intellegens

WHITE PAPER - MCAP Project Summary

Combining machine learning with physics and chemistry models

Executive summary

Physics and chemistry simulation and analytical methods are now standard tools in the development of improved materials and formulations, and the use of machine learning (ML) is increasing. Both classes of method have advantages and drawbacks. In this project, we developed a framework that allows such methods to be combined and proved the effectiveness of such integration with case studies using CALPHAD and molecular descriptors calculated using quantum chemistry methods from SMILES strings. This enabled 'feature engineering' for the ML methods, delivering improvements for measures of model quality when compared to models trained on the original datasets, reflecting the ability to make more accurate predictions. The new descriptors were shown to be important for achieving specific model targets as well as aiding in data visualisation via dimensionality reduction. The developed solution is easily extendable and flexible, and allows the integration of new scientific methods beyond the scope of the project. Future work will focus on further collaboration with partners to improve model performance to help achieve the desired Materials 4.0 roadmap set by the Henry Royce Institute.

Intellegens Ltd., The Studio, Chesterton Mill, Cambridge, CB4 3NP, UK

Introduction

Historically, the design of new materials and formulations has been driven by real-world experimentation and iteration. As materials research continues to evolve and adopt new tools, a paradigm shift is taking place that prioritises the use of digital methods to aid in optimising processes and improving materials. A wide range of simulation and analytical methods are now very well established, applying fundamental physics or chemistry knowledge to compute structure or behaviour. There is also increasing interest in applying machine learning techniques, which do not rely on physical science models, but instead learn from available data to construct their own model of the system being studied. Alchemite™, the novel machine learning (ML) method developed by Intellegens, is one such tool being adopted by industry leaders [1] [2].

The integration of varied predictive modelling tools with real-world data is a key characteristic of the Industry 4.0 vision for digitalisation of the manufacturing sector [3]. The Henry Royce Institute has produced its Materials 4.0 Roadmap [4], which provides a similar vision for the materials community. The Materials Challenge Accelerator Programme (MCAP) is a round of funding hosted by Royce to develop solutions against their roadmaps, including this pursuit of Materials 4.0. This white paper outlines results from an MCAP project, headed by Intellegens, that focused on the integration of machine learning with physics/chemistry methods. Intellegens set out to develop a prototype product that creates a reusable framework for enabling such integration while also demonstrating the effectiveness of combining models for two specific use cases.

The Problem: Limitations of different model types

Physics or chemistry models based on analytical equations or simulation methods often provide researchers with invaluable insight and predictive power, but they are not without their drawbacks. Setting them up and analysing the results can be complex. They may include assumptions that are hidden to the user, or which exclude factors that are in fact important in the system being studied. They can be computationally expensive and time-consuming to run. And, if you need to develop a new method or adapt an existing one for the system you are studying, this can be a major research project in itself.

Machine learning models avoid many of these drawbacks by generating insights quickly from the available data with no assumptions or pre-conceptions. This can be a very powerful tool in the context of material and formulation design. But ML also has its limitations. One constraint is that training ML models tends to fail where the data contains lots of gaps (is sparse) or where the data is noisy. The Alchemite™ method is built to overcome this limitation. It can be trained using sparse and noisy datasets that are typical of real-world experimental and process data [1]. However, in common with all other ML methods, it can only learn relationships from the data provided. Useful relationships that are established domain knowledge may not be detected, or may not be fully exploited, perhaps because the dataset is too small for them to be highlighted or is missing key information [5].

The Solution: Integrating chemistry and physics methods with ML

In this project, we set out to address some of these limitations of physics, chemistry, and ML models by creating a framework to use them in combination. We tested our approach using two different physical science models together with Alchemite™ machine learning. In both cases, we used the physical science models to enrich a dataset, adding additional input data that better describes the system (in machine learning terms, ‘features’) before the Alchemite™ analysis. In the machine learning world, this approach is known as ‘feature engineering’. Note that there are many alternative approaches to combining methods. For example, for a very computationally-expensive physics method, we might first use ML to narrow down a list of candidate systems to those most likely to succeed and then only simulate those.

CALPHAD

CALPHAD (CALculation of PHAse Diagrams) [6] has over 50 years of formula development by experts in the field of alloy manufacturing and can be used to predict thermodynamic information for materials. In our project, the open source pycalphad libraries [7] were used to generate new columns that enriched a dataset by describing the phase transitions for the alloy at different temperatures, pressures, and composition balances.

