Virtual Laboratories: Transforming research with AI

Arto Klami[1,3], **Theodoros Damoulas**[2,4], **Ola Engkvist**[5,6], **Patrick Rinke**[1,7], **Samuel Kaski**[1,2,8,9]

[1]*Finnish Center for Artificial Intelligence FCAI,* [2]*Alan Turing Institute,* [3]*Dept. of Computer Science, University of Helsinki,* [4]*Depts. of Computer Science and Statistics, University of Warwick,* [5]*Molecular AI, Discovery Sciences, R&D, AstraZeneca,* [6]*Dept. of Computer Science and Engineering, Chalmers University of Technology,* [7]*Dept. of Applied Physics, Aalto University,* [8]*Dept. of Computer Science, Aalto University,* [9]*Dept. of Computer Science, University of Manchester*

***Abstract*—New scientific knowledge is needed more urgently than ever, to address global challenges such as climate change, sustainability, health and societal well-being. Could artificial intelligence (AI) accelerate the scientific process to meet global challenges in time? AI is already revolutionizing individual scientific disciplines, but we argue here that it could be more holistic and encompassing. We introduce the concept of *virtual laboratories* as a new perspective on scientific knowledge generation and a means to incentivize new AI research and development. Despite the often perceived domain-specific research practices and inherent tacit knowledge, we argue that many elements of the research process generalize across scientific domains, and that it is possible to build a common software layer that serves different domains and provides AI assistance. We outline how virtual laboratories will make it easier for AI researchers to contribute to a broad range of scientific domains, and highlight the mutual benefits virtual laboratories offer to both AI and domain scientists.**

1 Introduction

Merriam-Webster defines a *laboratory* as *"a place equipped for experimental study in a science or for testing and analysis"* or more broadly as *"a place providing opportunity for experimentation, observation, or practice in a field of study"* [1]. The definition refers to a physical environment that exists for the purpose of making new discoveries. While laboratory tasks are now frequently carried out on computers, or on more and more automated synthesis and measurement devices, the laboratory itself remains surprisingly similar to its 19th century form. In our increasingly digital world, we think it is time for a paradigm shift to *virtual laboratories* (VLs).

[1]https://www.merriam-webster.com/dictionary/laboratory (28 May 2022)

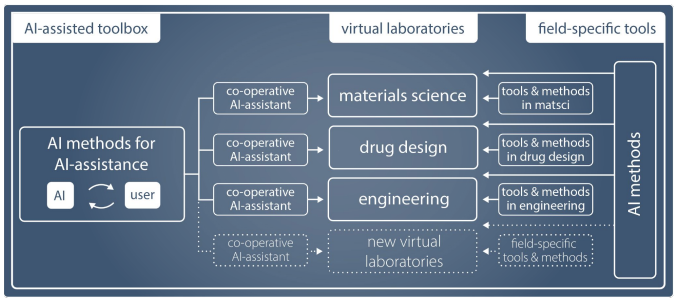

Fig. 1: AI methods enable generalizing across field-specific virtual laboratories, each using a mixture of field-specific and general methods.

The starting point for a virtual laboratory are the computational methods and tools that are already an integral part of modern scientific practices. These include computational simulations, digital twins of various instruments, robotic measurement devices, and methods for experimental design, data analysis and statistical estimation. In most scientific disciplines, physical laboratories already heavily use these computational tools, and research combines computation and real-world experiments. The new digital technologies already provide scale-advantages and improve reproducibility and reliability.

In this perspective, we argue, however, that the current tools are not yet sufficient for building virtual laboratories, and two aspects need to be addressed. First, the current toolkit needs to be updated. The tools of today are typically field-specific, each designed to address specific narrowly defined tasks, and deployment decisions are still primarily made by researchers almost as if the measurements were still carried out by laboratory scientists. Instead, the tools could be designed to better serve the scientific research process itself, and to offer better assistance, which becomes necessary with increasing workflow, tool and research complexity.

The second step required for reaching the full potential of virtual laboratories is to consider what can be done differently now that the computational backbone exists in different disciplines. Could we develop new types of tools (Fig.1)

by thinking across laboratories, and in particular, could we benefit from advances in AI methodology? If the tools were not developed independently in each field but would instead pool the creativity, ingenuity and resources from a variety of fields, progress would be faster and VLs could become a reality sooner. Such generalization and acceleration is precisely the promise AI-based tools offer.

In this paper we present a vision for AI-assisted virtual laboratories: Digitalization of research and development will move from isolated digital twins to AI-assisted support of the scientific innovation process. In the future, new innovations are made in virtual laboratories, where researchers seamlessly operate with physical and virtual measurements in close collaboration with AI, accelerating the pace and improving the quality of research. The virtual laboratories are supported by a common software library.

Virtual laboratories provide a conceptual frame of reference, and in this paper we outline practical directions for the transition from real to virtual laboratories. This paper is a call for both AI researchers and domain scientists to join forces. Section 2 introduces the main concept of virtual laboratory and outlines the high-level goals and challenges. In Section 3 we present the main actions we think should be taken by different parties. Lastly, we motivate the proposed developments by reviewing the state of emergent VLs in three different fields in Section 4, outlining for instance how drug design is already largely done in a virtual realm but using field-specific tools.

