doi:10.2533/chimia.2021.1 Chimia 75 (2021) 1–8 © Swiss Chemical Society

FH HES

Universities of Applied Sciences

Fachhochschulen – Hautes Ecoles Spécialisées

Digitalization in Processes

Olivier Vorlet[a], Lukas Neutsch[b], Christian Kronseder[c], and Alexandre Kuhn*[d]

*Correspondence: Prof. Dr. A. Kuhn[d], E-mail: alexandre.kuhn@hevs.ch
[a]Institute of Chemical Technologies, School of Engineering and Architecture, University of Applied Sciences and Arts Western Switzerland (HEIA-FR), CH-1700 Fribourg, Switzerland; [b]Institute of Chemistry and Biotechnology, Department of Life Sciences and Facility Management, Zurich University of Applied Sciences (ZHAW), CH-8820 Wädenswil, Switzerland; [c]Institute for Chemistry and Bioanalytics, School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland (FHNW), CH-4132 Muttenz, Switzerland; [d]Institute of Life Technologies, School of Engineering, HES-SO University of Applied Sciences and Arts Western Switzerland (HES-SO Valais Wallis), Rue de l'Industrie 19, CH-1950 Sion, Switzerland

Abstract: Digitalization is having an increasing impact on all industrial sectors, including the chemical and biotechnological industries. Aiming for innovative research and development, the Swiss Universities of Applied Sciences play a pivotal role in transferring academic knowledge and know-how to industrial practice. We review selected examples of projects related to the digitalization of processes and bioprocesses at four different institutions across Switzerland. These developments cover the whole spectrum of digital technologies, including big data, connectivity, analytics and automation. They are conducted in close collaboration with industrial partners and aim to support the growth of this important industrial sector.

Keywords: Big data · Bioprocess · Digital technologies · Industry 4.0 · Technology transfer

Digitalization is the use of digital technologies to transform business operations. Here we focus on operations used in the chemical and biotechnological industries, that we refer to as processes and bioprocesses, respectively. Digitalization has been a major theme for several years now, in both private and public sectors. Indeed, ongoing technological progress has opened new, disruptive possibilities. These enabling technologies include industrial internet of things, big data and analytics, cloud computing, advanced automation, digital twins and augmented reality. On the other hand, the idea that digital technologies can translate into significant economic growth has turned 'successful digitalization' into a priority, for economic and political decision makers alike.

With their focus on teaching professional skills, performing innovative applied research and offering high-value services, the Universities of Applied Sciences can play a key role in the ongoing digitalization of the industry. In Switzerland, the chemical, pharmaceutical and biotechnological industries together generate approximately 5% of the gross domestic product (which amounts to about 20% of the Swiss industrial production). For about 10 years, this industrial sector has been the leader of Swiss exports and in 2020 it reached 52% of total exports.[1] What is the current and future role of digitalization in this context?

As a country, Switzerland seems to compare favorably in terms of digitalization. For instance, IMD's World Digital Competitiveness Ranking 2019 placed it fifth in the world in terms of digital competitiveness.[2] This assessment was based on three main areas: 'knowledge', 'technology' and 'future readiness'. Looking closer at individual areas, it is the 'knowledge' area (defined as 'know-how necessary to discover, understand and build new technologies') that scored particularly well. This is confirmed by other studies showing that for instance research on digitalization is particularly active in Switzerland.[3]

Advanced knowledge and know-how, however, does not automatically translate into the widespread implementation of these technologies in industrial production: The Organisation for Economic Co-operation and Development (OECD), in its 2019 Economic survey of Switzerland, assessed that the take-up of new technologies by Swiss firms was around European average.[4] Interestingly, McKinsey Global Institute compared the adoption of digital technologies between different industrial sectors worldwide. The differences were large: Notably, the pharmaceutical industry scored the lowest (score 13), far behind other goods industries like the automotive (score 31) for instance.[5]

Taken together, this suggests that there is a significant potential for further digitalization in the chemical and biotechnological industries in general, and in Switzerland in particular. The Universities of Applied Sciences can play a pivotal role in this regard, keeping up the strong tradition of innovative developments and, in parallel, boosting digital technology transfer. The present article showcases selected digital technology projects at four Universities of Applied Sciences. HEIA-FR has transformed a pilot production hall to demonstrate the deployment of advanced digital technologies in chemical production. HES-SO Valais-Wallis shows the potential of high-throughput DNA sequencing and bioinformatics for developing faster and safer bioprocesses. ZHAW combines cutting-edge analytics, modeling and automation to develop smart bioprocessing solutions. Finally, FHNW uses machine learning to advance the automated interpretation of NMR spectra. Altogether these contributions demonstrate that the field is under very active development, covering the whole spectrum of digital technologies. These tools will prove to be critical in sustaining the continued growth of the chemical and biotechnological sectors.

HEIA-FR, School of Engineering and Architecture, Fribourg

At the School of Engineering and Architecture Fribourg (HEIA-FR), we are attentive to new digital challenges as we strive to meet the needs of the chemical industries. In discussion with our industrial partners, their interest is focused on aspects of predictive maintenance, reduction of production downtime and process deviation detection. To be accepted, the solutions must be integrated with existing equipment and bring rapid gains with a relatively moderate investment. As most production lines already have a good level of automation, the strategy is to use the data already available in the MES (Manufacturing Execution System) with a Big Data approach, add, if necessary, new measurement points by non-invasive Internet of Things (IoT) solutions and develop scalable algorithms using machine learning, deep learning or digital twin tools. Developing a digitization strategy requires interdisciplinary skills. At the HEIA-FR, with teaching departments along with research and development institutes active in the fields of computer science, mechanical engineering and chemistry, we have all the necessary resources to develop a digitalization strategy.

To demonstrate our skills, five HEIA-FR research institutes are currently developing a demonstration platform. The project, led by the Institute of Chemical Technology (ChemTech), aims to deploy digital technologies in our pilot production facility. The goal is to inter-connect users, factory building and process equipment to provide predictive maintenance tools and algorithms for process deviation detection (Fig. 1). Technologically, we use contactless sensors, RFID tags or energy consumption measurements to power our algorithms. On one of our reactors, a digital-twin approach makes it possible to detect a slight deviation in the flow of fluids in a few seconds. The consumption of energy networks (water, vacuum, nitrogen, steam) are monitored by a machine learning algorithm to detect abnormal behavior and identify the source and cause of the deviation. To ensure interoperability between systems, the communication layer is based on the industrial protocol Open Platform Communications Unified Architecture (OPC UA). Our facilities are connected to operators via augmented reality tools to guide the maintenance technician to the source of the problem while providing the necessary procedure to resolve the problem. This demonstration platform is available to all our partners in order to experiment with the possibilities offered by digitalization in the chemical industry.

