WHITE PAPER

Alchemite™ – enabling applied machine learning

Overcoming key challenges in ML for R&D to accelerate innovation

intellegens.com | info@intellegens.com
Intellegens, The Studio, Chesterton Mill, Cambridge, CB4 3NP, UK

intellegens

Applied machine learning

Executive Summary

An 80 percent reduction in experimental effort. $millions saved in R&D costs. Getting a product to market six months earlier. Avoiding problems in scale-up of manufacturing processes. Dramatically lowering energy and resource usage or responding quickly to new regulations. These are all proven benefits of machine learning (ML) technologies applied to R&D in sectors such as chemicals, materials, FMCG, and life sciences. But, too often, these benefits are not realised – either because of constraints imposed by the available data or because of implementation challenges. In this white paper, we introduce the Alchemite™ machine learning method and software, discuss how it solves these problems, and provide examples of its success in accelerating innovation.

Contents

1. Machine learning challenges

Machine learning (ML) is the branch of **artificial intelligence (AI)** in which an algorithm builds a model by learning from training data, enabling the model to make predictions or decisions without being explicitly programmed. ML is already delivering on its enormous promise to enable scientific breakthroughs and empower business processes. An Accenture report on the chemicals industry noted that *"usage of AI is advancing, with 61 percent having gone beyond the pilot stage to start implementing machine learning or other forms of AI in their operations, and 91 percent agreeing or strongly agreeing that machine learning-enabled processes help them realise previously hidden or unobtainable value"* [1]. However, significant constraints limit the practical application of ML to R&D in sectors such as chemicals, materials, FMCG, and life sciences. At Intellegens, we focus on overcoming three classes of challenge.

> For 91 per cent of chemical companies, machine learning-enabled processes realise previously hidden or unobtainable value

1.1 High-dimensional, sparse, and noisy data

The first key challenge is the inherent difficulty of solving complex, **high-dimensional problems** from real-world data that is **sparse and noisy**.

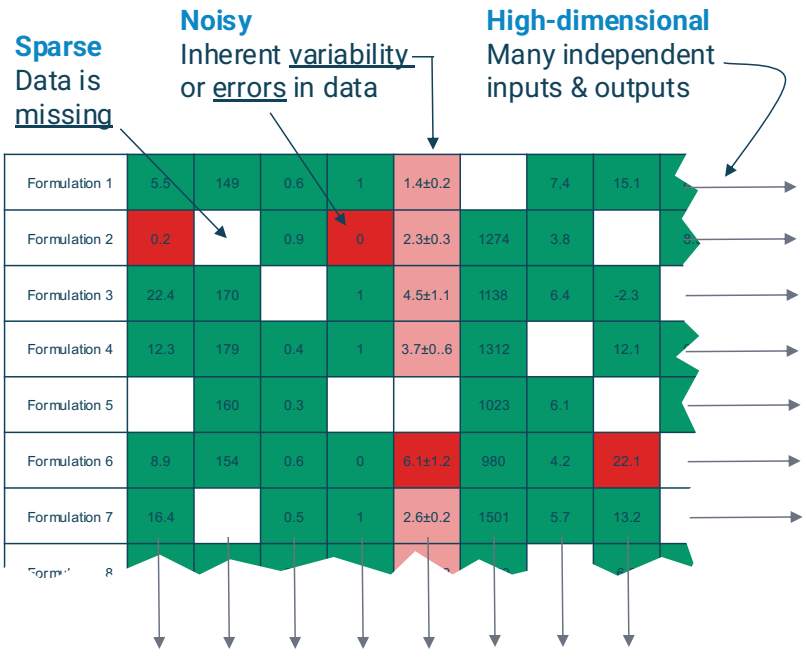


Figure 1. What do we mean by a high-dimensional, sparse, noisy dataset? Conventional machine learning approaches fail when faced with such training data.

Most real experimental or process datasets have *high dimensionality* - that is, they concern systems with many inputs and outputs. Think of a chemical formulation, for which we can vary numerous ingredients and processing in pursuit of an optimal combination of many different properties – things like density, viscosity, cost, stability, and compressibility. For a

manufacturing process, multiple factors such as heat, light, pressure, and time may all be generate the right strength, colour, finish, or environmental impact metrics.