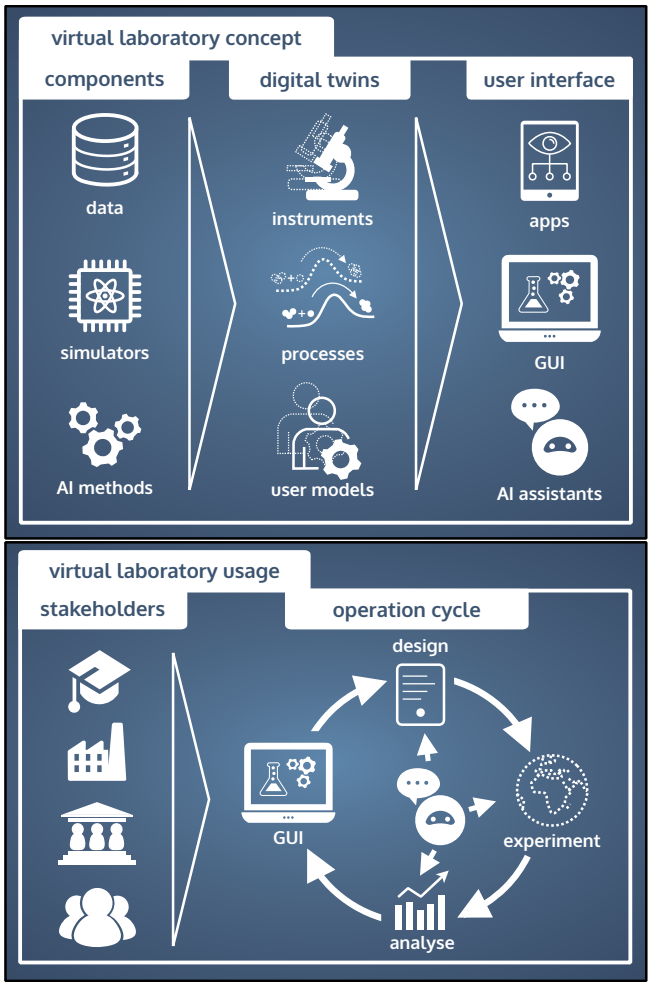

Fig. 2: Top: Elements of a virtual laboratory; bottom: Stakeholders from academia, industry, government and the public interact with the virtual laboratory. Assisted by AI assistants, they design, perform and analyze virtual experiments. The word "experiment" is used as a placeholder here for different functions and features of the virtual laboratory.

2 Virtual Laboratory Concept

2.1 Virtual Laboratory

Following our laboratory definition in the introduction, a *virtual laboratory* (VL) is the *in silico* equivalent of a physical laboratory. A VL exists primarily in a virtual space, or at least mediates the interaction of stakeholders with the VL remotely through a digital user interface. In practical terms, a VL is a collection of interconnected *digital twins* and a digital user interface (see Fig. 2). In our opinion, AI assistance is a critical element of VLs that facilitates navigation of the complex VL environment and enhances the research process.

Digital twins are faithful computational representations of real-world entities or processes [1, 2, 3]. We here consider a wider definition of digital twins than usual and distinguish between three types: a) assets, b) processes and c) human interactions. In a), physical assets is an umbrella term for scientific instruments, measuring devices and equipment that manufacture goods, fabricate materials and synthesise substances. In b), computational models and simulators aim to capture physical or chemical processes. In c), we refer to *user models* of human behaviour and human-machine interactions. Combined, these three types of digital twins transfer real-world data into the virtual realm, where it is processed by simulators and AI methods.

As implied by the word *virtual*, one purpose of VLs is to transfer the experimentation and discovery process from the real into the virtual realm. In this *virtual mode*, users interact with the digital twins instead of their real-world manifestations to derive new knowledge, educate themselves or to receive assistance in complex decision making. This usually offers significant time and resource savings compared to directly operating in a physical laboratory. The digital twins interact with the real-world, when necessary, to stay up-to-date and react to changing conditions. In the *real mode*, the VL has a direct physical outcome, e.g., a material or drug. The VL facilitates, accelerates or even enables the design and development of the physical outcome.

2.2 Elements of Digital Twins

Although each digital twin serves a specific purpose, several aspects are common to digital twins that have already been realized: a) live coupling between the physical asset and its digital twin via multiple streaming data sources originating from live sensing of the physical process, b) access to

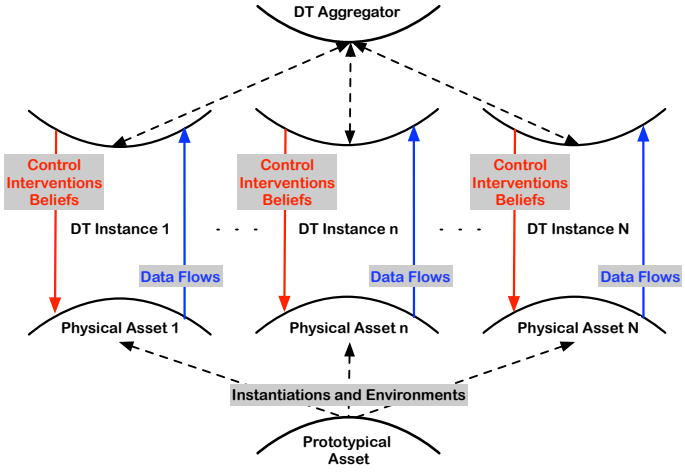

FIG. 3: Schematic of digital twins depicting the key information flow and quantities of interest. Several DTs could be aggregated in a VL. The sub instance could be different instruments combined into one digital twin or different realizations of a device in different labs around the world.