Molecular descriptors from SMILES strings

A challenge for ML methods in chemistry is how to incorporate information about molecular structure and geometry. One route is to use SMILES (Simplified Molecular Input Line Entry System) [8] strings, which are chains of chemical notation that describe the structure of molecules. In our project we collaborated with Software for Chemistry & Materials (SCM), a leading company in the field of computational chemistry. With the python library PLAMS [9] in their Amsterdam Modeling Suite, we easily create a workflow to generate 3D structures from SMILES strings and calculate descriptors at the electronic structure level with density functional-based tight binding (DFTB) [10], which can be passed on straightforwardly to extend the dataset and train an Alchemite™ model. Once more, this process can embed expert knowledge of the domain into the dataset and create more columns for Alchemite™ models to learn from.

Ichnite™

To accomplish the goal of linking scientific methods with machine learning models, a generic, expandable, and flexible framework was developed by Intellegens. Ichnite™ is a method-aggregation tool, previously used internally at Intellegens, that can pipe datasets through multiple layers of data processing methods. Once the data has been processed, it can be used to train an Alchemite™ model or as part of another workflow. During the project, Ichnite™ has been further developed and integrated with the Alchemite™ Analytics browser-based user interface, creating a tool that could be reused on other projects and beyond the Intellegens team. New scientific methods can be easily slotted into the framework, so custom workflows can be created quickly to meet the needs of any domain.

Aggregating methods improves performance

How do we evaluate performance?

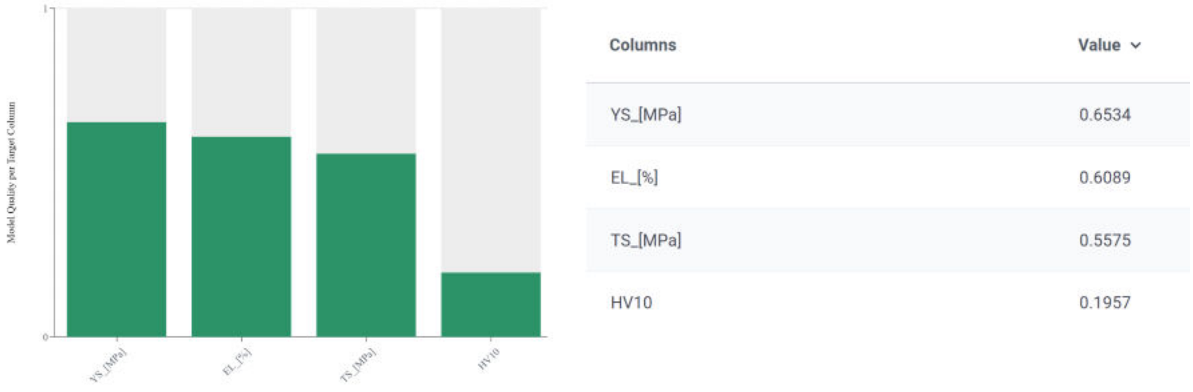
To evaluate the performance of these feature engineering methods on Alchemite™ models, 3 key metrics were considered:

- The coefficient of determination (R^2) [11] of a model reflects the accuracy of its predictions. An R^2 close to 1 means the model is predicting an attribute well, whereas values closer to and below 0 indicate that the model cannot find a useful relationship between inputs and outputs.
- The Importance Matrix graph shows how important Alchemite™ considers one column in the data when predicting another. This is most useful for showing which descriptors are best for predicting which targets; descriptor columns are input columns to a model and targets are the columns to be predicted.
- The Dimensionality Reduction plot is achieved by running Uniform Manifold Approximation and Projection (UMAP) [12] method to reduce a multi-dimensional dataset to a 2-dimensional plot. This can be a powerful tool in understanding trends, as rows considered similar will be grouped together into different 'clusters'.