Any high school mathematics student could tell you that solving high dimensional problems requires sufficient data. But what if data is limited? Or if, in assembling all available data, we find that it has gaps in it? This is inevitable in the real world, particularly in a discovery project where the exploration of new ground means there will inevitably be less existing data. In the chemical formulation case, for example, you would almost never set out to test all possible combinations of ingredients and process parameters against all possible properties. A dataset where some of the possible data is missing is *sparse* and **sparse data** is a problem for machine learning. Most ML approaches can't build useful models without a critical mass of data for which all the possible inputs and outputs are complete. They give poor results, and these problems are amplified if the data is *noisy* - i.e., contains errors or natural variability. Again, this is inevitable in much real experimental data.

Can we make ML methods that work for complex problems with real-world data?

1.2 The MLOps challenge

The second key challenge is what we might call the 'Machine Learning Operations' or 'MLOps' challenge.

Up to 80 per cent of data scientists' time is spent in data preparation

Data science teams or individuals with expertise in ML provide businesses with vital insight. But too much of their valuable time (up to 80 per cent according to one survey [2], [3]) is spent preparing data – often to work around problems with sparsity or noise. Or they spend time writing code to handle generic data problems. With better 'off the shelf' algorithms they could focus on adding value for the problems specific to their business. Solving such problems also means close collaboration with domain experts – scientists, engineers, analysts – who are the end users for the results of ML models. Ideally, data scientists would develop and tune models that can then be applied by these colleagues. But the barriers to use for ML among this wider community are often too high, so valuable data science never gets effectively deployed and fails to deliver its full impact.

The barriers to use are too high for many end users

What limits **end users** in benefitting from ML, whether using models from their data science teams or applying ML tools? ML often requires too much *preparation and setup*. For example, avoiding sparse data challenges in materials science may require work to fit that data into existing models that build-in *assumptions* about how parameters are related. Such setup requirements mean many commercial ML tools are hard for non-experts to use, relying on complex user interfaces or

on programming and scripting. And, even where they are applied, these factors mean that the way in which ML methods are used is often not consistent enterprise-wide, leading to problems with reproducing or understanding results.

Can we provide standard ML methods that speed up the work of data science teams and can also be deployed for end users with minimal assumptions, and strong usability? Added benefits of such standardisation would be reduced costs and time in onboarding new team members, and more effective collaboration across and between teams.

1.3 Trust

The final challenge is **trust**. Scientists are used to models that they can inspect, perhaps in the form of an equation. Even very complex analytical models provide the reassurance (sometimes, the illusion) that the user could to 'get their head around' them – or, at least, that someone else has done so. Machine learning models, particularly the most useful ones that model high-dimensional space, are impossible to inspect in this way. Of course, that is their strength. They don't require anyone to spend months in development and validation, yet they capture complex, often non-linear relationships that cannot be described analytically. But scientists understandably view such a '**black box**' sceptically. They want to understand what the model means in physical terms and to be aware of its assumptions and potential biases.

Explainable AI tools are key to building trust

Such reassurance can be provided, and there is much industry focus on '**Trustworthy AI**' [4]. It requires education and informed use of ML methods that are coupled to so-called '**Explainable AI**' tools – analytics and uncertainty quantification capabilities that enable the user to build confidence in the model and understand its limitations. Many ML implementations lack such tools.

2. Commercial implications

The motivation for overcoming these challenges is strong. Here are some industrial problems typical of those encountered by the Intellegens team:

"My competitor could formulate a winning new product tomorrow."

Formulators need to stay ahead in the 'machine learning arms race'

ML is ideal for formulated products (foods, inks, cosmetics, pharmaceuticals, plastics), due to its ability to mine experimental data for optimal ingredient / process combinations. Getting these recipes right has spectacular business impact, as the classic tale of Coca-Cola's failed reformulation shows [5]. ML technology also enables more effective 'reverse engineering' – negating competitive advantage by using computational methods to find new formulations that mimic or improve on a winning product. Formulators need to stay ahead in this 'ML arms race'.

"Experimental programs can cost $10m and take twice as long as we want."

