ROYAL SOCIETY
OF CHEMISTRY

PAPER

Check for updates

Composition-property extrapolation for compositionally complex solid solutions based on word embeddings†

Lei Zhang, [ID] *[a] Lars Banko,[b] Wolfgang Schuhmann, [ID] [c] Alfred Ludwig [ID] [b] and Markus Stricker [ID] [a]

Mastering the challenge of predicting properties of unknown materials with multiple principal elements (high entropy alloys/compositionally complex solid solutions) is crucial for the speedup in materials discovery. We show and discuss three models, using experimentally measured electrocatalytic performance data from two ternary systems (Ag–Pd–Ru; Ag–Pd–Pt), to predict electrocatalytic performance in the shared quaternary system (Ag–Pd–Pt–Ru). As a starting point, we apply Gaussian Process Regression (GPR) based on composition as the feature, which includes both Ag and Pd, achieving an initial correlation coefficient for the prediction (r) of 0.63 and a determination coefficient (r^2) of 0.08. Second, we present a version of the GPR model using word embedding-derived materials vectors as features. Using materials-specific embedding vectors significantly improves the predictions, evident from an improved r^2 of 0.65. The third model is based on a 'standard vector method' which synthesizes weighted vector representations of material properties as features, then creating a reference vector that results in a very good correlation with the quaternary system's material performance (resulting r of 0.94). Our approach demonstrates that existing experimental data combined with the latent knowledge of word embedding-derived representations of materials can be used effectively for materials discovery where data is typically scarce.

1 Introduction

Materials science is a driver of technological progress by development of innovative materials that enable advancements across industries from electronics to aerospace.[1,2] Novel materials are the driver because of new properties or property combinations or by replacing existing critical or expensive materials with less critical ones while at the same time not sacrificing performance. Discovering new materials (fast) requires accurate prediction of material properties, particularly in compositionally complex materials with four or more primary elements. Such systems show promise as Discovery Platforms, *e.g.* for electrocatalysis.[3] However, they pose significant challenges for discovery since the possible combinations of

elements and their compositional ratios render brute-force screening approaches practically impossible. Additionally, predicting their properties is difficult due to their complex compositional interactions and the intricate ways in which these interactions affect material behavior.[4] As such, the acceleration of the discovery process for new materials necessitates the development of new methods to navigate complex composition-structure–property relationships of promising material systems.

The integration of computational power and data analysis is necessary in overcoming the challenges presented by these material systems.[5] Machine learning has emerged as a useful tool, providing a path for materials scientists to predict and understand the properties of materials systems.[6,7] This transition from traditional, heuristic approaches to data-driven, computational strategies signifies a transformation of the field,[8,9] aligning with the complexity of the possible materials of interest.

Among data-centered approaches, Gaussian Process Regression (GPR) has demonstrated exceptional versatility and efficacy across multiple domains, illustrating its capacity to model complex relationships.[10] The adaptability of GPR stems from its non-parametric approach which allows to adjust its complexity based on the dataset, a feature that sets it apart from models like neural networks.[11] This flexibility renders it

[a] Interdisciplinary Centre for Advanced Materials Simulation, Ruhr University Bochum, Universitätsstraße 150, 44780 Bochum, Germany. E-mail: lei.zhang-w2i@rub.de; markus.stricker@rub.de

[b] Chair for Materials Discovery and Interfaces, Institute for Materials, Ruhr University Bochum, Universitätsstraße 150, 44780 Bochum, Germany. E-mail: lars.banko@rub.de; alfred.ludwig@rub.de

[c] Analytical Chemistry – Center for Electrochemical Sciences (CES), Faculty of Chemistry and Biochemistry, Ruhr University Bochum, Universitätsstraße 150, 44780 Bochum, Germany. E-mail: wolfgang.schuhmann@rub.de

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5dd00169b

particularly valuable in applications for complex non-linear relationships in high-dimensional data spaces.

However, the usefulness, *i.e.* the predictive power, hinges on available data and meaningful representations of materials. Often, sophisticated adjustments to such models are necessary to effectively capture complex correlations.[12] Possible modifications include appropriate accounting for noise in data and customization of kernel functions.