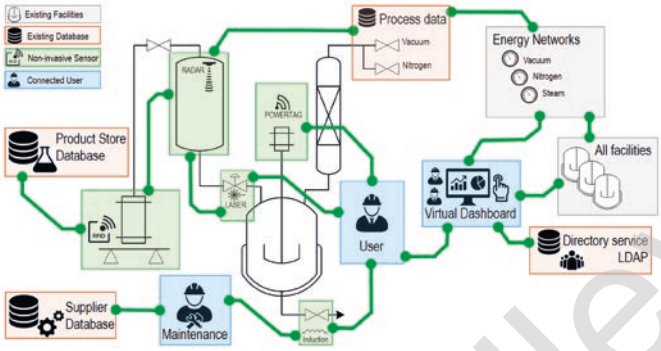

Fig. 1. Digitalization strategy of the pilot production facility of the HEIA-FR.

HES-SO Valais-Wallis, School of Engineering, Sion

Industrial activities in the chemical and biotechnological sectors have a long and rich tradition in the Swiss Canton of Valais. These activities play an important economic role for the region but also at the national level. The Canton has now made digitalization a priority of its socioeconomic development plan.

As a University of Applied Science, one of the missions of HES-SO Valais-Wallis is to foster innovation and boost regional economic activities. Last year, the institution signed a 10-year strategic partnership with the biotechnology company Lonza. The agreement aims at mutually beneficial developments in scientific and educational areas. Importantly, digitalization and smart biomanufacturing represent a key axis of the collaboration. Hence, the initial scientific projects focus on digital biotechnology, specifically in the areas of smart sensors and data science. Beyond this important partnership, data science in general is under very active current development at the School of Engineering of the HES-SO Valais-Wallis, including the Institute of Systems Engineering and the Institute of Life Technologies.

Specificity of Bioprocess Digitalization

Worldwide, the megatrend digitalization is making its way in industrial biotechnological processes. Like in other industries, digitalization can benefit bioprocesses through big data, connectivity, integration, modeling and automation (Fig. 2). A key difference with other industries, however, is the reliance of the biotech industry on biological organisms for production. This represents a great challenge but an equally great opportunity. Indeed, more than four decades into modern biotechnology, our understanding of the organisms used for bioproduction is still too scarce to model them accurately. We cannot predict their response efficiently and this constrains the potential of digitalization in bioprocesses.

The combined action of four driving forces, however, can change the situation: First, high-throughput characterization of cells and genomes, second continued improvements in bioinformatics and modeling, third efficient and precise genetic engineering and fourth, automated development cycles. This paradigm, sometimes referred to as 'engineering biology', is applied for

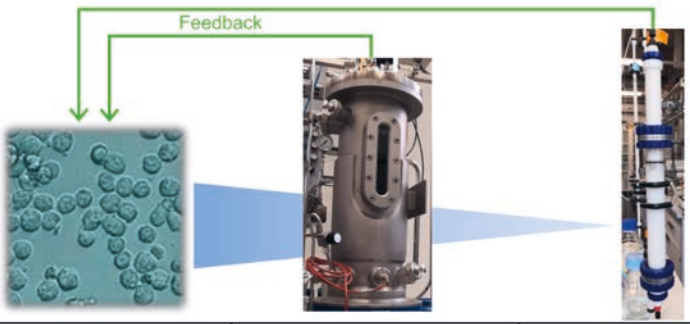

Fig. 2. Digitalization-related opportunities along the three steps of the bioprocess (*i.e.* strain development, upstream and downstream processing). The table lists examples of enabling technologies, sorted according to various aspects of digitalization. The arrows emphasize integration and feedback along the bioprocess: Strains can be developed to fit culture conditions and to facilitate product purification (blue triangle). Conversely, data about upstream and downstream processing can help to optimize strains (green arrows) in order to increase the efficiency of the whole bioprocess.

Digitalization	**Strain/cell line development**	**Fermentation/culture** (upstream processing)	**Product isolation** (downstream processing)
Big data acquisition	Genomics, transcriptomics, metabolomics, proteomics («omics» technologies)	Gaz analyzer, spectroscopic measurements, high-throughput sequencing	At-line and In-line sensors, spectroscopic measurements
Integration and interconnectivity	«Omics» data integration	Heterogeneous data integration, databases	Industrial Internet of Things
Data analysis	Bioinformatics, sequence analysis	Statistical learning, soft sensors	Signal processing, soft sensors
Modeling	Metabolic modeling	Carbon balance, Predictive modeling	Chemometrics, digital twin
Automation	Parallel, small-scale cultures, engineering biology	Design of Experiments, robotic facilities, adaptive control	Real-time product assessment, RFID

strain development but it can be integrated with the rest of the bioprocess: Data acquired during fermentation or downstream processing can be used to iteratively modify the cells, allowing to further optimize bioproduction. We briefly introduce two recent developments showing how big data and data analytics can benefit cell line development and upstream processing in a bioprocess.

Characterization of Cell Lines

As a first example, we are establishing new solutions to efficiently characterize modifications made to the genome of cells used for bioproduction. In industrial biotechnology, the insertion of foreign DNA into the cells remains, to a large extent, an uncontrolled process. The genomic location and the sequence of the integration, however, contribute to the expression level of the product and the stability of the cells. Thus 'transgene characterization' is often a regulatory requirement (for instance in the case of therapeutic protein production). In the long run, better transgene characterization is essential for developing better cell lines and increasing productivity.

Currently, most companies rely on several different experimental methods. These methods generally require a significant amount of manual work and each of them only yields partial information (Fig. 3a). Instead, we apply high-throughput DNA sequencing methods (specifically third generation, 'long read' technologies) and bioinformatic analyses to obtain a complete picture of a transgene integration sites. This includes the genetic architecture of the insertion and its precise location in the host genome (Fig. 3b). As opposed to conventional methods, high-throughput DNA sequencing and bioinformatic analysis can be automated to a large extent. We envision that correlating the genomic features of transgenes with cells' productivity can cut on lengthy cell line selection procedures: It can help to predict high producing lines and fast-track them for further development.