additional information about the modelled process, such as geometry, topology, physical laws or 3-D characteristics for physical assets, c) AI models utilising the aforementioned data sources and prior knowledge to accurately predict or simulate future states of the physical twin from these, d) some ability to perform what-if scenarios and counterfactual reasoning over the process, e) a decision-making mechanism (typically with a human-in-the-loop) for acting on the asset/process given the model and any what-if reasoning abilities therein.

A digital twin can also be composed of several sub digital twins. These sub digital twins could be in different physical locations, e.g. different real-world laboratories. The VL would integrate (or aggregate) all these sub twins into one digital twin as illustrated in Fig. 3. We are now moving to the realm of decentralised/distributed inference and borrowing statistical and causal strength across experiments. Methodological frameworks such as multioutput/multitask learning [4], transportability and data fusion [5, 6], federated learning [7], physics-informed ML [8] and semi-parametric statistics bridging to traditional numerical methods [9] are central for such interconnected VLs.

These interconnections are graphically depicted in Fig. 3 using the gemini symbol (♊) as a playful abstraction of a digital "twin". On the bottom hull the true *data generating process* of the real-world component is generating data that is noisily sensed via multiple sensor networks with potentially different characteristics. These are streamed upwards to the digital twin where inference over a model or a model family is performed with some level of model misspecification that may be estimated. In turn, the posterior beliefs of the digital twin over the model and/or the model parameters are utilised to compute expectations of functions of interest with respect to these posteriors, and subsequent decision-making or simulation of what-if scenarios is performed leading to actionable interventions back on the physical asset.

2.3 A virtuous cycle of VL development

A *virtuous cycle* is a recurring chain of events with a positive outcome. If we can start such a virtuous cycle in VL development, in which advances in AI and domain specific knowledge benefit each other, we can increase the pace, quality and cost-efficiency of scientific research. In this cycle, the VLs are the catalysts for facilitating the interaction and the research environment.

Once kick-started, the virtuous cycle will produce first success stories. These will trigger an increased interest from both AI researchers and VL users, which, as the cycle progresses, should then result in a self-sustaining community effort.

2.4 VL Library – a common software platform

We argue that a key requirement for a successful, virtuous cycle is a common software platform for VLs across fields - a *virtual laboratory library* (VLL). The software platform should be built such that AI advances can be developed independently in a modular fashion and taken immediately into use across all VLs with minimal effort. This platform provides the technical realization of the VLs to produce scientific and societal impact. We believe that the development of the VLL is necessary not only for the permeation of VLs across fields, but also for incentivizing VL developers, VL users and AI researchers to join forces.

3 Towards Real Virtual Laboratories

In this section, we outline the main steps we consider necessary, from a technical perspective, to build VLs and to generalize the concept across disciplines.

3.1 Virtualization

Transferring key components of the scientific method, such as hypothesis generation, experimentation, confirmation and discovery, from the physical to the virtual setting is the central objective of VLs that enables acceleration, reproducibility, and scalability of research. The primary vehicles for virtualization of such components are digital twins of assets, processes and human interactions that are interlinked inside the virtual lab. Significant resources are already being dedicated to improving the quality and versatility of digital twins as components of the VL and the transition will directly benefit from the results of these efforts, but dedicated research will be needed for virtualization of the research process and the human elements of that.

Many key AI technologies and research areas are necessary for the virtualisation process itself: from simulators, emulators, artificial agents and their data calibration and optimisation, to reinforcement learning and robotics for automated measurement devices. Some of these, such as robotics, target the automation of physical measurements while other areas are necessary for exploring and optimising virtual measurements and for counterfactual reasoning.

We note that a large body of AI research including experimental design, Bayesian optimization (BO), reinforcement learning (RL), causal inference (CI), bandits, probabilistic modeling, probabilistic numerics (ProbNum), uncertainty quantification (UQ), and physics-informed ML (Φ-ML) will be central in enabling full virtualization.

3.2 Human in the loop

AI tools are predominantly used to automate tasks and supplement or replace human-derived insight with data-driven models. The evolution towards 'robot scientists' [10] has been invoked, but in reality human scientists remain involved, in two ways. They drive the scientific process, by instantiating, designing and applying AI methods, and they provide knowledge.

Through *human-in-the-loop* machine learning [11], prior human knowledge could be directly integrated into VLs. Human-in-the-loop methods elicit knowledge from human users to maximally improve AI models with minimal user effort.

Current human-in-the-loop methods are not compatible with the other reason humans are involved—that they drive the research process. The current methods treat humans as passive data sources instead of active agents. For VLs, we need to develop human-centric AIs and human-AI collaborations [12]. Multi-agent modelling methods from human-robot interaction [13] are a start, but work is still needed for formulating assistants which are useful to human scientists while leaving them in full control [14]. For this the assistants will need to infer their user's goals and then recommend actions in a way they understand—in other words, they would need models of human users to efficiently collaborate with them. In Section 2, we referred to these models as *user models* or digital twins of human-machine interactions. With user models of scientists, AI assistants will be able to anticipate their actions and aid them in the scientific discovery process.