A critical part of any data-based approach is the representation[13–15] of the input. In particular, the challenge of how to represent a material. A simple approach is to just use the composition.[16] This is often sufficient for interpolation. However, if the goals is to predict into unknown spaces, any existing knowledge about a material or similar materials and their properties is desirable.

In this, the vast expanse of scientific literature represents a rich, yet not fully exploited, resource.[17] Through literature mining[18,19] and vector analysis,[20] we can convert the latent knowledge contained in scientific texts into formats amenable to machine learning in form of representations.[21] The integration of word embedding-based vector analysis, derived from literature mining, with machine learning models like GPs, represents a new path for improving predictive capabilities in for materials discovery, particularly for complex systems such as ternary and quaternary materials.

In our example, we present the problem of predicting the performance of a quaternary materials system for electrochemical applications, specifically the oxygen reduction reaction (ORR). Here, "performance" is defined as the current density of electrocatalysis of the ORR at an overpotential of 850 mV. We use existing measurements of ternary systems in conjunction with representations of materials and properties based on word embeddings. We examine three distinct approaches: standard GP modeling based on composition, GP augmented with material vectors based on word embeddings, and our 'standard vector method'.

Our approach improves the prediction capabilities for compositionally more complex materials by combining measured data from compositionally less complex materials, combined with advanced representations of materials through word embeddings. We illustrate its predictive power and compare it with the reference approach that solely relies on materials representations based on composition.

2 Methods

2.1 Dataset description

For our demonstration we use two datasets from two different overlapping ternary systems (Ag–Pd–Ru and Ag–Pd–Pt) to train models for property prediction of a shared quaternary system (Ag–Pd–Pt–Ru). The basic idea is to use compositionally less complex systems (ternary materials systems) to predict the behavior of more complex ones (quaternary) in the context of electrocatalysis, specifically the ORR.[22–24]

Two ternary datasets are used to fit models that capture their correlation with electrocatalytic properties, specifically a current at a fixed applied overpotential. These models are then used to predict the electrocatalytic properties of the shared quaternary system, which includes all the elements present in the ternary systems.

The experimental data is sourced from composition-spread materials libraries (CSML) and described in detail elsewhere.[25] Nevertheless, we provide a brief description here for completeness. The materials libraries were fabricated by co-sputtering thin films on 100 mm diameter sapphire wafers (*c*-plane) from 4 elemental targets. The targets were confocally aligned to a 100 mm substrate (target–substrate distance approx. 12 cm). Target materials had a purity of 99.99,%. Ar (99.9999%) was used as a sputter gas. The deposition pressure was 0.667 Pa. The film thickness was 100–150 nm. The chemical composition of the materials libraries was measured by energy dispersive X-ray spectroscopy (EDX) with an acceleration voltage of 20 kV. 81 measurements were done on a regular grid of 9×9 (8.5 mm spacing) on each library. Linear regression was used to interpolate the composition over the 342 measurement areas of a 4.5 mm grid that were electrochemically characterized using scanning droplet cell (SDC) experiments.

Electrochemical measurements were conducted with the use of a high-throughput SDC. The SDC head incorporates counter (Pt wire) and reference (Ag| AgCl| 3 M KCl) electrodes and a Teflon tip with 1 mm diameter. The materials library is connected as working electrode, *e.g.* the surface of the investigated sample in every spot where the tip touches the sample. The electrolyte was replaced for every measurement area. Linear sweep voltammograms were measured in 0.05 M KOH, pH 12.5, with a scan rate of 10 mV s^{-1}. All potentials are reported *versus* the RHE according to the following equation: U_{RHE} (V) $= U$(Ag| AgCl| 3 M KCl) $+ 0.210 + (0.059$ pH), where U(Ag| AgCl| 3 M KCl) is the potential measured *versus* Ag| AgCl| 3 M KCl reference electrode, 0.210 V is the standard potential of the Ag| AgCl| 3 M KCl reference electrode at 25 °C. Note that 0.059 is the result of $(RT) \times (nF)^{-1}$, where R is the gas constant, T is the temperature (298 K), F is the Faraday constant, and n is the number of electrons transferred during the reaction.

2.2 Modeling approaches

2.2.1 Method 1: Gaussian Process (GP) model with elemental composition. A Gaussian Process (GP) model based of elemental composition derived from the ternary datasets is fit to predict the electrochemical current at a potential of 850 mV for the quaternary system. This sort-of traditional approach provides a reference for predictions about electrocatalytic performance. In materials science, GP models have been effectively applied to predict various properties, including thermal conductivity[26] and electronic structure.[27] This model serves as our baseline, allowing us to evaluate the models with more nuanced representations against a reference standard.