High-throughput DNA Sequencing and Bioinformatics in the Bioprocess

High-throughput DNA sequencing, bioinformatic and modeling have the potential to transform the bioprocess in several important ways (Fig. 4). These methods can be used to not only characterize transgenes but the entire genome of the host. This has allowed to us to develop efficient ways to assess the genetic homogeneity of a cell population, which is a critical parameter of the bioprocess. After a cell line has been engineered, a single progenitor cell is usually isolated and grown into a cell bank. The cell bank is the source of all subsequent production cells. Clonal derivation of the cell bank thus helps to ensure homogeneity of the production cells and reproducibility of the bioprocess.

Based upon data obtained from whole-genome sequencing, we have developed a statistical model that can reliably infer if a cell bank was indeed derived from a single progenitor cell (clonal derivation).[6] Given the technical difficulties and costs associated with clonal derivation of cell banks, this method offers new avenues to streamline and de-risk cell line development. Moreover, we have found that, through cellular cloning, each cell line acquires a unique genetic signature (in the form of a set of naturally occurring point mutations).[6] This signature can be read and, like a barcode, it allows for authentication of the cell bank. This can be performed during a production run or at the end of it. Data from high-throughput sequencing can also be used to sensitively detect contamination (for instance viruses), a major risk of biotechnological production in general.

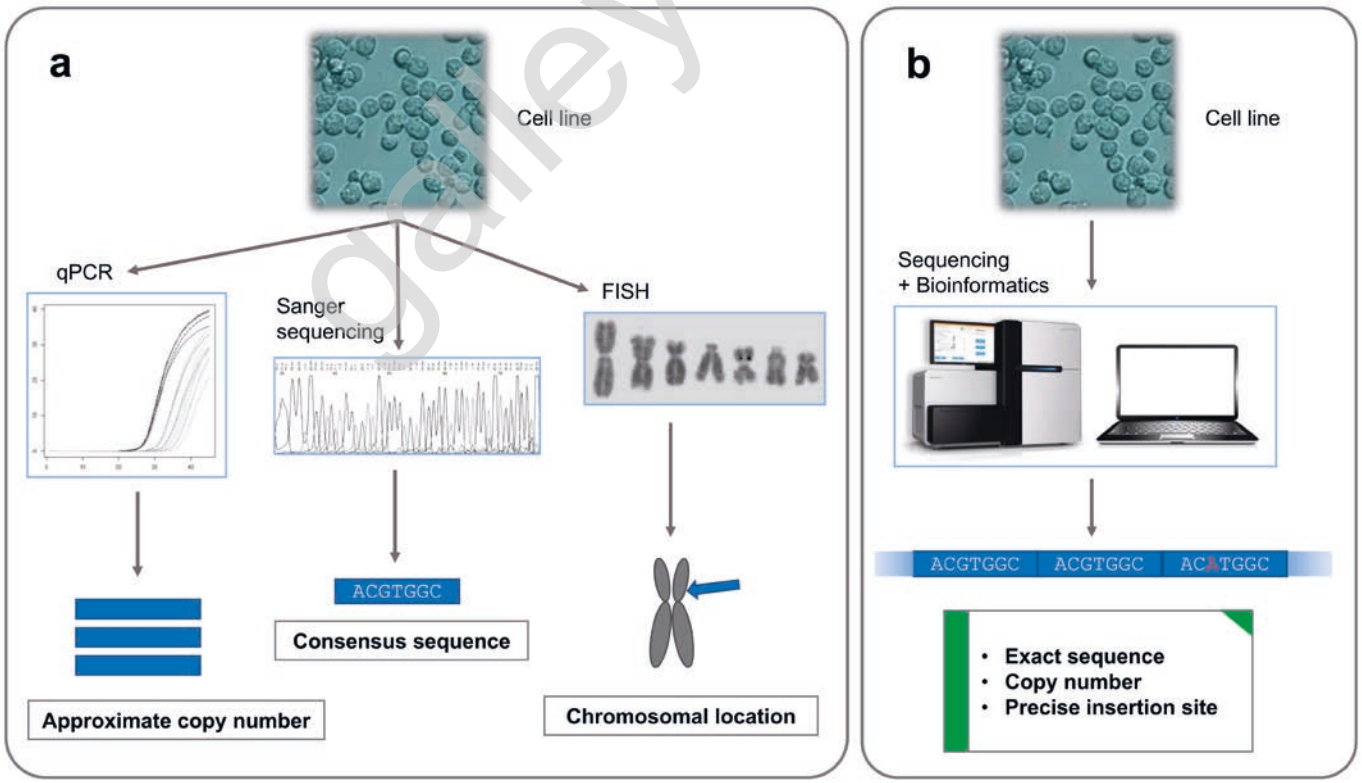

Fig. 3. Characterization of transgene integration sites in a cell line. In animal cell lines for instance, transgenes often integrate as concatemers that are difficult to characterize using standard molecular biology techniques. a. The conventional procedure is based on several methods including quantitative PCR (qPCR) to estimate transgene copy number, Sanger sequencing to establish a consensus sequence and Fluorescence In Situ Hybridization (FISH) to obtain the approximate chromosomal location. b. High-throughput DNA sequencing (including technologies yielding reads up to 100s kb long) and bioinformatic analysis can precisely locate the genomic integration site and fully resolve its architecture (here three head-to-tail transgene copies, for illustration purpose). The rightmost transgene copy contains a G to A mutation (red) that would be undetectable using conventional Sanger sequencing.