It was expected that models using feature engineering would have higher R^2 values for model targets, show that the added domain knowledge derived from the use of physical models was important in making new predictions, and show stronger clustering trends compared to the original dataset

CALPHAD Phase Descriptors

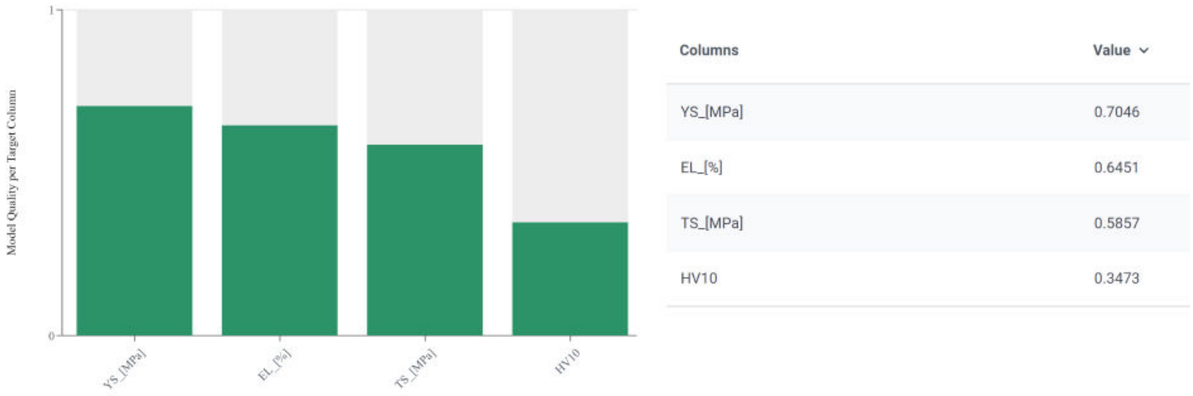
To test out the implementation of CALPHAD into Ichnite™, an aluminium-titanium alloy dataset was used. This contained 2005 rows and 27 columns, of which 23 were model descriptors. These ranged from minor solute additions to different alloy temperatures. The 4 remaining columns were set as prediction targets for Alchemite™: yield strength (YS_[MPA]), elongation (EL_[%]), tensile strength (TS_[MPA]), and hardness (HV10). Figure 1 shows the R^2 metric for each on a model trained on this original dataset. Alchemite™ did well at predicting the top 3 targets, although it struggled to predict alloy hardness.



Columns	Value ⌄
YS_[MPa]	0.6534
EL_[%]	0.6089
TS_[MPa]	0.5575
HV10	0.1957

Figure 1. R^2 for a model trained on the original alloy dataset.

An Ichnite™ chain using CALPHAD and Alchemite™ in sequence was developed to achieve the results in Figure 2. This process generated three new model descriptors: Ti3Al, TiAl and TiAl2. These columns described the state of the titanium-aluminium alloy at different thermodynamic phases. As shown, all model targets increased in predictive accuracy, especially alloy hardness. Alchemite™ is able to leverage the expert knowledge from these columns to make more accurate predictions – the core goal of the MCAP project.



Columns	Value ⌄
YS_[MPa]	0.7046
EL_[%]	0.6451
TS_[MPa]	0.5857
HV10	0.3473

Figure 2. Improved R^2 for the dataset with CALPHAD feature engineering.

Figure 3 shows the Importance Matrix from Alchemite™ Analytics for this feature engineered model, ranked by how important each column is. The x-axis represents the model descriptors and the y-axis represents the model targets. Alchemite™ is mainly using titanium (Ti) and aluminium (Al) to predict the 4 targets, which is expected given they are the main elements of the alloy. When ranked, the CALPHAD descriptors place 5th, 7th and 9th most important. These new descriptors are regarded as very important by Alchemite™ for predicting elongation and tensile strength, and also contribute to the quality of predictions for the other targets from the increase in R^2 across the board.