VLs will be a fitting environment for mixed human-AI research teams. Before fully fledged AI assistants become available, AI tools that give better recommendations would already be beneficial. To reach this point, advances in both AI and human-computer interaction are required.

3.3 Software layer

Building a common software layer for VLs will be critical. We only benefit from up-scaling, if multiple VLs use the same underlying platform, so that AI researchers can easily develop and evaluate their methods for multiple use-cases and VL hosts can easily integrate new VL elements.

The VL software layer mediates the scientific process in the virtual realm and provides the link to the physical realm, but should not be specific to any particular laboratory type. It needs to represent digital twins, moderate data flows between digital twins (as well as their physical counterparts), and enable human-AI collaboration. This requires a modular architecture that communicates with domain-specific databases and models, so that all elements of AI-assistance and DT operation are provided as independent modules.

We are not aware of any general VL software development even though many libraries for the individual VL components and for the automation of data analysis workflows [15, 16] already exist. Besides a modular architecture, an emerging software layer should:

- re-use existing (or future) libraries and avoid replicating any functionality specific to a given domain or general-purpose algorithms (e.g. optimization or machine learning tools).
- run on standard cloud architectures and databases.
- be free and open-source, but licensed to enable commercial support for broad use in industry and research.
- be designed from the start as shared community effort, and eventually establish new standards for information transfer between digital twins and laboratories.
- actively support the FAIR (Findable, Accessible, Interoperable & Reusable) data principles [17], required for community development of the AI elements and reliable operation of VLs.
- support compartmentalization of public and private data and models, so that sensitive data and proprietary simulators and digital twins can be excluded from externalised AI development.
- support extracting data and virtual running environments outside of the VL (so that AI researchers can run them on their own machines) as well as running external algorithms on local computing resources (so that predictions can be evaluated on internal private data and using proprietary models).

3.4 Enabling and encouraging VL research

VLs generate added value from the synergy between AI and research in other domains. To create this synergy, the barrier for contributions from AI researchers and VL domain scientists should be lowered. The common software layer is necessary but not yet sufficient for this. Numerous examples demonstrate a clear benefit from lowering the contribution barrier: ImageNet data [18] revolutionized computer vision and MuJoCo [19] and OpenAI gym [20] reinforcement learning. We need similar success stories for VLs.

A key difference between VLs and the above examples is that VLs are linked also with physical reality and many of the interesting research questions involve humans, as explained in Section 3.2. This introduces additional challenges but we have not identified any immediate show-stoppers that could not be overcome by combining different approaches. Many AI elements can be developed in purely digital laboratories, using simulated human activity if needed. For example, `ChemGymRL`[2] offers a reinforcement learning environment for a purely virtualized chemical laboratory, Trubucco *et al.* provide a virtual environment

[2] https://github.com/chemgymrl/chemgymrl

for design problems [21], and many elements of cognitive models of researchers can be trained with non-experts in crowd-sourcing experiments, for instance models of working memory and decision-making [22].

We believe the most important step towards realizing our vision will be the activation of the research community. Providing computational platforms, theoretical concepts and individual AI modules is a community effort, both in terms of sufficient resourcing, but also to ensure open standards and broad applicability across different fields. This is best achieved by an open initiative for supporting virtual laboratories. The initiative would bring AI researchers and domain scientists together to design and develop the software platform, to determine incentive structures and funding models for VL hosts e.g. by extending the practices currently in place for data releases, and to work towards key standards. The initiative should also thrive to increase awareness of the concept via workshops series organized alongside the leading AI conferences and challenges designed for steering the efforts of AI researchers, motivated by e.g. the effect the Netflix prize had on recommendation engine research [23].

4 Examples

No fully operational VLs exist today in the natural sciences, but significant progress is being made. To make the concept more concrete and to highlight ongoing research and potential outcomes, we discuss three examples from three different scientific disciplines. For ease of communication, we use examples from the authors' research domains, but emphasize that the VL concept is general and applicable from cognitive science to climate research.

4.1 Materials science

While no fully fledged virtual laboratories have emerged in materials science yet, the components are in place. The earliest databases date back to 1965. Their number has risen exponentially since the Materials Genome initiative [24] was launched in the United States in 2011 [25]. Databases evolved via data centers into materials discovery platforms by incorporating data analysis and machine learning tools. The Materials Project [26], the Novel Materials Discovery (NOMAD) laboratory [27] and Citrine Informatics [28] are prominent examples of such materials discovery platforms and could be viewed as virtual laboratory incubators.

Digital twins are more common in engineering and industry (see Section 4.3). They are slowly emerging in materials science, too, with battery development leading the way [29]. Ngandjong *et al.* recently proposed a digital twin of a Li-ion battery manufacturing platform that combines modeling approaches at different scales [30]. Thomitzek *et al.* added a battery cell production digital twin based on digitalization and mechanistic modeling [31]. Regarding scientific instruments, Passananti *et al.* developed a digital twin of a chemical ionization atmospheric pressure interface time-of-flight mass spectrometry (CI-APi-TOF-MS) that facilitates the analysis of molecular cluster formation events in the atmosphere [32].

