https://doi.org/10.1595/205651322X16270488736796 *Johnson Matthey Technol. Rev.*, 2022, **66**, (2), 130–136

JOHNSON MATTHEY
TECHNOLOGY REVIEW

www.technology.matthey.com

Accelerating the Design of Automotive Catalyst Products Using Machine Learning

Leveraging experimental data to guide new formulations

Thomas M. Whitehead*
Intellegens Ltd, Eagle Labs, Chesterton Road, Cambridge, UK

Flora Chen, Christopher Daly
Johnson Matthey, Orchard Road, Royston, Hertfordshire, SG8 5HE, UK

Gareth J. Conduit
Intellegens Ltd, Eagle Labs, Chesterton Road, Cambridge, UK; and Theory of Condensed Matter, Department of Physics, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, UK

***Email:** tom@intellegens.ai

PEER REVIEWED

Received 6th May 2021; Revised 8th July 2021; Accepted 22nd July 2021; Online 23rd July 2021

The design of catalyst products to reduce harmful emissions is currently an intensive process of expert-driven discovery, taking several years to develop a product. Machine learning can accelerate this timescale, leveraging historic experimental data from related products to guide which new formulations and experiments will enable a project to most directly reach its targets. We used machine learning to accurately model 16 key performance targets for catalyst products, enabling detailed understanding of the factors governing catalyst performance and realistic suggestions of future experiments to rapidly develop more effective products. The proposed formulations are currently undergoing experimental validation.

Introduction

Domestic and commercial vehicles are leading sources of global pollution, with vehicle emissions risking the health of communities near roads (1). Fine and ultrafine particulate matter, oxides of nitrogen, hydrocarbons and carbon monoxide are key road traffic pollutants that are associated with adverse health effects (2). Catalytic converters have been used since the 1970s to reduce the emission of these pollutants by catalysing their reaction into less-toxic substances, typically carbon dioxide, nitrogen and water (3). However, current catalytic converters are not 100% efficient in their reactions of pollutants and moreover have variable efficiency at different operating temperatures.

This work uses machine learning modelling to analyse current catalytic converter performance and identify which future experimental tests would add most value to the ongoing development of improved catalytic converters. Previous work using machine learning in the catalysis domain has tended to focus on either augmenting quantum mechanical models of catalyst function (4–8), screening potential new catalysts (7–11), or predicting properties from carefully-selected chemical descriptors of catalysts (6, 8, 12–14). In contrast, in this work we focus on modelling catalyst properties from the formulation ingredients and processing variables of the catalyst. The ingredients and processing conditions of samples are easily accessible during the development process, lowering the barrier to application of machine learning in active development projects. In the following section we discuss the project objectives, detail the machine learning methodology used and the results it delivers, before looking forward to potential future applications of machine learning for materials science in the automotive field and beyond.

Objectives

We collated data on 612 catalytic converter test sets that have been manufactured and experimentally tested by Johnson Matthey as part of an ongoing catalyst development project. The data contained information on the formulation used for the catalysts, including amounts and properties of 34 ingredients; 10 test parameters describing the testing process for each catalyst; and 16 experimentally measured properties for each catalyst including target gas conversions and selectivities. These output properties consisted of eight sets of tests, with each test run at both a high (approx. 500°C) and low (approx. 225°C) temperature on different samples of the same catalyst formulation. Each experimental property was reported as a steady-state average over 50–100 s of gas stream.

Using this data, we aimed to build understanding of the performance of this class of catalyst, using a machine learning model trained on the data to extract information on which input features of the formulation and processing parameters have most impact on the performance. Using this model, we then designed catalysts that offer both high performance and also add value to the machine learning model, which once made and measured can be added to the training dataset to enable more accurate modelling of high performance catalysts.

Methods

To model the catalyst data we used the Alchemite™ multi-target machine learning platform. This method is described in detail in the literature (15–17), but in brief consists of iteratively generating predictions for all data series, both input and output, and using these predictions to impute missing data on the input side, before the final iteration of predictions are reported as the predictions for the output series. This method is designed to handle sparse input data, as was found in this work where up to 10% of the catalysts were missing information on each of the input properties. As the method is multi-target, generating predictions for all output properties simultaneously, we trained a model to predict all 16 experimentally measured properties at once. Alchemite™ also generates estimates of the uncertainty in each prediction, which is vital to prioritise suggestions for future experiments that are most likely to achieve specified objectives. To test the performance of the model, data on 61 catalysts (10% of the data) was randomly held back; the model was trained on data for the remaining 551 catalysts. Hyperparameters of the model were optimised using Bayesian Tree of Parzen Estimators *via* five-fold cross-validation within the training set only (17, 18).