Experiment is expensive and time consuming. Adaptive Design of Experiments (DoE) is a new paradigm in experimental design [6], powered by ML. It guides experimental programs in order to get the same results with far fewer experiments – reductions of 80%+ over standard approaches are feasible in many circumstances.

Adaptive Design of Experiments powered by ML offers the potential to reduce experiment by 80%+

"How do we meet environmental targets and grow market share?"

Among the many reasons that a product may need to change is the requirement to meet new environmental targets or regulations. But changing the content or processing of a product while also making it perform better so as to retain or grow market share is a complex, high-dimensional problem. And there is often limited data relating to the new constraints. Sustainability, recycling, and safety of real-world products are key ML applications.

"50% of our downtime could be avoided by learning from data we already have."

How do businesses maintain complex production processes or networks of assets with limited resources? They often have data available on these processes and assets, particularly with the emergence of the Internet of Things. This data is vital intellectual property (IP), potentially delivering great competitive advantages. But, too often, that advantage cannot be unlocked due to the sparsity of the data. This might be because sensors lose connection or because different devices are recording different outputs. What if ML could use this valuable sparse data to direct maintenance resources more productively?

Too often, the value in existing corporate data cannot be unlocked

"We're losing vital corporate knowledge when our specialists retire or leave"

Much of the competitive advantage for knowledge-based businesses is locked-up in the experience of key staff members: which inputs to vary in order to 'tweak' the outputs of a particular process, for example, or the subtle variations in formulation ingredients sourced from different suppliers. The risk of losing this IP when specialists retire or leave is substantial. Of course, we cannot replace the full value of intrinsic human experience. But ML models can capture many of the key insights developed by experts in a form that can be shared and re-used, as well helping those experts to augment their insights by identifying subtle relationships that they might otherwise have missed.

3. Alchemite™ machine learning

Alchemite™ is a set of advanced machine learning algorithms and tools developed by Intellegens to meet key ML challenges, delivering the benefits discussed above. It originated in the work of Dr Gareth Conduit and his group in the Theory of Condensed Matter Group at the Cavendish Laboratory (Dept. of Physics) of the **University of Cambridge**. The following are some of its key technical characteristics. Together, they offer a unique combination. Alchemite™ solves problems where other ML approaches fail.

Alchemite™ solves problems where other ML approaches fail

3.1 Value from sparse, noisy data

Alchemite™ is based around a powerful, self-consistent, iterative machine learning algorithm that imputes sparse data. The method combines deep neural network techniques with an imputation framework [7] - [11]. Sparse data are input to the algorithm, which generates an initial best guess of the missing values. The full matrix is then reinserted into the algorithm and this process is repeated until the algorithm converges. The approach **enables useful models to be generated even from very sparse data** – it has been successfully applied in drug discovery, for example, where up to 99% of the data are absent in the first instance [10], [12].