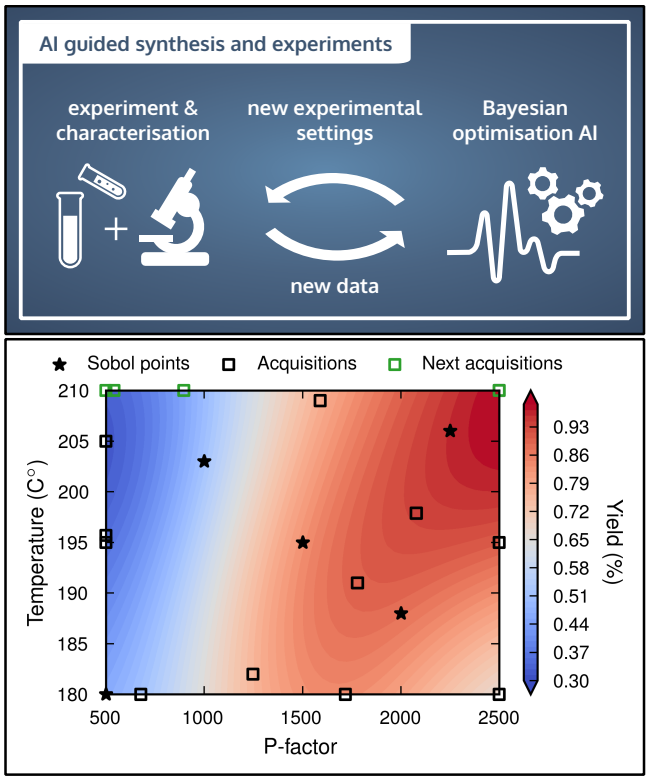

Fig. 4: Top: Conceptual illustration of AI-guided materials synthesis and characterization. Bottom: Biomaterials example, in which AI guided the extraction and characterization of lignin from birch wood with Bayesian optimization. With very few data points (black and green squares and stars) lignin properties (here the yield) can be correlated to the experimental control variables (here temperature and reactor severity (P-factor).

In the Finnish Center for Artificial Intelligence (FCAI), we are developing AI-guided experimentation and synthesis techniques [33, 34]. An example is presented in Fig. 4. A Bayesian optimization based AI requests data from scientists who synthesize and characterize materials. The example shows the extraction of the biopolymer lignin from birch wood and the characterization of the structural properties with 2D nuclear magnetic resonance (NMR) spectroscopy. The data is returned to the AI, which updates its surrogate model of the process and subsequently issues new data requests. The lower panel of Fig. 4 demonstrates that with relatively few datapoints (i.e., time consuming synthesis steps), the lignin yield can be maximized. In addition, the surrogate model provides an insightful visualization to the operating scientists of the relation between the extraction (or synthesis) conditions and the lignin (or materials) properties. Such AI-guidance tools are not only the first step towards autonomous experiments or fabrication, and thus the corresponding digital twins, but they also facilitate the collection of data that has traditionally been difficult to digitize due to its acquisition cost (e.g., human, process or computational time and instrument cost).

Akin to our proposition of a generalized software frame-

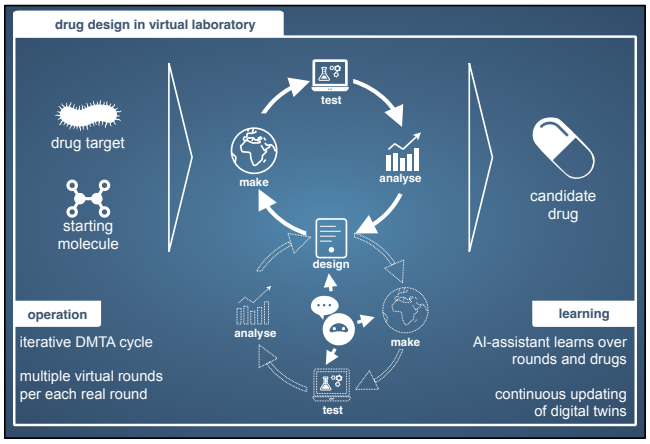

FIG. 5: Drug design is based on iterative cycles of Design-Make-Test-Analyze (DMTA). Within each round, several iterative cycles can be performed in a virtual laboratory (bottom cycle).

work for virtual laboratories, Deagen *et al.* recently proposed materials–information twin tetrahedra (MITT) [35]. The term "digital twin" is here used as an analogy between concepts in materials and information science and does not refer to a component of a virtual laboratory. With MITT they advocate a holistic, data-driven approach to materials science, which we believe could be further extended across scientific domains. In a similar vain, Suzuki *et al.* recently promoted a knowledge transfer from AI applications in pharmaceutical science to materials science through a generalized automated machine learning framework [36].

4.2 Drug design

Applying AI to drug design has become very popular in the last five years triggered by the innovations in AI [37]. Common application areas are molecular *de novo* generation, synthetic route predictions, and molecular property predictions. In drug design a starting molecule with typically poor properties is iteratively optimized until a molecule with properties suitable to start clinical trials is identified. The iterative cycle is usually called the Design-Make-Test-Analyze (DMTA) cycle (Fig. 5) [38].