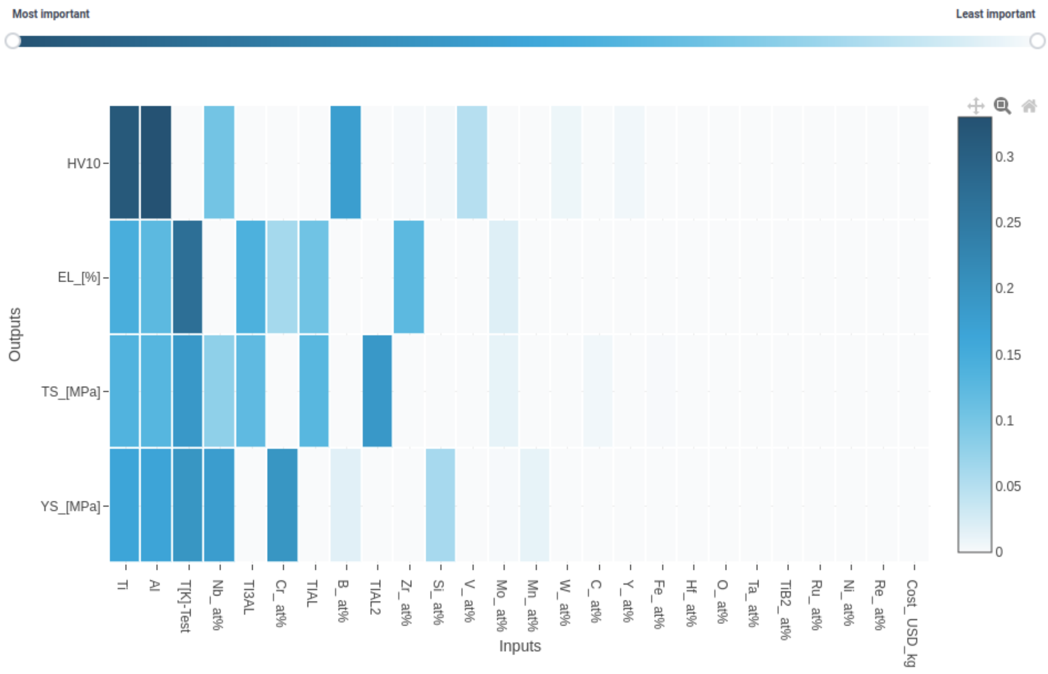
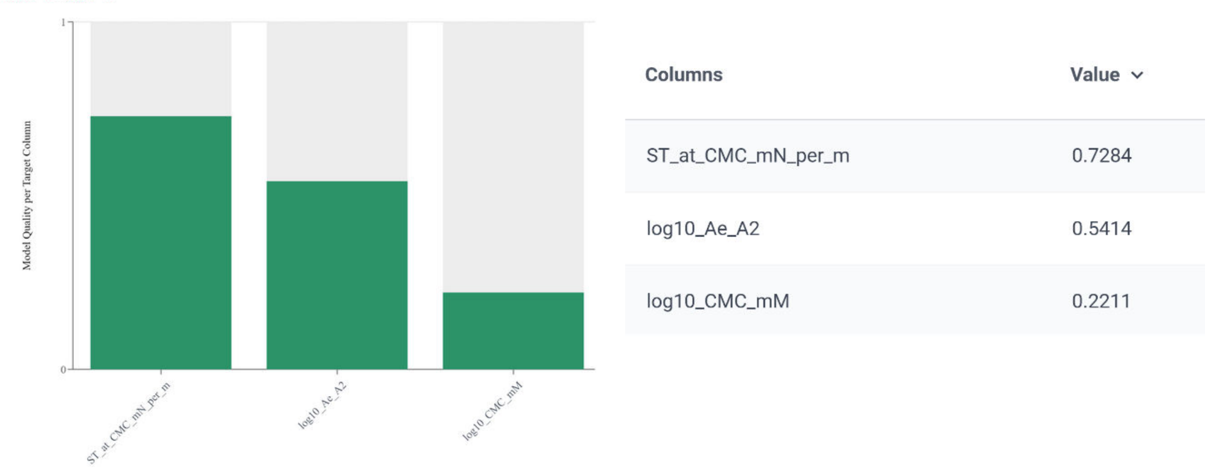


Figure 3. Importance matrix for feature-engineered data.

Molecular descriptors from SMILES

The dataset used for testing molecular descriptors derived from SMILES strings was 350 rows of surfactant data with 7 columns: 2 numerical descriptors, 2 categorical descriptors, and 3 numerical targets. These were surface tension (ST_at_CMC_mN_per_m), the alcohol ethoxylates (log10_Ae_A2), and the critical micelle concentration (log10_CMC_mM).