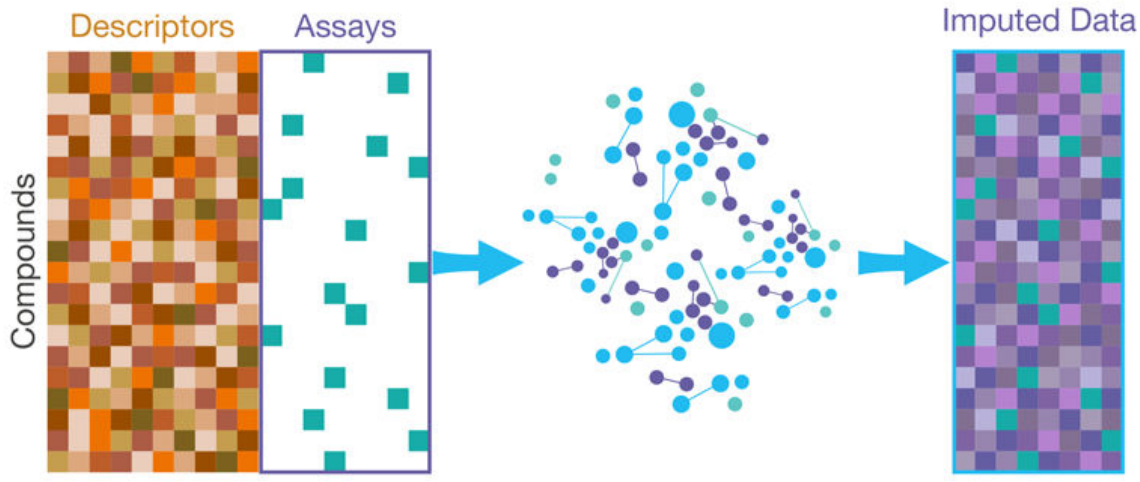


Fig 2. Drug discovery example. Despite the sparsity of the assay (experimental screening) data, Alchemite™ imputes data on compound performance, focusing effort on likely candidates.

3.2 Start fast with minimal assumptions

Because Alchemite™ works by generating a model from the available data that you can then refine and apply, there is minimal need for data cleaning and model setup. This **removes the high dependence on starting assumptions** that is typical of many ML methods.

For example, when predicting material properties, there is no need to specify what type of model to use. The user simply provides inputs and outputs from experiment or simulation and Alchemite™ finds useful patterns in that data [13], [14]. Of course, domain knowledge enables users to interpret results and to supply data that gives the best chance of success. But it is possible to get up-and-running fast. The aim is to support application of domain knowledge, without first requiring users to go a long way towards solving the problem through human insight.

A high dependence on starting assumptions is a key barrier to ML usage

3.3 The big picture – a global view of high-dimensional space

The ability to work across all available data to extract useful knowledge supports a global view. An excellent example is in formulation development. Experiments usually consider ingredients and process parameters separately, measuring different characteristics for each. This means that, when combined, the full dataset is sparse. Similarly, physical or chemical models usually analyse the impact of varying either ingredients or processing. Alchemite™ avoids these constraints and can **consider ingredients and process parameters in a single, efficient study** [15]. It also finds relationships that humans may miss.

Alchemite™ can **generate new solutions that satisfy multiple targets simultaneously** or identify the next experiment that will add the most value to a discovery project. So can many other ML approaches, but Alchemite™ is unusual in doing this quickly even when training data is sparse and noisy, with minimal assumptions. Beyond four or five dimensions, structured exhaustive searches become prohibitively expensive. Alchemite™ explores high-dimensional spaces using a guided Bayesian framework and proposes the most valuable next experiment [6]. This enables Adaptive DoE with its typical 80% savings in experimental cost and time.

3.4 Understand uncertainty to build confidence

A key tool for building trust in machine learning models is the ability to quantify uncertainty in their predictions – not simply to put a value on an unknown property, but to understand the likely range within which that property may vary, given the information held by the model. Alchemite™ offers advanced **uncertainty quantification based on nonparametric probability distributions** for many properties simultaneously [7], [10]. Such methods are more accurate than other approaches that compute uncertainty based on assumptions about probability distributions rather than using distributions actually computed from the data.

Accurate uncertainty quantification is essential for good decision-making

This makes it possible to more reliably understand risk and to make decisions, for example about what experimental route to take, not simply based on the best possible scenario, but on which candidates are *most likely* to succeed, focusing time and effort with a rational supporting business case.