The virtual drug design laboratory will consist of digital twins for the different components in the DMTA cycle. Several of the necessary digital twins are under development. Digital twins are developed for the design part through deep learning based molecular generation, for the make part through designing synthetic routes by deep learning, and for the test part through developing digital twins for the assays that are used to test the molecules.

An outstanding important research task is to find out how implicit knowledge residing with the scientist can be modelled through human-in-the-loop modelling, so that it can be included in the digital twin of the analysis step. It is important to keep in mind that the virtual laboratory is an approximation of a real drug design laboratory. Virtual molecules are optimised in the virtual laboratory and then actually synthesized and tested in a real laboratory in an iterative manner. An optimal laboratory would combine a virtual laboratory with a fully automated real laboratory. There are several efforts on-going to create autonomous automation systems for synthesizing and optimizing molecules [39]. Thus for drug design, virtual and real laboratories needs to exist in close collaboration, where as good compounds as possible are proposed by the virtual laboratory, the molecules are then synthesized as efficiently as possible in the laboratory, and the resulting data is fed back to the virtual laboratory.

4.3 Data-Centric Engineering

Engineering has recently witnessed a proliferation in data-centric techniques and digital twin development. While the concept and need for virtual laboratories to augment these efforts is in its infancy, we showcase two examples of recent engineering DTs developed at the Alan Turing Institute with academic and industrial partners, demonstrating how the VL concept extends beyond scientific research in natural sciences to design tasks in engineering.

The first one is the world's first 3-D printed steel bridge (MX3D bridge) depicted in Fig. 6 and currently situated in Amsterdam, Netherlands. The various sensor networks on the bridge, such as cameras, accelerometers and load cells, stream live data to its digital twin at the Turing Institute in the UK. The underlying DT model has been developed based on the StatFEM methodology that was recently introduced [9] to formally synthesize observational data and numerical models of its structure.

The second example, from the CROP project[3], is a digital twin of an underground farm in a tunnel situated in Clapham, London, UK. This is a hydroponics system with 2 aisles running in parallel in 23 zones and 2 meters long. Various environmental measurements and camera footage are live-streamed from sensor networks to monitor crop health, forecast yield and future conditions, and optimize all levels of operation including location of crops and environmental conditions. The underlying DT model here utilizes particle filtering for model calibration [40] and data synthesis.

5 Conclusion

We introduced the virtual laboratory concept to amalgamate scientific research and R&D in industry with AI technology and AI assistance. We highlighted the benefits of VLs for both research laboratories and AI researchers, and outlined key requirements of a common software layer and various research directions to proceed towards VLs. In our opinion, VLs are a community effort. To get the movement started, we are currently preparing for formation of an open initiative that brings AI researchers and scientists of other domains together to raise awareness for the VL concept and to work together towards realizing VLs.

[3] https://github.com/alan-turing-institute/CROP

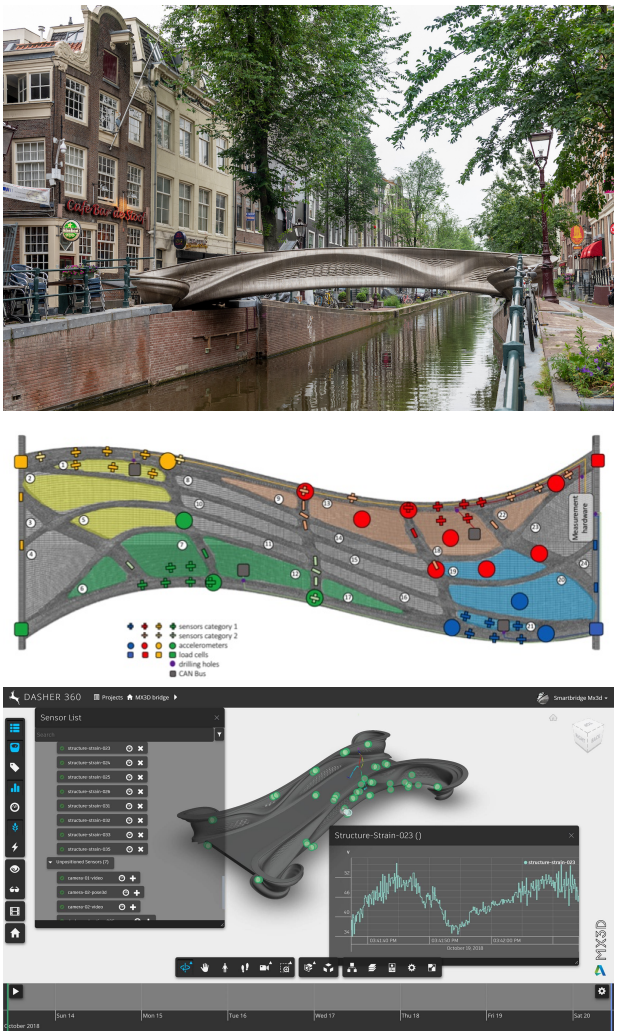

FIG. 6: The 3D-printed steel bridge currently installed in Amsterdam, Netherlands and its multiple sensing arrays that are streaming live data into the corresponding digital twin in The Turing, UK. Images by Joris Laarman Labs, Thea van den Heuvel, MX3D, and AutoDesk Research.