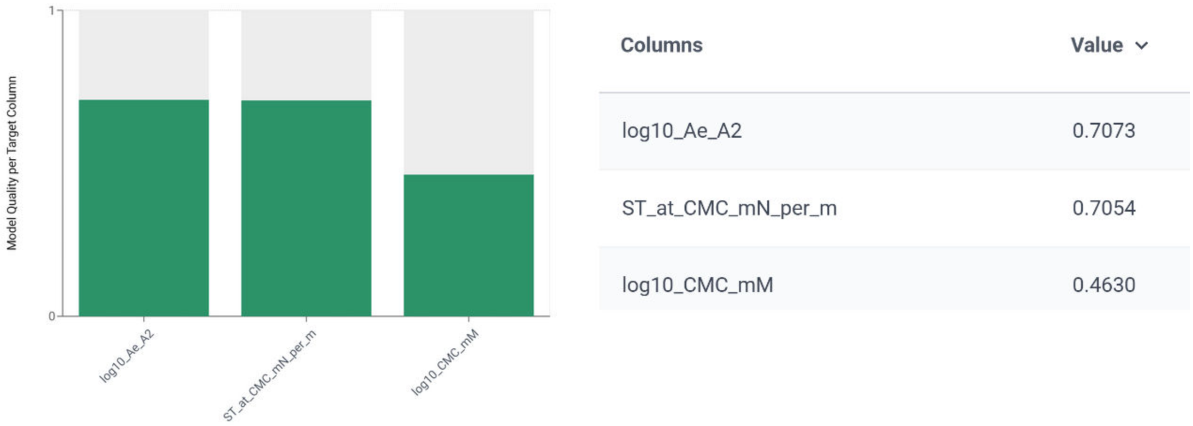
Columns	Value ⌄
ST_at_CMC_mN_per_m	0.7284
log10_Ae_A2	0.5414
log10_CMC_mM	0.2211

Figure 4. R^2 for a model trained on the original surfactant dataset.

Figure 4 shows the performance of Alchemite™ for predicting these molecular targets. The surface tension predictions performed well but the model struggled to predict the other 2 targets, understandably given the dataset lacks information about chemical structure.

We then extended the surfactant dataset with calculated bonding energies and dipole information at the DFTB level. This data is easily generated from SMILES strings with a PLAMS python script in AMS, which makes use of chemical simulation methods such as density function theory [13] and molecular dynamics [14]. The script generated five additional molecular descriptors for the dataset. One new column contained information on the bonding energy for each SMILES string. The remaining four described dipole information. Figure 5 shows the resultant R^2 for each target on a model trained with these new descriptors. The prediction accuracy became a lot more consistent across the model targets, with a large increase on the target with which Alchemite™ had initially struggled.

Figure 6 shows the dimensionality reduction (UMAP) plot for the original dataset reduced to 2 dimensions. The algorithm struggles to identify meaningful patterns, as a large collection of rows have been bundled into one cluster. This reveals no useful information about the surfactants, as niche differences between the data points are not exploited.



Columns	Value ⌄
log10_Ae_A2	0.7073
ST_at_CMC_mN_per_m	0.7054
log10_CMC_mM	0.4630

Figure 5. Improved R^2 for the dataset enriched by PLAMS calculations.

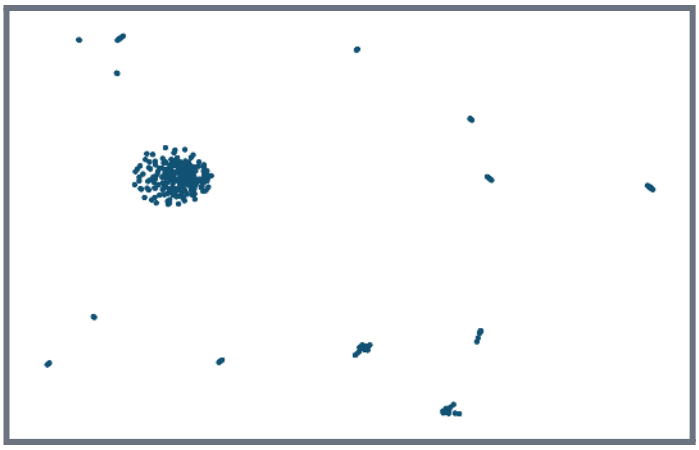


Figure 6. Dimensionality reduction plot for original surfactant dataset.

Figure 7 shows the UMAP plot for the dataset extended with PLAMS. The large cluster has now been broken down into multiple sub-clusters, demonstrating the effect of the new molecular descriptors. The expert information in these new columns has revealed more about each surfactant data point, which has allowed the UMAP algorithm to more specifically group rows of data in this new space.