2.2.2 Method 2: enhanced GP model with material vectors. The second model is different to the standard GP model by employing 'material vectors' instead of the elemental composition as a representation for materials. Material vectors are obtained from a Word2Vec model based on a comprehensive literature review.[28] We retrieve a 200-dimensional vector

representation of each pure element. Within this 200-dimensional space, we create representations for materials by a weighted linear combination of the elemental representations, in line with vector operations in word embedding space.[29] By employing material vectors, we use the latent knowledge from scientific literature and transform it into an explicit, quantitative form, to improve our model's predictive power. Like our baseline GP model, we predict each material's electrocatalytic performance, enabling a direct comparison between these two approaches.

2.2.3 Method 3: standard vector method. The third method is different from the GP-based models in two aspects. For one, we introduce a novel approach based on the concept of a 'standard vector'. Fig. 1 shows the process how we construct this 'standard vector'. The idea is to substitute representations of compositions based on word embeddings of elements and their linear combinations with a similarity vector obtained by comparison with known terms related to electrocatalysis, thereby encoding explicit domain knowledge in the representation of a material. The similarity of each word embedding

representation of the composition with the term constitute one dimension of the standard vector. The process begins with the assembly of a list of material properties relevant to electrocatalysis, from which vector representations are generated. Our property list include "electrocatalyst", "overpotential", "Tafel slope", "exchange current density", "stability", "durability", "surface area", "active site", "turnover frequency", "electrocatalytic activity", "faradaic efficiency", "charge transfer", "adsorption energy", "electronic structure", "electronegativity", "crystal structure", and "surface morphology" – a 17-dimensional space. Fig. 2 shows a dimensionality-reduced map of the vector representations of the listed terms using *t*-SNE.[30]

Each property in the list is chosen based on its known relationship to electrocatalytic performance and its role in determining the efficiency of the ORR. For instance, properties such as overpotential, Tafel slope, and exchange current density are critical for assessing the electrocatalytic performance of materials. Stability, durability, and surface area affect the longevity and effectiveness of catalysts in practical applications. Other properties like adsorption energy, electronic structure, and
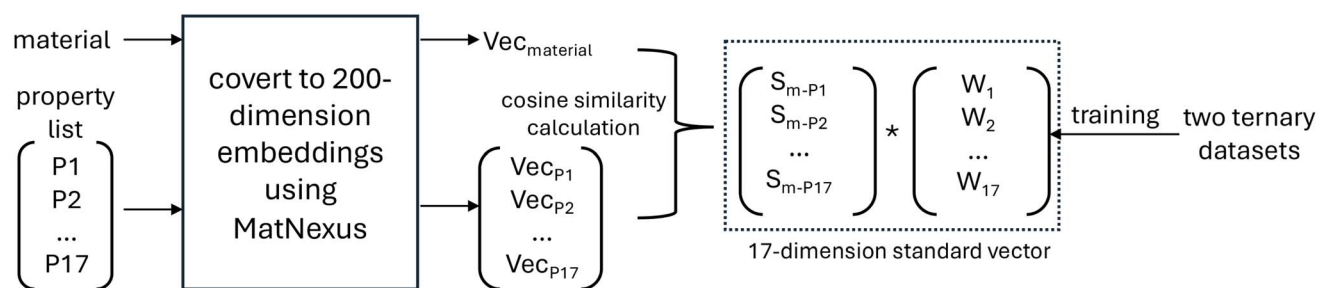