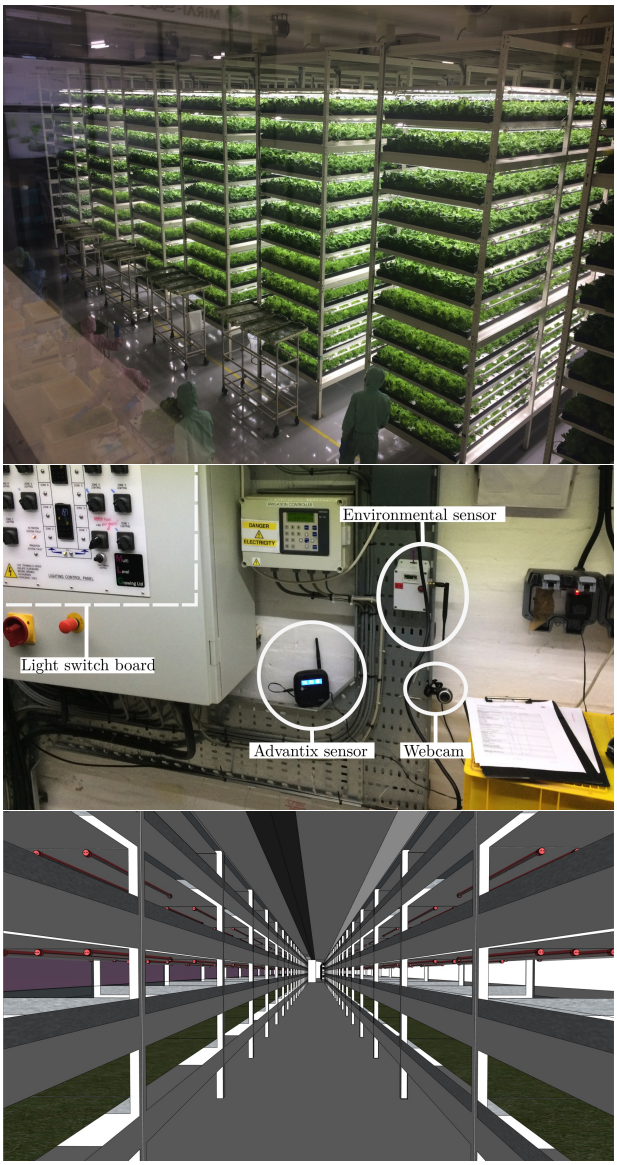

FIG. 7: The underground farm in Clapham, London, and its multiple sensing arrays that are streaming live data flows into the corresponding digital twin in The Turing and the University of Cambridge, UK. Images by Rebecca Ward, Flora Roumpani, and Zero Carbon Farms Ltd.

The goal of transforming research with AI is ambitious and the transformation will not happen fast. The domain scientists are already working towards this direction, as highlighted by the examples in this paper, and hence we conclude our work with words of encouragement for the AI researchers. In short, VLs provide AI researchers with incentives to contribute to the scientific efforts for solving the grand challenges we are facing. It is hard to think of a subarea of AI that would not be useful for VLs, and hence VLs will provide unique opportunities and cross-fertilization already within AI itself. In many areas, from reinforcement learning to constrained optimization and probabilistic modelling, the current techniques are already clearly sufficient for becoming core elements of VLs. In others, such as causal inference and probabilistic numerics, the VLs will provide concrete cases for testing the current solutions and identifying future research directions.

6 Acknowledgements

AK, PR, and SK were supported by the Academy of Finland from the Flagship programme: Finnish Center for Artificial Intelligence FCAI, and from projects 345604 (SK), 341732 (SK), 341589 (PR), 348180 (PR), and 336019 (AK). PR acknowledges support from the COST action CA18234, SK from the UKRI Turing AI World-Leading Researcher Fellowship (EP/W002971/1), and TD from the UKRI Turing AI Acceleration Fellowship (EP/V02678X/1).

References

[1] L. Wright and S. Davidson, "How to tell the difference between a model and a digital twin," *Advanced Modeling and Simulation in Engineering Sciences*, vol. 7, no. 1, p. 13, Mar 2020.

[2] G. S. Blair, "Digital twins of the natural environment," *Patterns*, vol. 2, no. 10, p. 100359, 2021.

[3] S. A. Niederer, M. S. Sacks, M. Girolami, and K. Willcox, "Scaling digital twins from the artisanal to the industrial," *Nature Computational Science*, vol. 1, no. 5, pp. 313–320, 2021.

[4] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[5] J. Pearl and E. Bareinboim, "External validity: From do-calculus to transportability across populations," *Statistical Science*, vol. 29, no. 4, pp. 579–595, 2014.

[6] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[7] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.

[8] Z. Mao, A. D. Jagtap, and G. E. Karniadakis, "Physics-informed neural networks for high-speed flows," *Computer Methods in Applied Mechanics and Engineering*, vol. 360, p. 112789, 2020.

[9] M. Girolami, E. Febrianto, G. Yin, and F. Cirak, "The statistical finite element method (statFEM) for coherent synthesis of observation data and model predictions," *Computer Methods in Applied Mechanics and Engineering*, vol. 375, p. 113533, 2021.