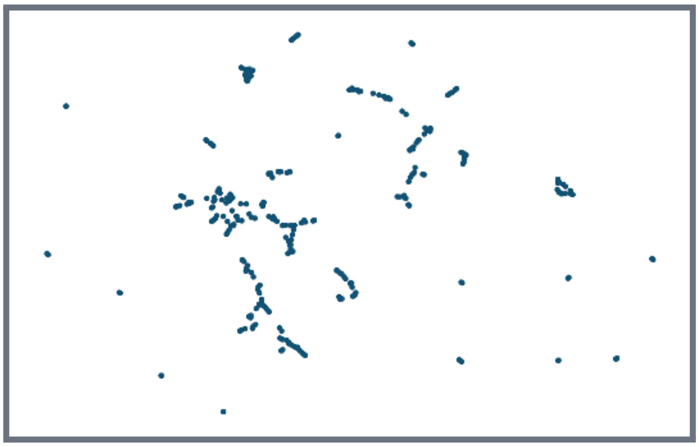


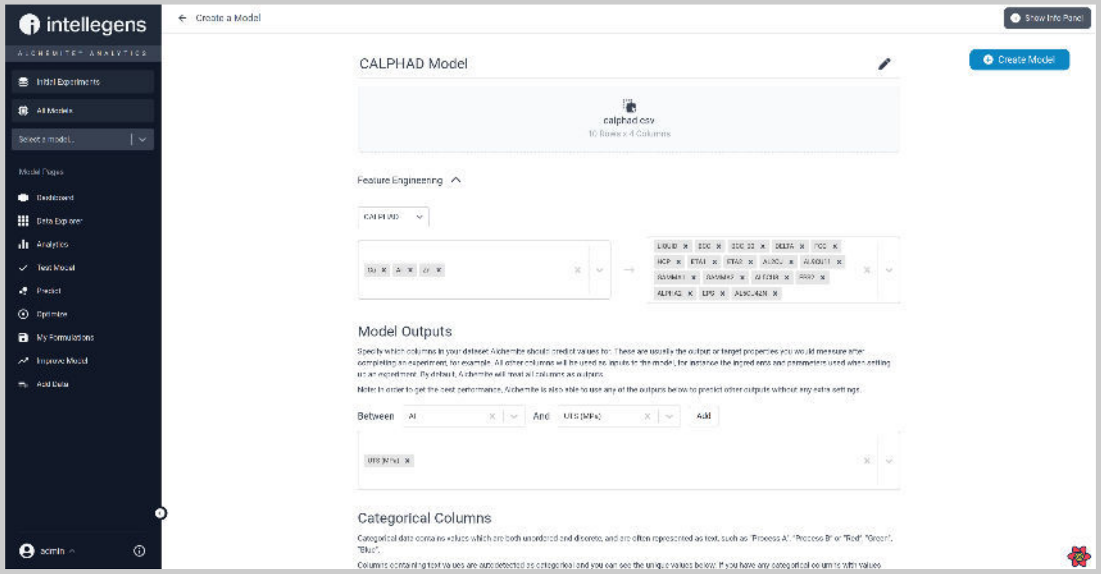
Figure 7. Dimensionality reduction plot for enriched surfactant dataset.

Commercial applications

Feature engineering through the use of physical science methods has proven to improve model accuracy, reveal important relationships between targets and descriptors, and aid in data visualisation. These are all important metrics for a user in understanding their dataset and how it can benefit them in their chosen domain. Experiments and suggestions made by Alchemite™ using these scientific methods in the background will be more reliable and backed by expert reasoning, which will save time and help to identify new solutions in the lab or the factory. The improved ability to extract value from data and make accurate predictions will allow users to achieve their goals faster and more consistently, realising the vision of Materials 4.0.

In our project, we were able to integrate feature engineering seamlessly into the existing Alchemite™ Analytics commercial software product, ensuring that users will be able to benefit from this method integration with minimal effort. Figure 8 shows a screenshot from the prototype delivered by Intellegens for the MCAP project. Users are able to select their engineering method and the required inputs in their dataset. The columns are then validated and the expected outputs shown. It is planned to allow the chaining of these data manipulation methods via this interface, meaning outputs from one method can act as inputs to the next.

Figure 9 shows the 'Predict' page for the prototype. The CALPHAD outputs are tagged with a gear symbol and set as read-only, as their values are generated from user inputs for aluminium (Al), copper (Cu), and zinc (Zn). Alchemite™ Analytics processes the user input live to calculate the phase descriptors for the data, which are then used to make predictions for the target UTS (ultimate tensile strength).