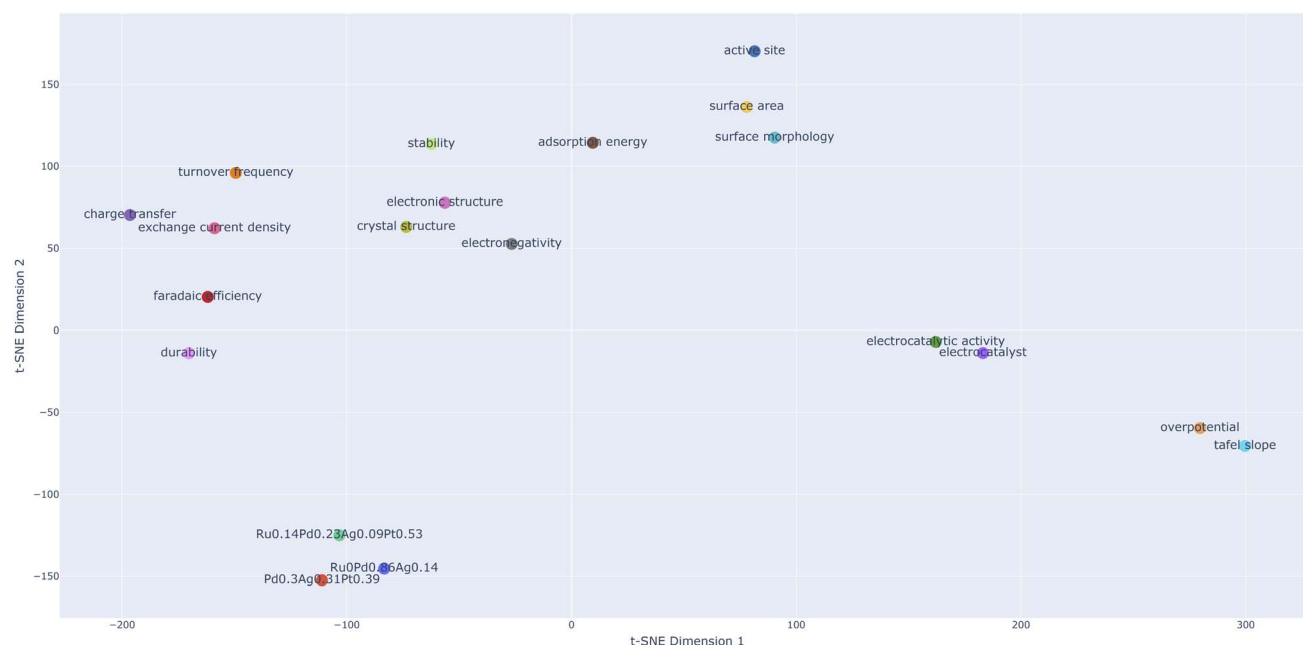Fig. 1 Illustration plot of standard vector method.

Fig. 2 Dimensionality reduced (*t*-SNE) map of vector representations for the chosen electrocatalytic properties and materials.

crystal structure offer deeper insights into the interaction mechanisms at the molecular level which might influence catalytic behavior and performance. The relative distance of word embeddings of materials to these properties capture the co-occurrence, and therefore proximity in embedding space. Our hypothesis is that proximity of properties and materials representations in embedding space captures correlations and thereby provides an improved representation of materials, not based on their composition, but based on their latent properties and their relationships.

However, the novelty of our approach is in how these property vectors are combined. Instead of simply merging the individual 17 similarity values, we calculate a 'standard vector' that represents an ideal electrocatalyst by weighting each property vector based on the experimental data for the two ternary systems to reflect its importance w.r.t. known catalytic activity in this material system.

In essence, we create a reference vector based on measured data which represents optimal characteristics for ORR performance for the given materials system. The weighting step, a fitting procedure, is a minimization with constraints. The weights are adjusted to minimize the squared difference between 'experimental indicators' (current at potential) and similarity dimensions. In our case, we use measured activity as experimental indicator, but any reliable known data for materials correlating with the predicted property could be used in general.

We then assume that materials which are 'closer' in vector space to this standard vector – measured by similarity metrics such as cosine similarity – are more likely to exhibit good electrocatalytic performance. By evaluating materials based on their proximity to this 'ideal' vector, we predict and identify promising electrocatalysts without relying solely on compositional or structural data features.

Once defined, the standard vector based on the two ternary systems is a benchmark representation for evaluating materials in the shared quaternary system. Rather than predicting performance by predicting the (measured) current directly, we apply similarity measures to pinpoint materials that align closely with the ideal standard vector, thereby identifying candidates with potentially high electrocatalytic performance.

2.2.4   Mathematical details of the standard vector method. To further clarify the Standard Vector Method, we formulated the process as follows.