[10] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, and A. I. Cooper, "A mobile robotic chemist," *Nature*, vol. 583, no. 7815, pp. 237–241, 2020.

[11] R. Monarch, *Human-in-the-Loop Machine Learning*. Manning, 2021.

[12] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, "Cooperative AI: machines must learn to find common ground," *Nature*, vol. 593, no. 7857, pp. 33–36, 2021.

[13] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.

[14] M. M. Celikok, F. A. Oliehoek, and S. Kaski, "Best-response Bayesian reinforcement learning with Bayes-adaptive POMDPs for centraurs," in *Proc. AAMAS 2022, 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022.

[15] P. Singh, "Airflow," in *Learn PySpark*. Springer, 2019, pp. 67–84.

[16] F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz *et al.*, "Sustainable data analysis with Snakemake," *F1000Research*, vol. 10, 2021.

[17] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[19] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.

[20] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," *arXiv preprint arXiv:1606.01540*, 2016.

[21] B. Trabucco, A. Kumar, X. Geng, and S. Levine, "Design-Bench: Benchmarks for data-driven offline model-based optimization," 2021. [Online]. Available: https://openreview.net/forum?id=cQzf26aA3vM

[22] H. R. Brown, P. Zeidman, P. Smittenaar, R. A. Adams, F. McNab, R. B. Rutledge, and R. J. Dolan, "Crowdsourcing for cognitive science–the utility of smartphones," *PloS one*, vol. 9, no. 7, p. e100662, 2014.

[23] J. Bennett, S. Lanning, and N. Netflix, "The Netflix Prize," in *In KDD Cup and Workshop in conjunction with KDD*, 2007.

[24] Materials Genome Initiative. Accessed 02/2022. [Online]. Available: https://www.mgi.gov/

[25] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, "Data-driven materials science: Status, challenges, and perspectives," *Adv. Sci.*, vol. 6, no. 21, p. 1900808, 2019.

[26] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, "The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013.

[27] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lüders, M. Oliveira, and M. Scheffler, "Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats," *npj Comput. Mater.*, vol. 3, no. 46, 2017.

[28] J. O'Mara, B. Meredig, and K. Michel, "Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access," *JOM*, vol. 68, no. 8, p. 2031–2034, 2016.

[29] E. Ayerbe, M. Berecibar, S. Clark, A. A. Franco, and J. Ruhland, "Digitalization of battery manufacturing: Current status, challenges, and opportunities," *Advanced Energy Materials*, p. 2102696.

[30] A. C. Ngandjong, T. Lombardo, E. N. Primo, M. Chouchane, A. Shodiev, O. Arcelus, and A. A. Franco, "Investigating electrode calendering and its impact on electrochemical performance by means of a new discrete element method model: Towards a digital twin of Li-Ion battery manufacturing," *Journal of Power Sources*, vol. 485, p. 229320, 2021.

[31] M. Thomitzek, O. Schmidt, G. Ventura Silva, H. Karaki, M. Lippke, U. Krewer, D. Schröder, A. Kwade, and C. Herrmann, "Digitalization platform for mechanistic modeling of battery cell production," *Sustainability*, vol. 14, no. 3, 2022.

[32] M. Passananti, E. Zapadinsky, T. Zanca, J. Kangasluoma, N. Myllys, M. P. Rissanen, T. Kurtén, M. Ehn, M. Attoui, and H. Vehkamäki, "How well can we predict cluster fragmentation inside a mass spectrometer?" *Chem. Commun.*, vol. 55, pp. 5946–5949, 2019.

[33] S.-A. Jin, T. Kämäräinen, P. Rinke, O. J. Rojas, and M. Todorović, "Machine learning as a tool to engineer microstructures: Morphological prediction of tannin-based colloids using Bayesian surrogate models," *MRS Bulletin*, 2022.

[34] J. Löfgren, D. Tarasov, T. Koitto, P. Rinke, M. Balakshin, and M. Todorović, "Lignin biorefinery optimization through machine learning," *ChemRxiv*, 2022, https://chemrxiv.org/engage/chemrxiv/article-details/61370b74b817b46c9a14e7cd.

[35] M. E. Deagen, L. C. Brinson, R. A. Vaia, and L. S. Schadler, "The materials tetrahedron has a "digital twin"," *MRS Bulletin*, Feb 2022.

[36] H. Suzuki, S. Kurosawa, S. Marcella, M. Kanba, Y. Koretaka, A. Tsuji, and T. Okumura, "How AI application in pharmaceutical industries is beneficial to materials science," *Journal of Physics D: Applied Physics*, vol. 55, no. 24, p. 243002, feb 2022.

[37] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today*, vol. 23, no. 6, pp. 1241–1250, 2018.

[38] "Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle," *Drug Discovery Today*, vol. 17, no. 1, pp. 56–62, 2012.

[39] C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison, and K. F. Jensen, "A robotic platform for flow synthesis of organic compounds informed by AI planning," *Science*, vol. 365, no. 6453, p. eaax1566, 2019.

[40] R. Ward, R. Choudhary, A. Gregory, M. Jans-Singh, and M. Girolami, "Continuous calibration of a digital twin: Comparison of particle filter and Bayesian calibration approaches," *Data-Centric Engineering*, vol. 2, 2021.