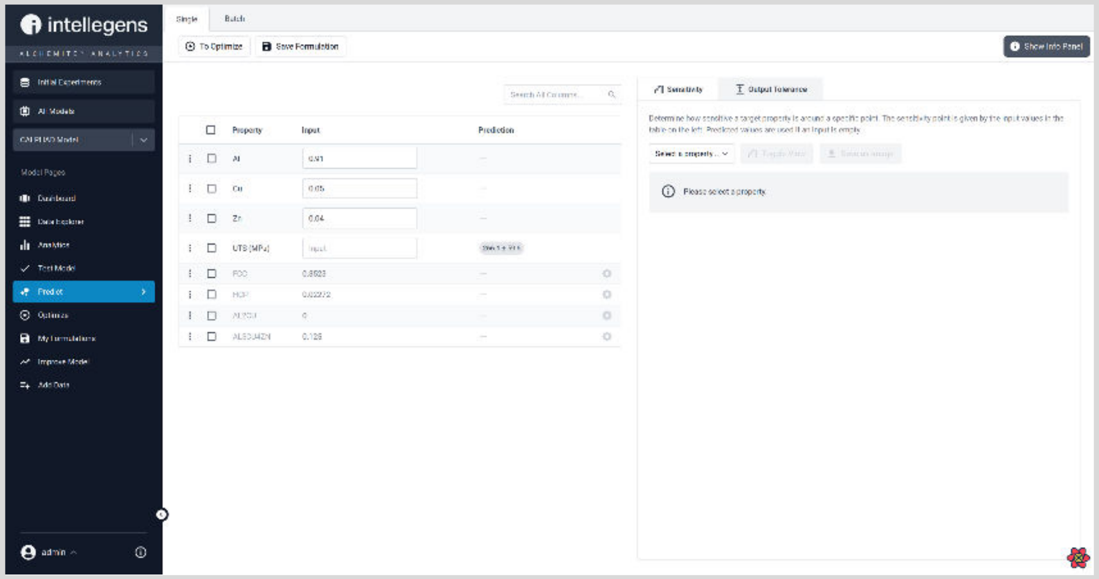
Specify which columns in your dataset Alchemite should predict values for. These are usually the output or target properties you would measure after completing an experiment, for example. All other columns will be used as inputs to the model, for instance the ingredients and parameters used when setting up an experiment. By default, Alchemite will treat all columns as outputs.

Note: In order to get the best performance, Alchemite is also able to use any of the outputs below to predict other outputs without any extra settings.

Categorical data contains values which are both unordered and discrete, and are often represented as text, such as "Process A", "Process B" or "Red", "Green", "Blue".

Columns containing text values are autodetected as categorical and you can see the unique values below. If you have any categorical columns with values

Figure 8. Selecting the physical science method in Alchemite™ Analytics.



	□	Property	Input	Prediction	
⋮	□	Al	0.91	—	
⋮	□	Cu	0.05	—	
⋮	□	Zn	0.04	—	
⋮	□	UTS (MPa)	Input	[illegible]	
⋮	□	FCC	0.8623	—	⚙
⋮	□	HCP	0.02272	—	⚙
⋮	□	ALSCU	0	—	⚙
⋮	□	ALSCU4ZN	0.128	—	⚙

Determine how sensitive a target property is around a specific point. The sensitivity point is given by the input values in the table on the left. Predicted values are used if an input is empty.

Please select a property.

Figure 9. Setting up an ML prediction using feature-engineered data as inputs.

The implementation of feature engineering for further physics and chemistry methods is an ongoing project at Intellegens. The structure of the Ichnite™ framework allows the quick development of new methods, which can then be used to generate more powerful Alchemite™ models. Future collaboration with prospective and existing customers will help integrate bespoke data processing methods into Ichnite™, removing the need to run custom scripts outside of the Alchemite™ platform. This kind of collaboration will help improve the quality of results for customers and the efficiency of the digital experiment workflow.

Conclusion

The project has created a framework for using physical science methods in combination with the Alchemite™ machine learning method, which can be made easy to use via the Alchemite™ Analytics user interface. The Alchemite™ machine learning tool delivers improved results when equipped with domain knowledge generated by these algorithms. Integration of feature engineering in this way allows users to improve models without having to set up custom scripts on their data. Collaboration with Intellegens can ensure the best scientific methods are implemented with priority, which will help extract even more value from the Alchemite™ platform.