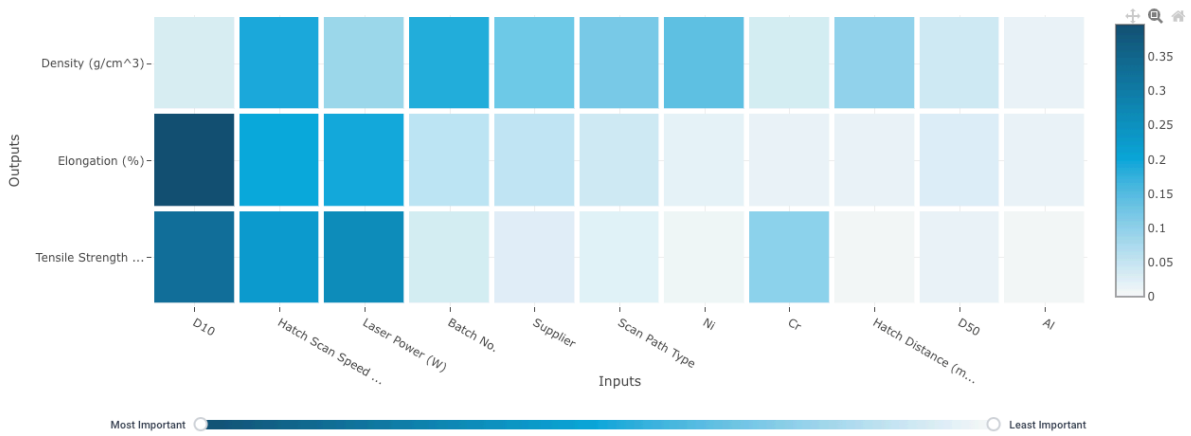


Figure 3. Identifying the importance of relationships in multidimensional space using Alchemite™

3.5 Explainable AI tools

We have seen how high-dimensional spaces and the models that describe them are impossible for humans to understand intuitively. But 'Explainable AI' tools facilitate this understanding. One such analysis tool coupled to Alchemite™ is the **importance chart**, which shows exposes the key relationships that the model has discovered. It shows which input parameters are most important in driving predictions of the output parameters (Figure 3). Alchemite™ provides a range of such analytics. Another is the **sensitivity plot** (Figure 4), enabling users to understand quantitatively the impact of each input on a particular output, identifying opportunities for fine-tuning the inputs; this contrasts with the more global view of the importance plot.

Figure 4. Sensivity plot – how variations in the inputs impact an output.

3.6 Speed and scale-up

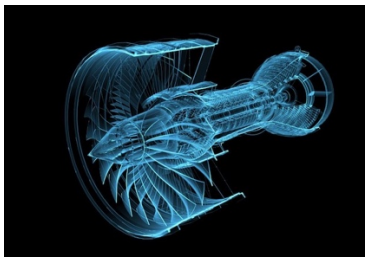
Computations have a light memory and CPU footprint, so Alchemite™ is **lightning fast and able to handle huge databases** [16].

4. Validation and deployment

4.1 Case studies

The Alchemite™ technology is proven for a wide variety of tough applications in the physical sciences, life science, and engineering and has broad applicability to any numerical or categorical dataset or to data (e.g., images, time series, chemical structures) that can be pre-processed into a numerical/categorical format.





Here are some examples:

- The methodology was used in collaboration with **Rolls-Royce** to design a new aero alloy [7], finding a novel solution, experimentally verified to satisfy 11 physical criteria.
- **Domino Printing Sciences** cut key experimental timescales from months to minutes in the reformulation of specialist inks [17].
- Working with **AstraZeneca** and drug discovery partner **Optibrium**, Intellegens modelled pharmacokinetic drug properties, important in translational medicine [18].
- In a project with **Boeing** and the **Advanced Manufacturing Research Centre**, Alchemite™ optimised additive manufacturing processes with fewer experiments [19].
- Food technology leader **Yili** gained valuable insights into its formulations that were not available by other routes, saving time and optimising ingredients [20].
- **NASA** validated the Alchemite™ method for use in design of shape memory alloys and heat exchanger components [21].

For a more on these and other examples, see our separate white papers: *Seven Examples of How Materials & Chemicals Companies Innovate with AI* [13] and *Five ways machine learning can power life science data analysis* [14].

4.2 Using Alchemite™ in practice

The ability of Alchemite™ to function based on minimal initial assumptions removes many of the barriers to usability. Alchemite™ is deployed through two complementary products:

- **Alchemite™ Analytics** is for end users, providing a simple desktop user interface. Users upload data in standard formats (e.g., as a spreadsheet), and can immediately generate useful models from that data. A range of graphical visualisations empower users to interact with the results and refine the model.
- **Alchemite™ Engine** provides advanced API-based access to the computational engine for data scientists. Data science teams can integrate Alchemite™ methods into their own codes, scripts, and workflows and, moreover, fine-tune Alchemite™ parameters for their own data and problems. Models that have been configured for their organisation can then be deployed to end users via Alchemite™ Analytics.