To test the performance of the model we simultaneously predicted all 16 output properties for each of the 61 held-back catalysts and measured the coefficient of determination R^2, for each output property. The coefficient of determination is defined as Equation (i):

$$R^2 = 1 - \frac{\Sigma_i(y_i - f_i)^2}{\Sigma_i(y_i - \bar{y})^2} \quad \text{(i)}$$

where i indexes each catalyst in the validation set; y_i are the true experimental values, with mean $\bar{y}$; and f_i are the model predictions. A value of 1 indicates a perfect fit between model and experimental values; a value of 0 indicates a fit that is no better than random chance; and negative values indicate predictions that are worse than random. The performance of the model is shown in light blue in **Figure 1**. The median R^2 across all the output properties is 0.71, indicating highly successful predictive accuracy. In **Figure 1** we also compare to two robust standard machine learning approaches: support vector regression with radial basis function kernel and K nearest neighbours with 20 neighbours, implemented in scikit-learn (19), which were trained on a mean-imputed version of the ingredient and test parameter data and achieve baseline median R^2 values of 0.52 and 0.49 respectively.

We observed that the predictions for Property 6, at both high and low temperatures, were poor: we identified that although changes in Property 6 are observable, a key physical mechanism directly influencing the value of Property 6 is driven by a chemical species not easily measurable by any analytical method and so is not fully captured in the dataset used to train the models. This explains the poor performance of the models in this aspect. The addition of (perhaps heuristic) descriptors to capture the physical mechanism may improve the modelling performance (14), but at the cost of increasing the barrier to usage of the method compared to taking only ingredients and processes as input.

Because the experimental tests on the catalysts are each repeated, run first at high temperature and then at low temperature, these results can be correlated so there is the possibility of increasing the efficiency of the testing process by using machine learning to replace one of the rounds of testing. To validate this, we trained a machine