2.2.4.1   Representation of compositions via word embeddings. Let $\mathbf{v}_i$ be the word embedding representation of element i, and let a material composition M consisting of elements $\{E_1, E_2, ..., E_n\}$ be represented as a linear combination:

$$\mathbf{v}_M = \sum_{i=1}^{n} c_i \mathbf{v}_i, \tag{1}$$

where c_i represents the fractional contribution of element i in the material.

2.2.4.2   Property-based similarity encoding. A set of domain-specific properties $\{P_1, P_2, ..., P_d\}$ with corresponding embedding representations $\mathbf{p}_j$ (where $j = 1, ..., d$) forms a basis for

similarity comparisons. The similarity score between a material M and a property P_j is computed using cosine similarity:

$$S(M, P_j) = \frac{\mathbf{v}_M \times \mathbf{p}_j}{\|\mathbf{v}_M\|\|\mathbf{p}_j\|}. \tag{2}$$

The vector $\mathbf{s}_M$ containing these similarity values forms a standard vector representation:

$$\mathbf{s}_M = [S(M,P_1), S(M,P_2), ..., S(M,P_d)] \in \mathbb{R}^d. \tag{3}$$

2.2.4.3   Construction of the standard vector. Instead of treating the similarity values independently, we define an optimal standard vector $\mathbf{s}^*$, which represents an ideal electrocatalyst. This vector is obtained through a weighted fitting procedure using experimental data. Given a set of materials $\{M_1, M_2, ..., M_k\}$ with experimentally measured catalytic activities y_i, the weight optimization problem is formulated as:

$$\min_{\mathbf{w}} \sum_{i=1}^{k} \left(y_i - \sum_{j=1}^{d} w_j S(M_i, P_j) \right)^2, \quad \text{subject to} \quad \sum_{j=1}^{d} w_j = 1. \tag{4}$$

Solving this constrained optimization problem provides an optimal weight vector $\mathbf{w}^* = [w_1^*, w_2^*, ..., w_d^*]$, which defines the standard vector:

$$\mathbf{s}^* = \sum_{j=1}^{d} w_j^* \mathbf{p}_j. \tag{5}$$

2.2.4.4   Evaluation of new materials. For a new material M', its proximity to the standard vector $\mathbf{s}^*$ is evaluated using cosine similarity:

$$\text{Similarity}(M', \mathbf{s}^*) = \frac{\mathbf{s}_{M'} \cdot \mathbf{s}^*}{\|\mathbf{s}_{M'}\|\|\mathbf{s}^*\|}. \tag{6}$$

Materials closer to $\mathbf{s}^*$ are predicted to exhibit superior electrocatalytic performance.

2.3   Model evaluation

The performance of the first and second GP model is quantitatively assessed using the Pearson's correlation coefficient (r) between the actual and predicted current densities, alongside the coefficient of determination (r^2), to gauge the models' ability to capture variance in the actual measurements.

The third model, employing the standard vector method, is assessed differently. Given the different nature of its output, we adapt our evaluation strategy using the correlation coefficient between the actual current densities and our predictions, the similarity scores. This metric reflects the model's performance in identifying materials with high electrocatalytic performance based on their conceptual proximity to the 'ideal' electrocatalyst as defined by the standard vector.

To further underscore the models' applicability to high-performance electrocatalysts, we introduce a filtering

criterion, focusing on data points where the current at 850 mV (current_at_850 mV) is below -0.2 mA cm^{-2}. This is designed to improve the models' ability to identify materials with significant electrocatalytic activity. By focusing on data points where the current at 850 mV indicates notable activity, we tailor our analysis to emphasize materials that, based on our dataset, stand out for their electrocatalytic performance. This method allows us to direct our model's focus and analytical efforts towards those candidates most likely to impact future electrocatalysts. In other words, for materials displaying low activity, we are not interested in 'how low'.

2.4 Model reproducibility

MatNexus[28] underpins our data processing, analysis, and visualization workflows. MatNexus supports the standardized handling of materials science data, ensuring the reproducibility of our findings through a workflow. We use it for all parts of the analysis: from initial data preprocessing to feature extraction, structuring for word embedding model training, and the visualization of datasets and analysis results.

We also use it to create a word embedding model to generate material vectors, which are then used in conjunction with the GP model as well as in the standard vector method for predictive analysis.