Would you like to collaborate with Intellegens with custom or existing feature engineering methods? We can work together to implement industry-standard processing methods that are commonly used in your domain, or even implement custom data processing scripts that only you can access. Get in touch!

info@intellegens.com

References

1. Stuckner, J., Whitehead, T.M., Parini, R.C., Conduit, G.J., Benafan, O. and Arnold, S.M., 2022. *Design of Materials with Alchemite* (No. E-20054).

2. Whitehead, T.M., Chen, F., Daly, C. and Conduit, G., 2022. Accelerating the Design of Automotive Catalyst Products Using Machine Learning Leveraging Experimental Data to Guide New Formulations.

3. Ghobakhloo, M., 2020. Industry 4.0, digitization, and opportunities for sustainability. *Journal of cleaner production*, *252*, p.119869.

4. The Henry Royce Institute, 2021, Material 4.0 Roadmap: Predicting and controlling materials' microstructures and performance. https://www.royce.ac.uk/collaborate/roadmapping-landscaping/materials-4-0/

5. Barbedo, J.G.A., 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and electronics in agriculture*, *153*, pp.46-53.

6. Saunders, N. and Miodownik, A.P. eds., 1998. *CALPHAD (calculation of phase diagrams): a comprehensive guide*. Elsevier.

7. Otis, R. & Liu, Z.-K., 2017. pycalphad: CALPHAD-based Computational Thermodynamics in Python. *Journal of Open Research Software*. **5**(1), p.1.

8. Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), pp.31-36.

9. Michał Handzlik, Bas van Beek, Patrick Melix, Robert Rüger, Tomáš Trnka, Lars Ridder, Felipe Zapata, Python Library for Automating Molecular Simulations, SCM

10. R. Rüger, A. Yakovlev, P. Philipsen, S. Borini, P. Melix, A.F. Oliveira, M. Franchini, T. van Vuren, T. Soini, M. de Reus, M. Ghorbani Asl, T. Q. Teodoro, D. McCormack, S. Patchkovskii, T. Heine, AMS DFTB, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, http://www.scm.com

11. Cameron, A.C. and Windmeijer, F.A., 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, *77*(2), pp.329-342.

12. McInnes, L., Healy, J. and Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

13. Koch W, Holthausen MC. A chemist's guide to density functional theory. John Wiley & Sons; 2015 Nov 18.

14. Hollingsworth, S.A. and Dror, R.O., 2018. Molecular dynamics simulation for all. *Neuron*, *99*(6), pp.1129-1143.

About Intellegens

Intellegens provides unique deep learning software, Alchemite™. Our focus is on making it easy to apply machine learning to accelerate innovation in materials, chemicals, manufacturing, and beyond. Alchemite™ can train machine learning models from real-world, sparse, noisy data. The method originated at the University of Cambridge and development is on-going at Intellegens. Successful applications include industrial R&D and process improvements in superalloys, additive manufacturing, chemical processes, formulated products, batteries, and drug discovery.

About the Henry Royce Institute

Operating with its Hub at The University of Manchester, Royce is a Partnership of nine leading institutions – the universities of Cambridge, Imperial College London, Liverpool, Leeds, Oxford, Sheffield, the National Nuclear Laboratory, and UKAEA. Royce's associate partners are the universities of Cranfield and Strathclyde. Royce coordinates over £200 million of facilities, providing a joined-up framework that can deliver beyond the current capabilities of individual Partners or research teams. Royce is the front door to the UK materials research and innovation community open to academia, industry and the public. Their research tackles some of the most pressing challenges facing today's society, from providing energy for future cities to decarbonisation and new recyclable materials.

About SCM

SCM develops and markets the Amsterdam Modeling Suite (AMS), a powerful computational chemistry software package. The SCM staff are passionate to help chemistry and materials researchers develop new and improved molecules, materials, and processes more quickly and with less wasteful experiments. Integrating with deep-learning tools such as Alchemite™ can further improve discovery time & costs.

The comprehensive AMS package includes electronic structure methods (DFT, DFTB), force field based methods (ReaxFF, Machine Learned potentials), kinetics, fluid thermodynamics (COSMO-RS), as well as an excellent user interface (GUI), python scripting tools for workflows and parametrization, and the central AMS driver for complex energy landscape exploration. AMS is being further expanded for multiscale modelling in catalysis, organic electronics, and batteries.

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