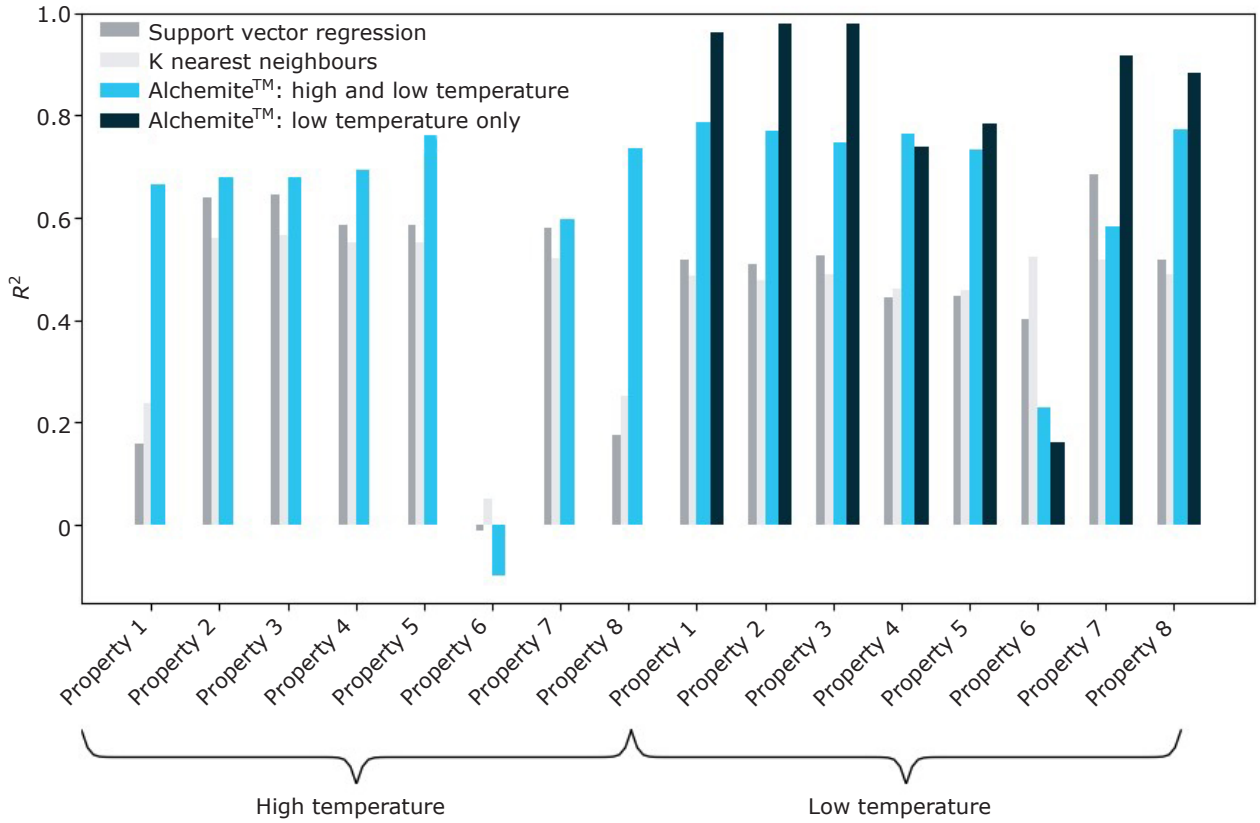

Fig. 1. The coefficient of determination in prediction of each output property against the holdout test set, showing predictions of both high and low temperature tests in light blue and predictions using the high temperature experimental results to help predict the low temperature results in dark blue. Results from support vector regression and K nearest neighbours models are shown in grey for comparison

learning model that took as inputs the formulation ingredients and test parameters as well as the experimentally measured results on all eight tests at high temperature, and predicted the results of the eight tests at low temperature. This order (using high temperature results as input to predict low temperature results) was selected to align with the current testing methodology.

The improved performance by using the high temperature measurements to help predict the low temperature performance is exemplified in dark blue in **Figure 1**. For five of the eight experimental properties the accuracy significantly increased (increase in R^2 of more than 0.1), and for Properties 1, 2 and 3 the resulting accuracy, with R^2>0.95, is effectively equivalent to the experimental uncertainty in the measurement. For these three properties in particular, machine learning predictions could reliably replace experimental measurements, offering a saving in the time and effort required to run the experimental tests on new catalysts. The three experimental properties that were not improved by using the high temperature measurements are all related to the same target gas' conversion rates, although it is not clear why these properties are not improved by access to increased experimental data. These three experimental properties are less commercially important than Property 1, which is the property with most commercial relevance.

Machine Learning Results

Now that we have confirmed the accuracy of the model we are well-positioned to extract actionable insights. Therefore, we first analyse the relationships that the model identified between inputs and outputs. To do so we examined which input features are used by the model when making predictions for each of the output properties, by evaluating the overall relative weights assigned to each input feature by the trained model, i.e. what fraction of the model prediction for each output is attributable to each input feature, on average across the whole model. This is calculated using the information gain attributable to each input feature (20). The results are summarised in **Figure 2**, separately for the model trained to predict both high and low temperature properties and the model

trained to predict low temperature properties only. Averaging across each of the output properties, we find that for the high and low temperature model the test parameters and formulation ingredients are utilised in the proportion 0.59:1, and for the low temperature only model the test parameters, formulation ingredients and experimental high temperature measurements are utilised in the proportion 0.60:1:1.19. The consistent ratio of 0.6:1 in utilisation of the test parameters and formulation ingredients between the two models indicates that the high temperature experimental measurements (especially Properties 1, 2 and 3) are adding distinct information to the model that it was not capable of identifying from either the test parameters or formulation ingredients.

The key operational insight derived from this analysis was that although the formulation ingredients provide important information for the simultaneous modelling of the high and low temperature results, the variation in the test parameters also provides a key contribution. Historically the test parameters have been controlled within specification ranges but the impact of variation within these ranges has not been considered. These results show that the test parameters have an impact on the resulting properties and that control and understanding of these parameters improves the value of the data.

Machine Learning Formulation Design

With increased understanding of the importance of the test parameters for measured catalyst performance, we used the machine learning model to design catalyst formulations. For performance targets, we focussed on the most commercially important property (Property 1), aiming to maximise its value at both high and low temperatures, and for that value to be stable with temperature. Although Property 1 is the most commercially important property, the values of the other properties are also required for product success.