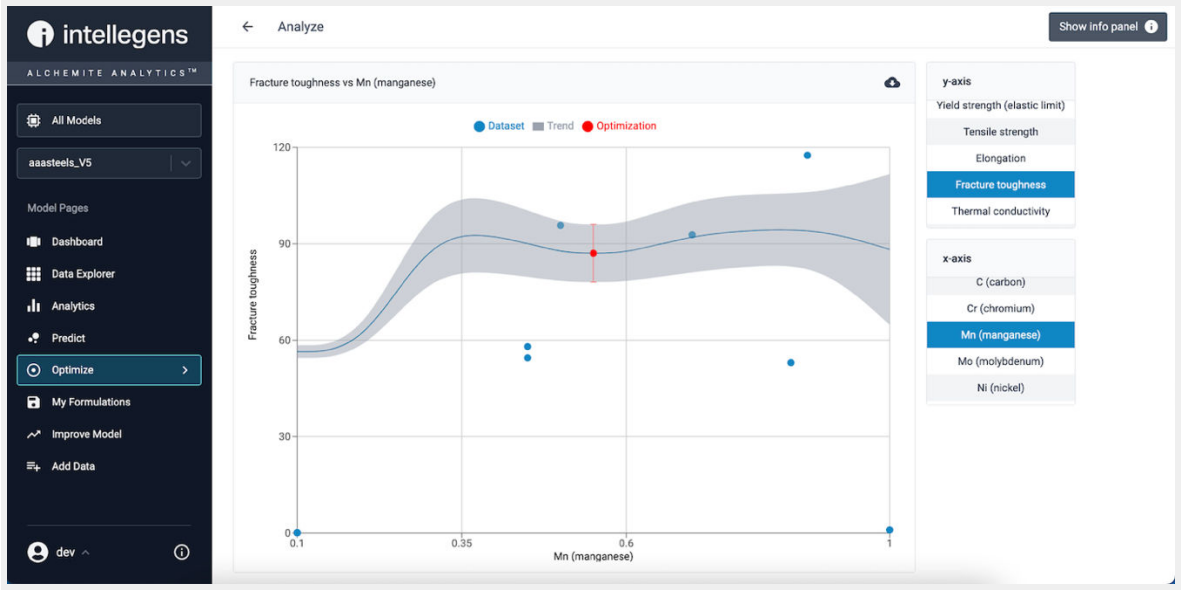


Figure 5. Alchemite™ Analytics visualisation – analysing the results of an Alchemite™ study to determine an optimal materials design.

At Intellegens our aim is to minimise the barrier to getting started with ML. Our **Alchemite™ Success** scientific team works with every customer to get them up-and-running quickly, and to transfer expertise to make client organisations self-sufficient.

5. Conclusions

Alchemite™ from Intellegens is a unique, proven, powerful solution for applying machine learning when solving complex problems using real-world, sparse and noisy data. Alchemite™ is also set up for easy and rapid deployment. We have outlined some of the key challenges that we overcome and provided details of the technical capabilities that meet these challenges – for more information see the references below.

Would you like to get ahead of your competitors by applying Alchemite™ now to optimise products and processes, save time and cost in discovery and development, or break through bottlenecks in data analysis? Get in touch!

References

[1] M. Panjwani, “Process Reimagined,” 2018. [Online]. Available: https://www.accenture.com/_acnmedia/PDF-79/Accenture-Chemicals-Process-Reimagined.pdf.

[2] “2016 Data Science Report,” Crowdflower Inc., San Francisco, 2016. [Online]. Available: http://www2.cs.uh.edu/~ceick/UDM/CFDS16.pdf.

[3] G. Press, “Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says,” *Forbes*, 2016.

[4] Intellegens White Paper, “Trustworthy AI in the chemicals and materials industries,” 2023. [Online] Available: https://intellegens.com/trustworthy-ai-in-the-materials-and-chemicals-industries/

[5] R. Gorman and S. Gould, “This mistake from 30 years ago almost destroyed Coca-Cola,” *Business Insider*, 2015.

[6] Intellegens White Paper, “Machine learning for Adaptive Experimental Design,” 2021. [Online]. Available: https://intellegens.com/machine-learning-for-guided-design-of-experiments-white-paper/

[7] B. D. Conduit, N. G. Jones, H. J. Stone, and G. J. Conduit, “Design of a nickel-base superalloy using a neural network,” *Mater. Des.*, **131**, 358–365, 2017, doi: 10.1016/j.matdes.2017.06.007.