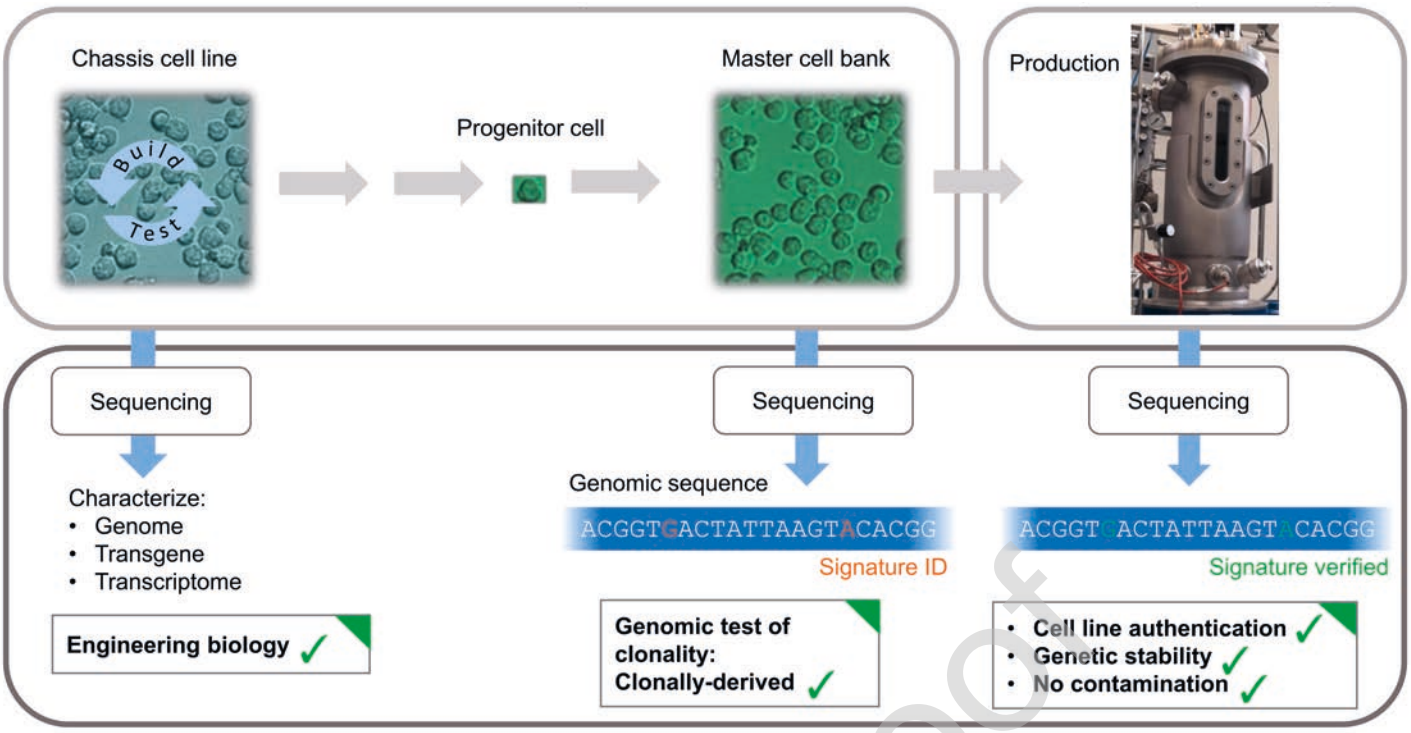

Fig. 4. High-throughput DNA sequencing and big data analytics can provide strong support during cell line development and upstream processing. The illustration shows a simplified bioprocess for a mammalian cell expression system. During cell line development, the cells are engineered iteratively to obtain optimal product expression. High-throughput DNA sequencing (simply referred to as "sequencing" in the figure) allows for genomic and transcriptomic profiling of candidate cells. These methods can support metabolic engineering and the rapid selection of promising cell lines. At the end of cell line development, successful clonal derivation of the master cell bank can be verified based on whole-genome sequencing and data analysis (via the genomic test of clonality). Simultaneously, the algorithm can identify the genomic signature of the cell line (based on naturally occurring point mutations represented by orange bases). During production (or at the end of the run), the signature can be used to authenticate the cell line. High-throughput DNA sequencing can also be used to assess genetic stability of the cells and to detect potential contamination.

Thus, adoption of whole genome sequencing can significantly improve bioprocesses, from cell line development to production efficiency to quality management.

In conclusion, digital biotechnology is a strategic axis of development at the HES-SO Valais-Wallis. Current efforts focus on big data acquisition and smart data analysis, addressing physical and chemical properties of the bioprocess, as well as its biological constituents. The School of Engineering and the Institute of Life Technologies work in close collaboration with industrial partners. As it is often the case with disruptive technologies, the necessary changes implied by digitalization can be intimidating. We thus not only focus on developing innovative solutions but we also pay particular attention to finding ways of lowering adoption barriers and easing transformation.

ZHAW, Department of Life Sciences and Facility Management, Wädenswil

The implementation of digital tools and workflows – further accelerated in the wake of the Covid19 pandemic – today is a dominant topic in the R&D roadmaps of the producing industries. The dynamic progress in this field exceeds the pace of other technological developments, and yet is still in comparatively early stages in sectors with 'classic' nature science orientation. Corresponding to the broad scope of use scenarios and methodical approaches, multiple centres and research groups at ZHAW currently engage in the design, development and prototyping of digital tools and integrated data solutions for robust, performance-optimized and sustainable processes. In the following, we briefly touch on the requirements, technical enablers, and implementation examples of process digitalization in biotechnology and biochemistry, exemplary for two core segments of the Swiss innovation landscape.

Integrated Systems for Enhanced Bioprocess Understanding and Control

Biopharmaceutical laboratories and production plants typically rely on work procedures with a significant share of manual handling steps, from material preparation to sample analysis. Large equipment fleets with stand-alone configuration, different data formats and communication protocols render the implementation of modern *Industrial Internet of Things (IIoT)* or *Industry 4.0* concepts challenging. Moreover, regulatory constrains disfavour changes in existing production setups or processes after validation and approval.

