intellegens

WHITE PAPER

Alchemite™ powered machine learning with “small data”

Executive summary

Traditional machine learning is well-known for needing a large amount of training data. However, real-life data derived from experiments is not only expensive but also time-consuming to collect, so there is never enough data available. The Alchemite™ deep learning software uses property-property correlations, uncertainty estimates, design of experiments, and validation to overcome inevitable lack of data. Each approach is tried and tested in the real world, backed up by experimental validation of materials and drugs. Alchemite™ allows machine learning to be applied to systems with apparently impossibly small amounts of data, opening opportunities to apply the power of machine learning to systems previously inaccessible due to the lack of data.

Intellegens Ltd., Eagle Labs, Chesterton Road, Cambridge, CB4 3AZ, UK

Challenge

Traditional machine learning approaches require a vast amount of data. However, real-life experimental data is expensive and time-consuming to collect, so it is often inevitable that only a few values are collected. Moreover, discoveries are necessarily in unexplored design space with little historical data. The novel Alchemite™ deep learning software from Intellegens embodies four approaches so that it can prosper when faced by “small data”.

The first two approaches get as much as possible out of data to which you already have access:

1) Exploiting property-property relations,
2) Understanding uncertainty to guide design in systems that are difficult to model.

The second two approaches help you to access more data as efficiently as possible:

3) Design of experiments to judiciously request additional results,
4) Validation of additional data streams.

We detail each approach below, outlining the strategy illustrated by a real-life case study.

Property-property correlations

Strategy

Although data for one crucial property may be sparse, information could be known about another related property, either for the material being studied or from a broader family of similar materials. This often arises when the related property is either cheap to measure or has been recorded historically for a long period of time. When this property is correlated to the prime property of interest, or a first principles computer simulation, Alchemite™ first performs an interpolation from the input variables to the proxy property, and then a second interpolation from the proxy property to the final property. This procedure effectively performs a dramatic extrapolation from the input variables to the final property.

Case study

An example use case was to design a nickel-base alloy for additive manufacturing. Initially, just ten results were available for the crucial property of how many defects are formed on additive manufacturing — far fewer than the 30-dimensional design parameter space of composition and processing variables. However, Alchemite™ was able to exploit a large database of weldability to guide the extrapolation of the analogous ability for additive manufacture. This allowed Alchemite™ to propose a material (see Figure 1), which then had eight properties experimentally verified, including the density of defects formed. More details can be read in the peer reviewed paper *Materials & Design* **168**, 107644 (2019).

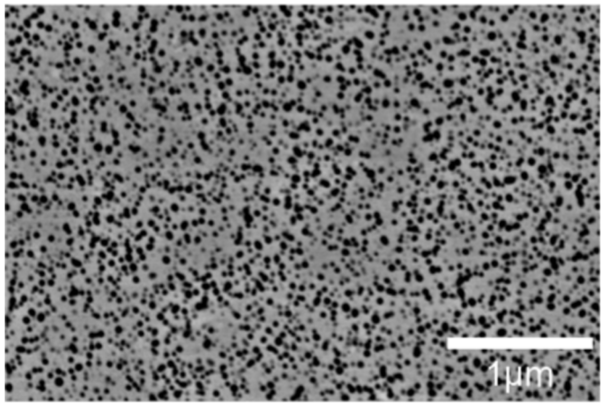

Figure 1: Secondary electron micrograph image of the nickel-base superalloy. The two-phase structure of the alloy can be seen with no visible defects. (Reproduced from Materials & Design 168, 107644 (2019))

Uncertainty

Strategy

When exploring virgin design space it can often be the case that there simply is not enough data available. In this case, although it is difficult to make predictions, Alchemite™ understands how uncertainty in predictions grows when extrapolating predictions to new space. Alchemite™ can then focus on those materials most likely to fulfill targets (see Figure 2), which combine good properties and small uncertainty. Users are pointed toward the material most likely to work when verified, avoiding costly cycles of trial and improvement.