As well as looking for the formulations that would be most likely to succeed against these performance targets ('exploitation' of the model) we also searched for formulations that, when measured, would increase the model's understanding of the formulation landscape and so improve future rounds of predictive modelling and formulation design ('exploration' of the model), as

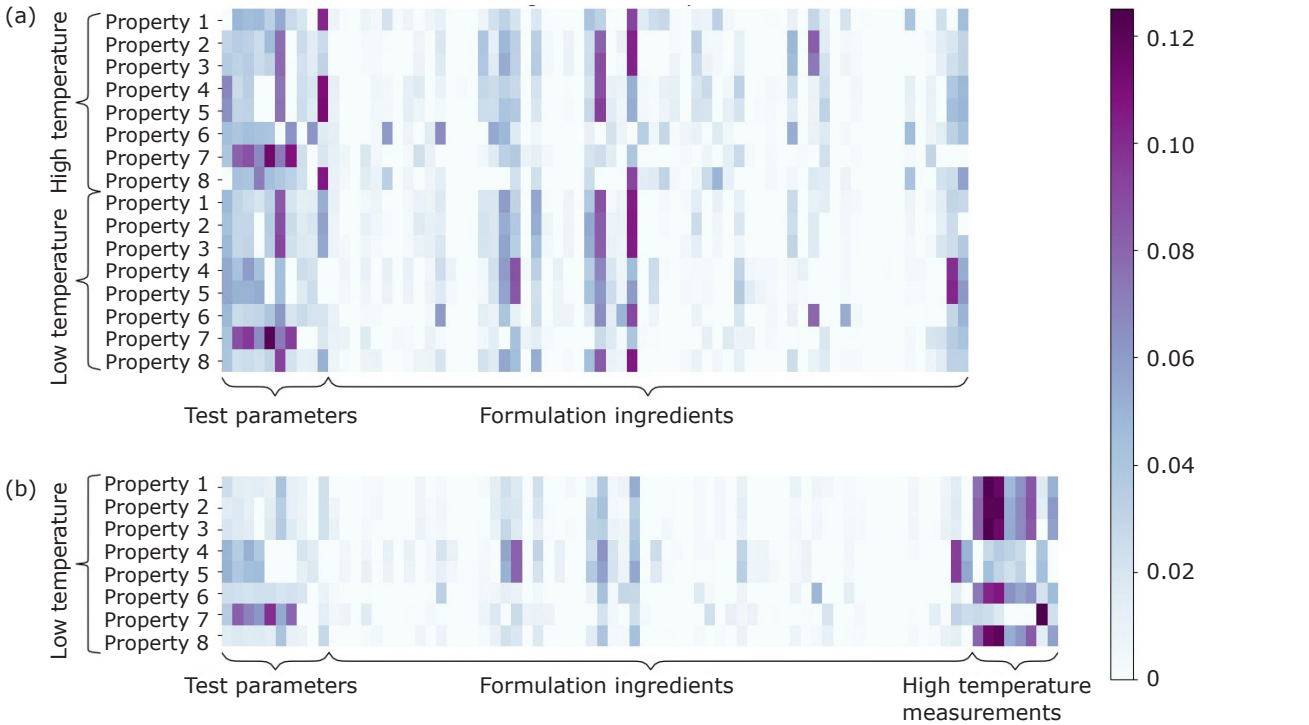

Fig. 2. Importance of each input factor (horizontal axis) for making predictions of each output property (vertical axis). The upper plot shows the model trained to predict both high and low temperature results, whilst the lower plot shows the model trained to use the high temperature results to help predict the low temperature results. Higher values (darker colours) indicate more importance given to a variable. The importance values sum to one for each output property

well as a balanced mix of the two objectives. We used a Bayesian search of the formulation space using Tree of Parzen Estimators (18) built into the Alchemite™ platform, taking as the cost function the probability of simultaneously achieving all the performance targets, including a contribution from the uncertainty in each formulation's predicted performance calculated as standard errors across the Alchemite™ platform's internal ensemble of sub-models (21). This cost function is the commercially relevant metric to aim to propose successful and useful new formulations. Exploitation-focused suggestions prioritise formulations with high probability of success, while exploration-focused suggestions prioritise formulations whose predictions are currently uncertain and will also help improve predictions over a wide range of formulation space.