The need to use data more efficiently is evident, and many larger companies run dedicated initiatives to advance the digitalization level in their production lines. *Model-based control*, *soft sensors* and *digital twins* are just some of the omnipresent keywords. The task posed to applied research is to make such tools accessible for a broad user base, including start-ups and small and medium enterprises (SME), *via* infrastructure and software concepts that are tailored to this application domain.[7]

Establishing the technical and methodical framework for bioprocess intensification *via* 'smart' digitalization and automation concepts is the goal of an Innosuisse-supported project between the Bioprocess Technology research group at ZHAW and Securecell AG, Urdorf. It explores the potential of unified monitoring, control and evaluation routines in a broad set of biotechnological value chains, from microbial to mammalian and microalgae cultivations. As one of the key deliverables, a fully integrated development and prototyping ecosystem for digitally enhanced workflows in biotech is created, termed *i2BPLab* (*Intelligent and Integrated BioProcessing Lab*). By bridging gaps between traditionally isolated unit operations, a holistic view on the entire processing chain is provided,

opening the door to 'advanced' methods for data exploitation (Fig. 5).[8]

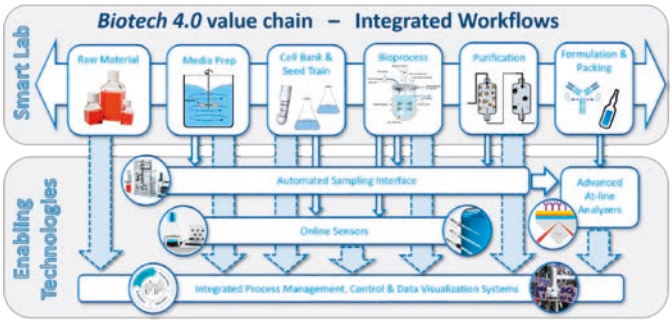

Fig. 5. Schematic representation of workflows, systems and data interfaces in an integrated lab environment for bioprocessing. To exploit the full potential of advanced, digital tools, data should be collected across the full value chain and consolidated in a central depository.[8]

The initial prerequisite to implement digital workflows is to ensure end-to-end data integrity.[9] In bioprocessing, this means that isolated measurement points and continuous signals from preparation phase to preculture data, cultivation runs and downstream purification steps have to be aligned in one central platform. Potent data mining strategies can then be applied to screen for interdependencies, *e.g.* between raw material properties, equipment specifications and product quality attributes. For valorisation in form of improved process guidance, the relevant information has to be extracted from the data stream in a fully autonomous way, contextualized and translated to the correct control operation in (near) real-time. Forward-looking concepts include the implementation of 'intelligent', sampling-on-demand strategies via model-based prediction. The process algorithm decides when an at-line analytical sample is required to maintain a given prediction accuracy, helping to keep manual efforts at a reasonable minimum. The *i2BPLab* tackles these different tasks *via* preconfigured sub-modules in the process management and information system, which allow for rapid adaptation to infrastructure and process layout in use.

A Closer Look at the Process with Real-time Modelling

High computation power and flexible interfaces to programming environments facilitate the inclusion of data-intense simulations and models in real-time control context. Particularly, process variables that are hard to determine my measurement, and previously were approximated from simplified reference experiments, can now be calculated with high spatial and temporal resolution for a better 'view inside the bioreactor'.

One basic example in bioprocesses is oxygen demand, typically expressed *via* the volumetric mass transfer coefficient (k_La) and transfer rate (OTR) as global variables. By coupling population balance models (PBM) with computational fluid dynamics (CFD), the dependency chain from air bubble size dispersion, local oxygen supply and culture physiology can be mapped.[10] *In silico* representations or '*Digital twins*' of the reactor system can be validated against physical experiments with high-precision analytics, such as laser doppler anemometry. Supported by real-time data, they can be run next to the 'actual' process to reflect the dynamics in a culture vessel in real time (Fig. 6). Researchers at the Competence Center for Biochemical Engineering and Cell Cultivation Techniques are exploring how the combination of model-derived, process engineering parameters with biological parameters can be exploited for improving equipment design and culture performance from shake flask to production scale.

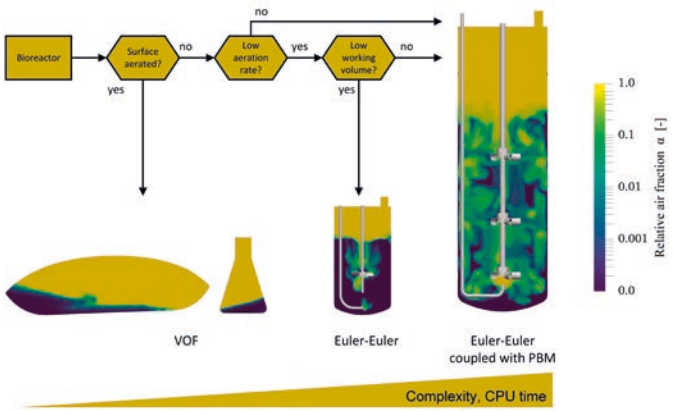

Fig. 6. Flow chart of two-phase models for the calculation of kLa values in bioreactors, with computing effort increasing from left to right.[10]

Digital Tools for Advanced PAT Concepts

Depending on source, noise level and structural complexity of process data, sophisticated algorithms can be necessary for preprocessing and correct interpretation. Spectroscopic data from RAMAN, impedance or other in-line sensors are primary examples.[11] In collaboration with leading suppliers of PAT equipment, experts on Sensor Technology and associated analytical research groups at the Institute of Chemistry and Biotechnology are designing mechanistic models that allow for tracking morphological and metabolic features of reactor content that cannot be covered by classical sensors. To comply with different reference analysis methods available in different process labs, a '*soft sensor suite*' featuring different levels of model calibration and validation has been developed. The user can opt between fully defined (mechanistic) estimators, and different stages of hybrid up to 'black box' models to derive the target variable. Application examples currently under investigation at ZHAW range from photobioreactors for microalgae cultivation to crystallization processes in chemical engineering. In the first example, bias by light irradiance and substantial changes in cell morphology have to be compensated by the model, while in the latter example monitoring becomes possible in high-pressure environments that preclude the use of standard sensor equipment.

Automation as Enabling Technology

Many digital solutions around processes are inherently connected to automation, either in data acquisition/exchange or in hardware operations that shall be triggered. In bioprocesses, it is usually not possible to obtain all relevant information from online sensors, either (i) due to a lack of suitable measurement technologies or (ii) due to bias caused by complex and varying matrix backgrounds. Sampling robots, connected to at-line analysers, may be used to autonomously perform the necessary processing steps and feed data back to the process. Bioprocess Technology at ZHAW has hosted several projects on the development and optimization of such sampling systems (NUMERA, by Securecell AG), resulting in an automated platform of multiple analytical devices, including HPLC, cell counters and *ex situ* sensors. Evolving from prototype stage in 2016, the system is now in routine operation and validated for hygienic (maintaining the sterility barrier) and mechanic (low volume handling) demands in bioprocessing. It delivers data to several of the advanced control concepts implemented in the *i2BPLab* and is constantly expanded by novel methods.