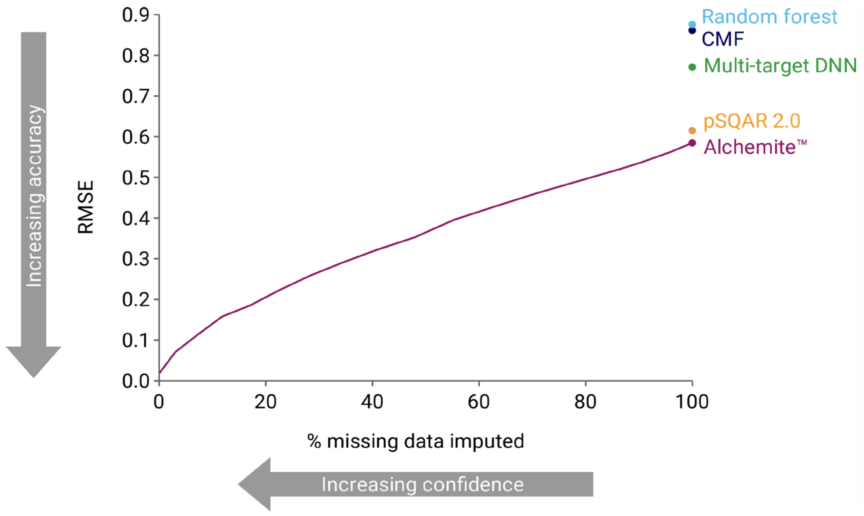

Figure 2: Decreasing average error when uncertainty is used to focus on the most certain predictions. Alchemite™ also delivers more accurate predictions than competing methods, which also cannot quantify their uncertainties. (Data from *Journal of Chemical Information and Modeling*, **59**, 1197 (2019).)

Case study

The Open Source Malaria competition challenged machine learning companies to propose new drugs to fight malaria, which were then synthesised and tested by the organizers of the competition.

Alchemite™ focused on the predictions most likely to be successful, proposing the molecule shown in Figure 3. Of all molecules synthesised, this was the only one that, when measured in an experiment, had activity against the malaria target. [Submitted & ChemRxiv.13194755]

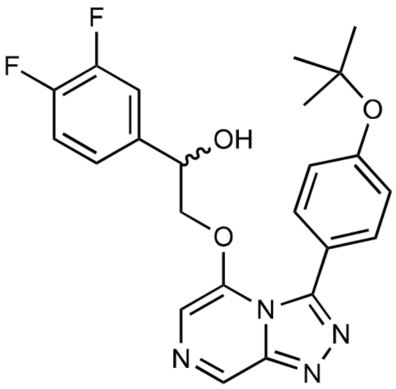

Figure 3: The successful anti-malaria drug designed by Intellegens and Optibrium using Alchemite™. The drug was independently verified to have an activity against *P. falciparum* of 0.46μM. (Reproduced from ChemRxiv.13194755.)

Design of experiments

Strategy

New ground is regularly pioneered by performing additional experiments to understand the landscape. Often these experiments are performed either on a grid, randomly, or distributed following a more judicious algorithm. Instead of starting out from a few rarefied experimental results, building a machine learning model, and then interrogating, Alchemite™ identifies the most valuable additional results to accelerate understanding. For example, Figure 4 shows how Alchemite™ delivers the same accuracy with many fewer measurements than other leading approaches.

Starting with no data

Pioneering studies start out from absolutely no data about the system. Alchemite™ provides generic advice of where to perform those first few experiments in order to gain a broad sweep of the system, before Alchemite™ trains a model based on those first few results to then guide where additional experiments should be performed.