A two-dimensional Uniform Manifold Approximation and Projection (UMAP) embedding (22) of the formulations is shown in **Figure 3**. The dark blue points show the historic experimental results, with more opaque points having higher performance against Property 1 and more transparent points having lower performance. We observe that there are several clusters of dissimilar formulations that had previously been measured, but that most of the formulations were relatively similar and are clustered in the centre of the plot (this clustering analysis being a key strength of the UMAP approach). **Figure 3** also shows the formulations proposed by the machine learning approach, labelled by whether they are focused on exploration, exploitation or a balanced mixture. We observe that, as expected, the exploitation-focused suggestions are clustered more tightly at the centre of the plot, demonstrating that they are attempting to exploit a class of formulations with a high probability (up to 60%) of achieving all of the design targets simultaneously. In contrast, the exploration-focused suggestions are more varied, focusing particularly on gaps in the existing coverage of the formulation space where additional information will improve the model. The balanced suggestions show aspects of both behaviours. A subset of the formulations suggested by the machine learning, including samples from the exploration, exploitation and balanced suggestions, are currently undergoing experimental validation.

Conclusions

In this work we have shown how machine learning analysis of catalyst formulations enables new insights into the factors that affect catalyst performance, including particularly that the test parameters more strongly impact the eventual performance than was initially anticipated: this will have operational significance for the future of this product development. We have also shown how the use of a machine learning platform, rather than a single predictive tool, can enable full design workflows, including prioritising exploration of the formulation space or exploitation of a model to achieve high product performance, accelerating the design process by enabling a holistic view of the formulation opportunities. Future progress in this project could focus on achieving multiple target properties simultaneously, beyond only Property 1, or utilising the accurate predictions

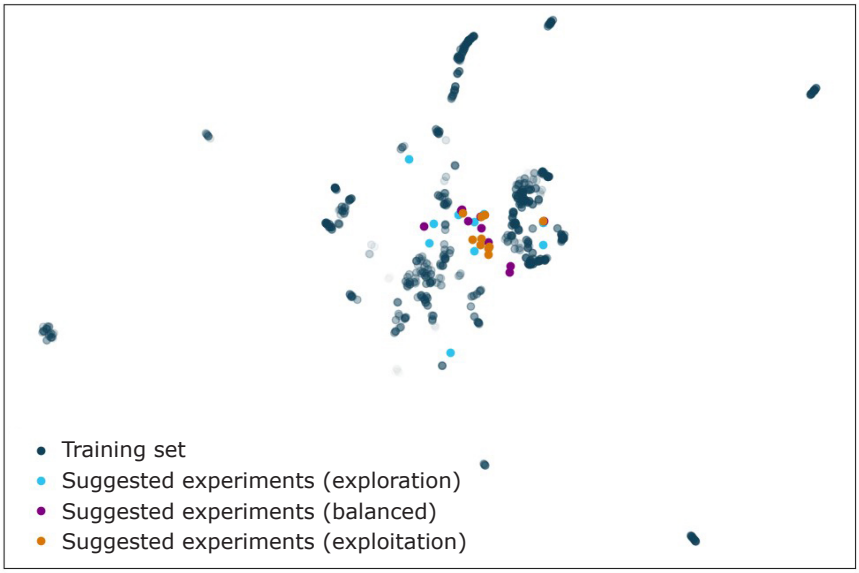

Fig. 3. Two-dimensional UMAP embedding of the training data (blue points), with darker points those with higher performance on Property 1. Also shown are the experiments suggested by the machine learning approach, in light blue (exploration focused), purple (balanced search) and orange (exploitation focused)

of low temperature measurements based on experimental high temperature measurements to halve the amount of experimental effort required when screening new formulations.

The machine learning approach here is applicable beyond catalytic converters, including the design of metal alloys (15, 23), batteries (24), and pharmaceutical drugs (21). A machine learning platform that can carry out the full cycle of formulation development, handling sparse real-world experimental data, building predictive models and proposing and interpreting new formulation designs adds value in each of these areas, with reduced barrier to entry by working directly on the composition and processing variables immediately accessible to project scientists.

Acknowledgements

Gareth Conduit acknowledges financial support from the Royal Society. There is Open Access to this paper online.