[8] B. D. Conduit, N. G. Jones, H. J. Stone, and G. J. Conduit, “Probabilistic design of a molybdenum-base alloy using a neural network,” *Scr. Mater.*, **146**, 82–86, 2018, doi: 10.1016/j.scriptamat.2017.11.008.

[9] P. C. Verpoort, P. MacDonald, and G. J. Conduit, “Materials data validation and imputation with an artificial neural network,” *Comput. Mater. Sci.*, **147**, 176–185, 2018, doi: 10.1016/j.commatsci.2018.02.002.

[10] T. M. Whitehead, B. W. J. Irwin, P. Hunt, M. D. Segall, and G. J. Conduit, “Imputation of Assay Bioactivity Data Using Deep Learning,” *J. Chem. Inf. Model.*, **59** (3), 1197–1204, 2019, doi: 10.1021/acs.jcim.8b00768.

[11] B. W. J. Irwin, S. Mahmoud, T. M. Whitehead, G. J. Conduit, and M. D. Segall, “Imputation versus prediction: applications in machine learning for drug discovery,” *Futur. Drug Discov.*, **2 (**2), FDD38, 2020, doi: 10.4155/fdd-2020-0008.

[12] B. W. J. Irwin, J. R. Levell, T. M. Whitehead, M. D. Segall, and G. J. Conduit, “Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data,” *J. Chem. Inf. Model.*, **60** (6), 2848–2857, 2020, doi: 10.1021/acs.jcim.0c00443.

[13] Intellegens White Paper, “7 Examples of How Materials & Chemicals Companies Innovate with AI,” 2020. [Online]. Available: https://intellegens.com/materials-chemicals-companies-innovate-with-ai/

[14] Intellegens White Paper, “Five ways machine learning can power life science data analysis,” 2020. [Online]. https://intellegens.com/five-ways-machine-learning-can-power-life-science-data-analysis/

[15] B. D. Conduit *et al.*, “Probabilistic neural network identification of an alloy for direct laser deposition,” *Mater. Des.*, **168**, 2019, doi: 10.1016/j.matdes.2019.107644.

[16] B.W.J. Irwin, T. M. Whitehead, S. Rowland, S. Mahmoud, G.J. Conduit, M. D. Segall, “Deep Imputation on Large-Scale Drug Discovery Data,” *Applied AI Letters* **2**, e31, 2021, doi: https://doi.org/10.1002/ail2.31

[17] Intellegens Webinar: “Efficient formulation design using machine learning,” December 2020. [Online]. Available: https://www.intellegens.com/webinars/.

[18] O. Obrezanova et al, “Prediction of In Vivo Pharmacokinetic Parameters and Time–Exposure Curves in Rats Using Machine Learning from the Chemical Structure”, *Molecular Pharmaceutics* **19**, 5, 1488-1504, 2022. doi: https://doi.org/10.1021/acs.molpharmaceut.2c00027

[19] Intellegens Webinar: “Deep learning for materials development and additive manufacturing,” February 2021. [Online] Available: https://www.intellegens.com/webinars/.

[20] Intellegens Webinar: “Formulation development – a food industry case study,” December 2022. [Online] Available: https://intellegens.com/developing-food-formulations-at-yili/

[21] Intellegens Webinar: “Material and component design with NASA”, September 2021. [Online] Available: https://intellegens.com/material-and-component-design-with-nasa/

About Intellegens

Our mission is to be the leading machine learning solution for real-world, sparse and noisy data problems in industrial R&D and manufacturing processes. Our focus is on making it easy to apply machine learning to accelerate innovation. Alchemite™ originated at the University of Cambridge and development is on-going at Intellegens, in close collaboration with our growing community of Alchemite™ customer organisations in sectors including alloys, additive manufacturing, aerospace, batteries, biotech, ceramics, chemical processes, composites, consumer products, cosmetics, drug discovery, energy, food and beverage, formulated products, paints, plastics, and printing technology.

www.intellegens.com | info@intellegens.com | @intellegensai