In a similar way, robotic platforms have become inevitable tools to merge *in silico* design strategies with high-throughput, miniaturized processes, *e.g.* in activity screenings of enzymes for directed evolution. The Competence Center of Biocatalysis at ZHAW is interlinking algorithm-assisted enzyme engineering, based on *e.g.* machine learning and Bayesian modelling concepts, with expression assays in fully automated incubation and analysis systems (Fig. 7).[12] Different enzyme classes, ranging from epimerases for antimicrobial peptide synthesis to PETases for plastic degradation, are currently being investigated in the platform.

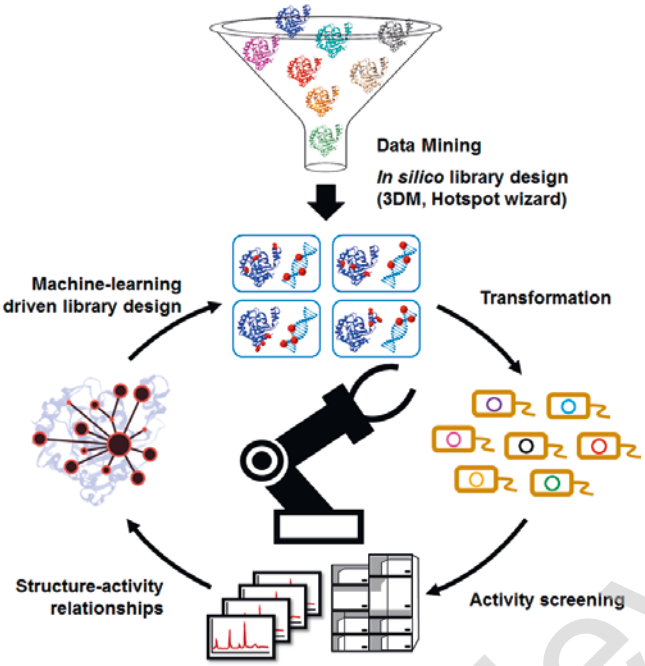

Fig. 7. Automated platforms for algorithm-assisted enzyme engineering with high-throughput activity screening. In iterative optimization, refined enzyme libraries are created by interlinking *in silico* methods with physi-

Data Visualization in Process Environments

The use of digital tools in process environments extends to supporting operators in taking informed decisions in the increasingly automated systems with hard-to-overview data load.[7] For efficient work, the key information on process and/or system status should be easily accessible from any point in the lab or production floor. Augmented reality (AR) applications have in recent years become much easier to implement and can resort to off-the-shelf technical components for building customized solutions. A cross-institutional project, anchored in the Digital Transformation thematic framework program of the Department of Life Sciences and Facility Management evaluates AR-supported visualization tools in bioprocess engineering (Fig. 8). Use cases from process control to system setup and maintenance are covered. Next to visualization technologies, special emphasis is laid on concepts for a context-sensitive differentiation between important and less critical information to be delivered to the operator.

Fig. 8. Concept for AR visualization of bioprocess data on mobile devices. In context-sensitive displays, the most relevant key information is directly accessible.

FHNW, School of Life Sciences, Muttenz

Introduction

A fully automated Nuclear Magnetic Resonance (NMR) workflow for structure elucidation of organic molecules is still one of the more challenging topics for scientists. The current status of NMR workflow digitalization has made enormous progress in the field of metabolomics and quality control. Fully automated high-throughput measurement workflows are established. Automated sample changers[13] and 24 h measurements exist and can be controlled through open source or commercial workflow software.[14–18]

In situations where the molecules are unknown the interpretation of NMR spectra is still mostly manual. Progress has been made to combine machine learning approaches with DFT calculations, which has at least shown in principle that a structure elucidation from first principles could work.

Research work for structure elucidation has shown that an automated workflow for the discrimination of diastereoisomers can be fully automated.[19] We have a research project at FHNW creating a robust workflow for automatic NMR spectrum interpretation.

Workflows in NMR Spectroscopy

NMR spectroscopy is one of the most established and widely used analytical instruments in scientific research. It is complemented by liquid chromatography-mass spectrometry (LC-MS). Contrary to LC-MS the analysis of NMR data is dominated by commercial software packages and it is argued that there is a fairly large amount of user intervention required for data processing.[18] This is the motivation why from time-to-time open-source projects are starting to address certain weaknesses of commercial packages.[18]

A generic workflow for NMR spectroscopy is shown in Fig. 9, which lends itself to a manual analysis of any NMR spectrum or a fully digitalized interpretation.

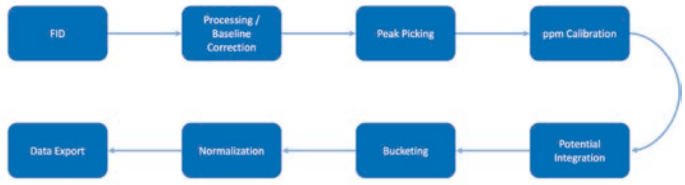

Fig. 9. Generic NMR workflow.

The challenge in high-throughput experiments is to automate or digitalize almost all steps of this generic workflow. NMR spectroscopy is a technology, which exhibits a small signal to noise ratio making certain steps in the pre-processing of the data difficult. This becomes apparent especially in metabolomics where either the concentration of the desired metabolite is low, the acquisition time is limited or the magnetic field is significantly distorted.

A high-quality Free Induction Decay (FID) can therefore not always be assumed. This has an impact on subsequent steps such as identifying peaks. If the signal to noise ratio is low, then it is highly likely that noise or artifacts are selected as peaks resulting in wrong outcomes of the interpretation.

Improving the Signal-to-Noise Ratio

Improving the signal-to-noise ratio (SNR) or denoising of NMR spectra is not a new problem and denoising filters have been applied for quite a while now. Wavelet transformations are one of the dominant tools to achieve substantial noise reductions[20] for FIDs across the board of NMR applications.

