review of methods for the calculation of solution free energies and the modelling of systems in solution. Physical Chemistry Chemical Physics. 2015;17(9):6174–6191.

[49] Bergström CA, Norinder U, Luthman K, Artursson P. Experimental And Computational Screening Models For Prediction Of Aqueous Drug Solubility. Pharm Res. 2002;19(2):182–188.

[50] Tetko IV, Engkvist O, Koch U, Reymond JL, Chen H. BIGCHEM: challenges and opportunities for big data analysis in chemistry. Molecular informatics. 2016;35(11-12):615–621.

[51] Huibers PD, Lobanov VS, Katritzky AR, Shah DO, Karelson M. Prediction of critical micelle concentration using a quantitative structure- property relationship approach. 1. Nonionic surfactants. Langmuir. 1996;12(6):1462–1470.

[52] Mokrushina L, Buggert M, Smirnova I, Arlt W, Schomäcker R. COSMO-RS and UNIFAC in prediction of micelle/water partition coefficients. Industrial & Engineering Chemistry Research. 2007;46(20):6501–6509.

[53] Godavarthy SS, Robinson Jr RL, Gasem KA. Improved structure–property relationship models for prediction of critical properties. Fluid Phase Equilibria. 2008;264(1-2):122–136.

[54] Palmer DS, O'Boyle NM, Glen RC, Mitchell JB. Random forest models to predict aqueous solubility. Journal of chemical information and modeling. 2007;47(1):150–158.

[55] Rowe R, Colbourn EA. Neural computing in product formulation. Chem Educator. 2003;8:1–8.

[56] Peremezhney N, Connaughton C, Unali G, Hines E, Lapkin AA. Application of dimensionality reduction to visualisation of high-throughput data and building of a classification model in formulated consumer product design. Chemical Engineering Research and Design. 2012;90(12):2179–2185.

[57] Childs CM, Washburn NR. Embedding domain knowledge for machine learning of complex material systems. MRS Communications. 2019;9(3):806–820.

[58] Wu T, Pan W, Chen J, Zhang R. Formulation optimization technique based on artificial neural network in salbutamol sulfate osmotic pump tablets. Drug development and industrial pharmacy. 2000;26(2):211–215.

[59] Hussain MA. Review of the applications of neural networks in chemical process control—simulation and online implementation. Artificial intelligence in engineering. 1999;13(1):55–68.

[60] Widrow B, Rumelhart DE, Lehr MA. Neural networks: applications in industry, business and science. Communications of the ACM. 1994;37(3):93–106.

[61] Pyzer-Knapp EO. Cognitive Chemistry: The Marriage of Machine Learning and Chemistry to Accelerate Materials Discovery. Materials Informatics: Methods, Tools and Applications. 2019;p. 223–251.

[62] Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic acids research. 2013;41(W1):W557–W561.

[63] McDonagh JL, Mitchell JBO, Palmer DS, Skyner RE. In: 3. In Silico methods to predict solubility; 2020. p. 71–112.

[64] CAS. Formulus is key to unlocking R&D productivity; 2020. Accessed: 21-01-2020. https://www.cas.org/products/formulus.

[65] Urbach N, Röglinger M. Introduction to Digitalization Cases: How Organizations Rethink Their Business for the Digital Age. In: Digitalization Cases. Springer; 2019. p. 1–12.

[66] Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. Journal of cheminformatics. 2018;10(1):4.

[67] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825–2830.

A Appendix: Random Forest Model Details

A.1 Dataset

In this work we have collected a data set of 87 molecules with CMC values known in the experimental literature. Where possible these have been taken from standard sources such as NIST, which has produced a well curated and research set of CMC values. The 87 molecules are all electrically neutral and represent a diverse set of surfactants. The data set contains several 2D representations of the molecular structure (SMILES and InChI) together with URLs and InChIkeys. These should enable anyone else to easily follow this work and use the dataset. We report the CMC in molar units and $\log_{10}$ molar units. Additionally, the experimental method of determination where available, temperature, references and any relevant notes on ambiguities and how data was found is included as meta-data.

For each molecule entry the Mordred[66] descriptor engine was used to calculate 2D descriptors from the SMILES strings. This led to over 1000 descriptors per molecule. Descriptors are a way to represent (usually numerically but not always) aspects of molecule. These generally constitute structural complexity, counts of atoms, polarity, and simple electronic properties. The ML algorithms can use these to correlate against the target experimental property.

A.2 Random Forest Model

The RF model is an ensemble learning algorithm. It has been used to great success in many applications. The centre of the model is a binary decision tree, which makes optimal splits of the data based on minimizing the Mean Squared Error (MSE) in this case. The forest is constructed of multiple of these trees, where each tree is built from a random subset of the data and descriptors. The final prediction of the CMC is given as the average of all of the trees predictions. RF is generally consider fairly robust to over fitting due to each tree only being trained on a subset of the data and to redundant descriptors, i.e. it is possible for several descriptors to largely convey the same information, this can lead some algorithms to consider this to be more important and weight it more highly.

In the current work we optimized the hyper-parameters of the RF model (number of trees and depth of each tree) using a grid search spanning number of trees from 100 to 1000 in steps of 100 and the tree depth from 5 to 15 in steps of 1. The optimal parameters were found to be 300 trees and a depth of 9. As this is a small data set we did not wish to use all descriptors as this would lead to an over determined problem. We applied Recursive Feature Elimination (RFE) to rank the importance of each feature. Scikit-learn was used to perform this model generation and predictions.[67]

Model training and predictions was carried out by the standard train test split methodology. This methodology means the data is partitioned into training and testing sets, the model is constructed using the training set and then tested on unseen data in the test set. In the current manuscript we held out 12 molecules as an external test set from the 87, these were molecules whose data we were confident of based on the authors previous work.[40]