References

1. K. Zhang and S. Batterman, *Sci. Total Environ.*, 2013, **450–451**, 307
2. D. Brugge, J. L. Durant and C. Rioux, *Environ. Health*, 2007, **6**, 23
3. C. Morgan, *Johnson Matthey Technol. Rev.*, 2014, **58**, (4), 217
4. K. Shakouri, J. Behler, J. Meyer and G.-J. Kroes, *J. Phys. Chem. Lett.*, 2017, **8**, (10), 2131
5. Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621
6. J. R. Kitchin, *Nat. Catal.*, 2018, **1**, (4), 230
7. W. Yang, T. T. Fidelis and W.-H. Sun, *ACS Omega*, 2019, **5**, (1), 83
8. B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel and C. Sutton, *AIChE J.*, 2018, **64**, (7), 2311
9. Z. Li, S. Wang, W. S. Chin, L. E. Achenie and H. Xin, *J. Mater. Chem. A*, 2017, **5**, (46), 24131
10. Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan and J. K. Nørskov, *ACS Catal.*, 2017, **7**, (10), 6600
11. T. Williams, K. McCullough and J. A. Lauterbach, *Chem. Mater.*, 2020, **32**, (1), 157
12. Z. Li, X. Ma and H. Xin, *Catal. Today*, 2017, **280**, (2), 232
13. I. Takigawa, K.-i. Shimizu, K. Tsuda and S. Takakusagi, *RSC Adv.*, 2016, **6**, (58), 52587
14. K. Suzuki, T. Toyao, Z. Maeno, S. Takakusagi, K.-i. Shimizu and I. Takigawa, *ChemCatChem*, 2019, **11**, (18), 4537
15. B. D. Conduit, N. G. Jones, H. J. Stone and G. J. Conduit, *Scr. Mater.*, 2018, **146**, 82
16. P. Santak and G. Conduit, *Fluid Phase Equilib.*, 2019, **501**, 112259
17. T. M. Whitehead, B. W. J. Irwin, P. Hunt, M. D. Segall and G. J. Conduit, *J. Chem. Inf. Model.*, 2019, **59**, (3), 1197
18. J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, 'Algorithms for Hyper-Parameter Optimization', NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems, 12th–15th December, 2011, Granada, Spain, Curran Associates Inc, New York, USA, 2011, 9 pp
19. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825
20. B. Frénay, G. Doquire and M. Verleysen, *Neural Networks*, 2013, **48**, 1
21. B. W. J. Irwin, J. R. Levell, T. M. Whitehead, M. D. Segall and G. J. Conduit, *J. Chem. Inf. Model.*, 2020, **60**, (6), 2848
22. L. McInnes, J. Healy and J. Melville, 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction', arXiv:1802.03426v3 [stat.ML], 18th September, 2020, *preprint*
23. B. D. Conduit, N. G. Jones, H. J. Stone and G. J. Conduit, *Mater. Des.*, 2017, **131**, 358
24. M.-F. Ng, J. Zhao, Q. Yan, G. J. Conduit and Z. W. Seh, *Nat. Mach. Intell.*, 2020, **2**, (3), 161

The Authors

Thomas Whitehead holds a PhD in theoretical physics from the University of Cambridge, UK, and is now leading the application of Intellegens' novel deep learning approaches to a wide variety of industrial applications. His work focuses on developing a series of application-specific machine learning modules to address high-value data analysis bottlenecks.

Flora Chen is the Data Science Lead at Johnson Matthey. She has 15 years' experience in global high-tech companies and has held technical and management roles spanning engineering, operations, R&D and quality. Since Flora joined Johnson Matthey in 2018, she has led several digital analytics projects, discovering and delivering the business value of data. Flora holds a PhD in Mechanical Engineering from Bristol University, UK, and is a chartered engineer.

Christopher Daly received an MChem (2008) and PhD (2012) in Chemistry from the University of Leicester, UK, where his research focused on the synthesis of organometallic compounds of the late transition metals and their applications in bifunctional catalysis. Since 2013 he has worked on automotive catalyst development at Johnson Matthey across several technologies, where he is currently a Senior Chemist.

Gareth Conduit has a track record of developing and applying machine learning methods to solve real-world problems. The approach, originally developed for materials design, is now being commercialised by startup Intellegens in materials design, healthcare and drug discovery. Gareth also has research interests in strongly correlated phenomena, in particular proposing spin spiral state in the itinerant ferromagnet that was later observed in CeFePO. Gareth's group is based at the University of Cambridge.