Convolutional Neural Networks (CNN) have been used to denoise NMR spectra of metabolites in the brain in order to record ^{1}H-spectra at higher resolution.[21] CNNs are well known for image classifications and can be used for a number of image driven classification tasks such as a peak picking.

Other applications of deep learning algorithms have been in arterial spin labelling, which is prone to low SNR.[22] Here a denoising autoencoder[23] model was used to improve the SNR by 62% and reduce artifacts caused by long measuring times. Denoising autoencoders (DAE) are deep learning algorithms used for handling corrupt data and restore as much information as possible.

It must be noted though that the application of deep learning methods for improving SNR is still in its beginning. The publications show that there are substantial benefits and we expect to see more activity in the near future.

Automated Peak Picking

The identification of peaks in NMR spectra is a research field for almost 30 years.[22] The potential of automated peak picking is recognised as an important aspect of the NMR workflow to build a high throughput pipeline.

As described the improvement of SNR through wavelet transformations is a key component to facilitate an efficient peak picking process. An improved SNR allows a threshold approach, which then doesn't need input by the user. This forms part of the peak picking routine in popular NMR software packages such as CCPN,[24] NMRView,[25] and XEASY.[26]

From a machine learning point of view, peak picking is a classification task. Early classification efforts were made by fitting ellipses to peaks, applying Bayesian statistics[27,28] or neural networks.[29] All methods are based on the fact that peaks have a different topology than noise or artifacts.

It was shown in principle that machine learning is also able to discriminate peaks, but given the lack of computing power and more sophisticated algorithms it was not until 2015[30] that through the application of Support Vector Machines (SVM) peak picking in NMR moved back into the focus for machine learning algorithms.

It is surprising that none of the researched methodologies for peak identification has been widely adopted. Despite that fact that for a chemical shift assignment, for example of macromolecules, a complete list of peaks is not required.

All methods described above are applied to multi-dimensional spectra, which are often used for structure elucidation of larger molecules. They contain more information such as chemical shift and scalar coupling constants between atoms, which in turn is helpful for the application of the methods described above. What seems to be easy for the trained human eye is an inherently difficult task for any type of peak picking algorithm. Often poor SNR, overlapping peak areas, baseline distortions or many other factors introduce noise.

It is therefore very promising that the application of CNNs as described in ref. [30] has achieved for the considered data set a high accuracy of about 90% in identifying peaks correctly. This can be considered as human level accuracy. For the analysis multi-dimensional spectra were used in order to have more data available and a threshold for signal intensity had to be applied.

Research is under way to achieve a similar level of classification for 1D-NMR spectra.[31]

Interpretation of NMR Spectra

Structure elucidation is, after the peak picking process, another difficult to automate task. Currently it is a mostly manual process, which is supported and guided by sophisticated commercial or open-source software. This process is difficult to speed up and a big hurdle for high throughput experiments where the results are not known in advance.

The current approach[32] is often a combination of machine learning together with DFT calculations. With the latter the chemical shifts are approximated and machine learning is then used to assign the predicted shifts against published spectra and data base entries. Usually a correlation coefficient, mean absolute error *etc.* is minimized in order to come to a conclusion about the structure of the molecule. A search in the so-called chemical space of potential structures based on DFT calculations is not only prohibitively time consuming, but also not an elegant approach to this problem.

An interesting approach was taken in ref. [33] to apply machine learning and DFT to determine the stereochemistry of a natural product. Here the principal structure of the product is known, but not its stereochemistry. The number of stereoisomers (or diastereoisomers in this particular case) is known, which allows the problem to be tackled by DFT calculations. The results are very precise and the authors report an impressive 60-fold increase in processing speed.

Current Research at FHNW Muttenz

From a conceptual mathematical point of view the interpretation of NMR spectra is an inverse problem. One solution to the problem is discussed in ref. [19] applying an inverse problem paradigm: If one considers a molecule as a graph, then the NMR spectrum reveals the vertices of this particular graph. The inverse problem to solve is then to find the labelled edges, *i.e.* the bonds, of this graph, given the properties of the vertices.

At FHNW we are following a different path to solve this inverse problem. The NMR spectrum is a manifestation of the desired properties of a particular molecule. The goal is to search for the ideal molecule structure matching the properties expressed in the spectrum. However instead of applying a discriminative model a generative model is applied.

The discussed approaches from above are all discriminative in their character. Such models try to determine the probability of observing a spectrum A given a particular molecule B: $p(A \mid B)$. Generative models on the other hand determine a joint probability $p(B,A)$, which is the probability of observing the spectrum A and the molecule B. This then allows to either determine the direct discriminatory approach by: $p(A|B) = \frac{P(B,A)}{P(B)}$

or for the inverse model $p(B|A) = \frac{P(A,B)}{P(A)}$.

Generative models are more demanding than direct machine learning approaches, but they have made quite some progress over the last couple of years and have been successfully used to design molecules. Currently three generative models are being researched: Variational encoders (VAE), reinforcement learning (RL) and general adversarial networks (GAN).

An additional challenge is the representation of the molecules. NMR shifts are a direct consequence of an atom being more or less magnetically shielded by its direct environment. The molecular representation, acting as input for the generative methods, should therefore carry this information. Currently different representations are being developed ranging from SMILE type of representations to weighted graphs, carrying information about

magnetic moments.

The immediate focus is on the interpretation of 1D-spectra and in order to understand the principles of the generative models ^{13}C-spectra in particular are investigated. The absence of multiplets makes the annotation of the training date simpler and it is hoped to understand more about the required molecular representations.

Summary and Outlook

The digitalization of NMR workflows, *i.e.* the acquisition, processing and interpretation of NMR data, has come a long way and is going to experience a further boost through machine learning. The promise of quick turnarounds for analytical purposes as well as building fully digitalized high-throughput NMR pipelines is very attractive.

Research on 1D-NMR spectra is scarce, but the short time frame to acquire 1D-spectra, *e.g.* sufficient for structure elucidation, is very low. Multi-dimensional spectra take time to acquire and make high-throughput scenarios difficult to achieve. In order to exploit the potential of machine learning further, more high-quality NMR spectra and data in one and higher dimensions must be made available in databases.