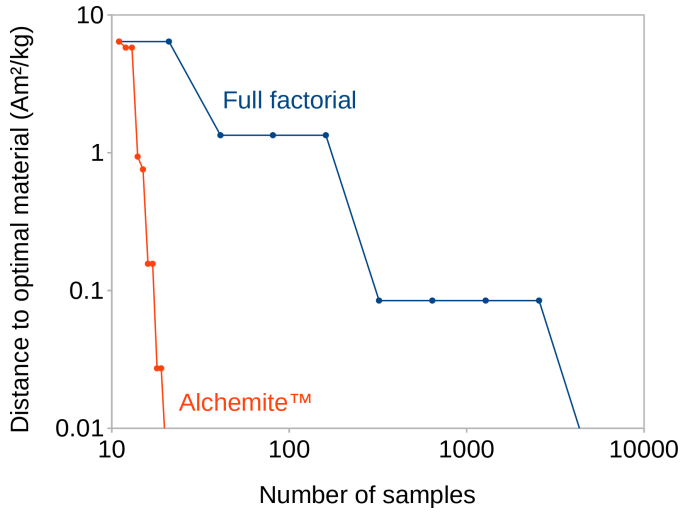

Figure 4: The error of identifying the phase transition in *Nature Communications* **9**, 2925 (2018), with number of samples, with improved performance being at the bottom of the graph. The blue curve is the accuracy when performing design of experiments with a traditional full factorial approach, whereas the red curve is the improved accuracy with Alchemite™. The Alchemite™ guided design of experiments requires fewer results to achieve the same level of accuracy.

Case study

A multinational manufacturer of batteries performs first principles computer simulations of the quantum motion of electrons to help design batteries. The computer simulations are time-consuming, so performing them judiciously improves efficiency. Alchemite™ design of experiments was used to cut the number of simulations required by a factor of 10 to achieve the

same level of modeling accuracy, which was then used to design real-life batteries, including the cells with atomic structure shown in Figure 5.

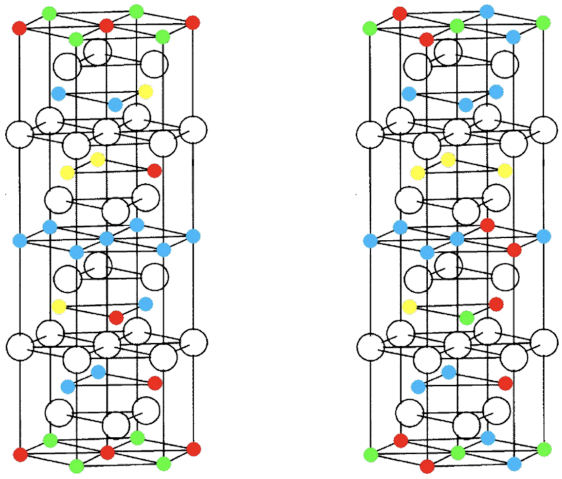

Figure 5: Two of the crystalline atomic structures studied for cathodes within Lithium-ion batteries. The white circles are Oxygen atoms, blue are Lithium, green are Cobalt, yellow are Manganese, and red are Nickel.

Validation of additional data

Strategy

It is sometimes possible to harvest further data from historical sources: either data within the company that has not been digitized, or from public sources. However, as the data is historical the precise approach behind the collection of the data may be unclear, its reliability questionable, and so its trustworthiness needs to be established. Alchemite™ will train a model on an already-established reliable dataset, and then validate each of the entries in the putative additional dataset. The user then has the evidence to assess whether each entry the additional data is trustworthy, and if so accept the values into the master database that can be used to train a better model.

Case study

Intellegens worked with the Granta materials business unit of Ansys to analyse and improve the quality of property data provided within its materials data products. The work is described in the peer reviewed paper *Computational Materials Science* **147**, 176 (2018). Ansys has since selected Intellegens as its machine learning partner for additive manufacturing. Alchemite™ has

also been applied to validate and improve materials data quality by specialist data provider, Matmatch, the case study is available on the Intellegens website.

Key outcomes

- Alchemite™ can train accurate models on **"small data"**.
- Gain access to more data through Alchemite™ guided **design of experiments** or **validation** of historical data.

Future opportunities

Alchemite™ Analytics empowers users to apply machine learning to real-life systems that often have little training data. This opens machine learning to a wide range of companies and applications, allowing them to benefit from the insights and predictive power of machine learning.

About Intellegens

Intellegens has developed a unique deep learning engine, Alchemite™ for training neural networks from the sparse and noisy data typical of real-world science and business challenges. The technique was first developed at the University of Cambridge where it has been used to develop aerospace alloys, guide the design of new drugs, and design next-generation battery technology. The tool is now being used to solve a wide range of industrial customer problems, optimising products and processes, saving time and cost in discovery and development, and enabling breakthrough insights.

www.intellegens.ai | info@intellegens.ai | @intellegensai