Through a combination of DFT calculation of machine learning algorithms it was shown that NMR spectra can in principle be interpreted automatically. The realization that NMR interpretation is an inverse problem can be considered as a step in the right direction. If the need for DFT calculations can be eliminated through a different approach to solve the inverse problem, then a high-throughput process can be most likely established. Applying generative models for NMR spectra interpretation could be the right step to avoid DFT calculations in the future.

Received: May 25, 2021

[1] Swiss Biotech Report **2021**, https://www.swissbiotech.org/report/, accessed May 19, 2021.
[2] IMD World Digital Competitiveness Ranking **2019**, https://www.imd.org/wcc/world-competitiveness-center-rankings/world-digital-competitiveness-rankings-2019/, accessed May 19, 2021.
[3] Digitalisation and research competences in Switzerland, https://www.sbfi.admin.ch/sbfi/en/home/dienstleistungen/publikationen/publikationen-bestellen/s-n-2019-4/s-n-2019-4i.html, accessed May 19, **2021**.
[4] OECD Economic Surveys: Switzerland **2019**, https://doi.org/10.1787/7e6fd372-en, accessed May 19, 2021.
[5] 'Ten insights to get your digitization strategy right', https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/twenty-five-years-of-digitization-ten-insights-into-how-to-play-it-right, accessed May 19, **2021**.
[6] A. Kuhn, V. Le Fourn, I. Fisch, N. Mermod, *Biotechnol. Bioeng*. **2020**, *117*, 3628, https://doi.org/10.1002/bit.27534.
[7] C. Herwig, O. F. Garcia-Aponte, A. Golabgir, A. S. Rathore, *Trends Biotechnol.* **2015**, *33*, 381, https://doi.org/10.1016/j.tibtech.2015.04.004.
[8] L. Neutsch, *Eur. Biopharm. Rev*. **2021**, *2021*, 57.
[9] F. Tao, Q. Qi, A. Liu, A. Kusiak, *J. Manuf. Syst.* **2018**, *48*, 157, https://doi.org/10.1016/j.jmsy.2018.01.006.
[10] S. Seidel, R. W. Maschke, S. Werner, V. Jossen, D. Eibl, *Chem. Ing. Tech.* **2021**, *93*, 42, https://doi.org/10.1002/cite.202000179.
[11] A. S. Rathore, R. Bhambure, V. Ghare, *Anal. Bioanal. Chem.* **2010**, *398*, 137, https://doi.org/10.1007/s00216-010-3781-x.
[12] R. Frey, T. Hayashi, R. M. Buller, *Curr. Opin. Biotechnol.* **2019**, *60*, 29, https://doi.org/10.1016/j.copbio.2018.12.004.
[13] LEAP NMR prepstation, https://www.trajanscimed.com/products/leap-nmr-prepstation, accessed May 24, **2021**.
[14] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, S. Mock, Proc. 16th Int. Conf. Scientific and Statistical Database Management, **2004**, pp. 423-424, https://doi.org/10.1109/ssdm.2004.1311241.
[15] rNMR: Open Source Software for NMR Data Analysis, http://rnmr.nmrfam.wisc.edu/, accessed May 24, **2021**.
[16] J. J. Helmus, C. P. Jaroniec, *J. Biomol. NMR* **2013**, *55*, 355, https://doi.org/10.1007/s10858-013-9718-x.
[17] NMR Software | NMR Technologies, https://www.bruker.com/en/products-and-solutions/mr/nmr-software.html, accessed May 24, **2021**.
[18] C. Beirnaert, P. Meysman, T. N. Vu, N. Hermans, S. Apers, L. Pieters, A. Covaci, K. Laukens, *PLOS Comput. Biol.* **2018**, *14*, e1006018, https://doi.org/10.1371/journal.pcbi.1006018.
[19] E. Jonas, 'Title?', Vol. 32, Eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, R. Garnett, Curran Associates, Inc., **2019**.
[20] S. C. Carl M. Edwards, SPWLA 37th Annu. Logging Symp. **1996**.
[21] H. H. Lee, H. Kim, *Magn. Reson. Med.* **2019**, *82*, 33, https://doi.org/10.1002/mrm.27727.
[22] G. J. Kleywegt, R. Boelens, R. Kaptein, *J. Magn. Reson.* **1990**, *88*, 601, https://doi.org/10.1016/0022-2364(90)90291-g.
[23] P. W. Hales, J. Pfeuffer, C. Clark, *J. Magn. Reson. Imaging* **2020**, *52*, 1413, https://doi.org/10.1002/jmri.27255.
[24] S. P. Skinner, R. H. Fogh, W. Boucher, T. J. Ragan, L. G. Mureddu, G. W. Vuister, *J. Biomol. NMR* **2016**, *66*, 111, https://doi.org/10.1007/s10858-016-0060-y.
[25] B. A. Johnson, in 'Protein NMR Techniques', Humana Press, pp. 313, https://doi.org/10.1385/1-59259-809-9:313.
[26] C. Bartels, T. Xia, M. Billeter, P. Güntert, K. Wüthrich, *J. Biomol. NMR* **1995**, *6*, 1, https://doi.org/10.1007/bf00417486.
[27] D. S. Garrett, R. Powers, A. M. Gronenborn, G. M. Clore, *J. Magn. Reson.* **1991**, *95*, 214, https://doi.org/10.1016/0022-2364(91)90341-p.
[28] A. Rouh, A. Louis-Joseph, J.-Y. Lallemand, *J. Biomol. NMR* **1994**, *4*, 505, https://doi.org/10.1007/bf00156617.
[29] C. Antz, K.-P. Neidig, H. Kalbitzer, *J. Biomol. NMR* **1995**, *5*, https://doi.org/10.1007/bf00211755.
[30] P. Klukowski, M. J. Walczak, A. Gonczarek, J. Boudet, G. Wider, *Bioinformatics* **2015**, *31*, 2981, https://doi.org/10.1093/bioinformatics/btv318.
[31] S. B. Federico Paruzzo Youssef Janjar Bjoern Heitmann1, C. Bolliger, Bruker BioSpin 0920 T181072, **2020**.
[32] M. W. Lodewyk, M. R. Siebert, D. J. Tantillo, *Chem. Rev.* **2011**, *112*, 1839, https://doi.org/10.1021/cr200106v.
[33] S. G. Smith, J. M. Goodman, *J. Am. Chem. Soc.* **2010**, *132*, 12946, https://doi.org/10.1021/ja105035r.