MatNexus is used to conduct targeted literature queries, focusing on articles indexed in Scopus with keywords 'electrocatalyst' and 'high entropy alloy' published before the year 2024. We restrict our search to Open Access (OA) articles. This approach not only aligns with our commitment to open science but also ensures compliance with copyright laws. Furthermore, in building our word embedding model, we limit our analysis to the abstracts of these papers, not the full texts, balancing depth of analysis with the accessibility of data (See the ESI† Bibliography document).

For details of the implementation of MatNexus and its functionality, refer to our MatNexus repository on PyPI (**https://www.pypi.org/project/matnexus/**).[28]

All relevant codes, experimental datasets, and model predictions are publicly accessible *via* GitHub (**https://www.github.com/lab-mids/ccss_word_embedding_prediction**), ensuring that our research can be validated, replicated, or expanded upon by others.

3 Results

3.1 Dataset overview

This section provides an overview of the datasets used for model training and prediction (Tables 1, 2), (Fig. 3, 4 and 5). The training datasets comprise two ternary systems (Ag–Pd–Ru; Ag–Pd–Pt), the prediction target data set is their shared quaternary system (Ag–Pd–Pt–Ru).

3.1.1 Ag–Pd–Ru system. The Ag–Pd–Ru system contains a range of element composition, with Pd showing the highest compositional range from 23% to 87%, followed by Ru ranging from 0% to 45% and Ag from 10% to 40%. In terms of electrochemical performance, this system shows a mean current in

Table 1 Comparative elemental composition across systems

System	Element	Minimum content (%)	Maximum content (%)
Ag–Pd–Ru	Ag	10	40
	Pd	23	87
	Ru	0	45
Ag–Pd–Pt	Ag	1	70
	Pd	0	47
	Pt	17	69
Ag–Pd–Pt–Ru	Ag	3	39
	Pd	0	28
	Pt	0	56
	Ru	7	67

Table 2 Comparative metrics of current at 850 mV across systems and their correlations with elements

Metric	Ag–Pd–Ru	Ag–Pd–Pt	Ag–Pd–Pt–Ru
Mean current (mA)	-0.278	-0.342	-0.159
Standard deviation (mA)	0.114	0.098	0.074
Minimum current (mA)	-0.673	-0.583	-0.366
25% Quantile (mA)	-0.348	-0.423	-0.195
Median (mA)	-0.248	-0.372	-0.131
75% Quantile (mA)	-0.189	-0.271	-0.110
Maximum current (mA)	-0.065	-0.063	-0.060
Correlation with Ag	$+0.766$	$+0.587$	$+0.440$
Correlation with Pd	-0.905	-0.771	-0.502
Correlation with Pt	N/A	-0.017	-0.771
Correlation with Ru	$+0.719$	N/A	$+0.719$

ORR of -0.278 mA at 850 mV. A correlation analysis reveals a significant negative correlation of Pd with electrochemical performance (-0.905), suggesting that higher contents of Pd lead to improved performance (lower current indicated better performance). Conversely, Ru and Ag show positive correlations, $+0.719$ and $+0.766$ respectively, indicating that increases in their contents may not favor performance. This suggests that optimizing Pd content while minimizing Ru and Ag could enhance the system's efficiency (Fig. 6 and 7(a)), in line with chemical intuition.[31]

3.1.2 Ag–Pd–Pt system. The Ag–Pd–Pt system exhibits a compositional range with Pd between 0% and 47%, Ag between 1% and 70%, and Pt between 17% and 69%.

The mean current at 850 mV for the Ag–Pd–Pt system is -0.342 mA, displaying a slightly better performance compared to the Ag–Pd–Ru system. The correlation analysis shows a strong negative correlation with Pd (-0.771) and a very weak negative correlation with Pt (-0.017), suggesting that Pt's influence on performance is minimal. Ag's positive correlation ($+0.587$) further implies that, similar to the Ag–Pd–Ru system, increasing Ag content does not benefit the system's performance (Fig. 8 and 7(b)).

3.1.3 Ag–Pd–Pt–Ru system. The quaternary system displays a spread of elemental composition with Ru vary from 7% to 67%, Pd from 0% to 28%, Ag from 3% to 39%, and Pt from 0% to 56% (Table 1). The performance metrics show a mean current of -0.159 mA at 850 mV, which is less negative than the other
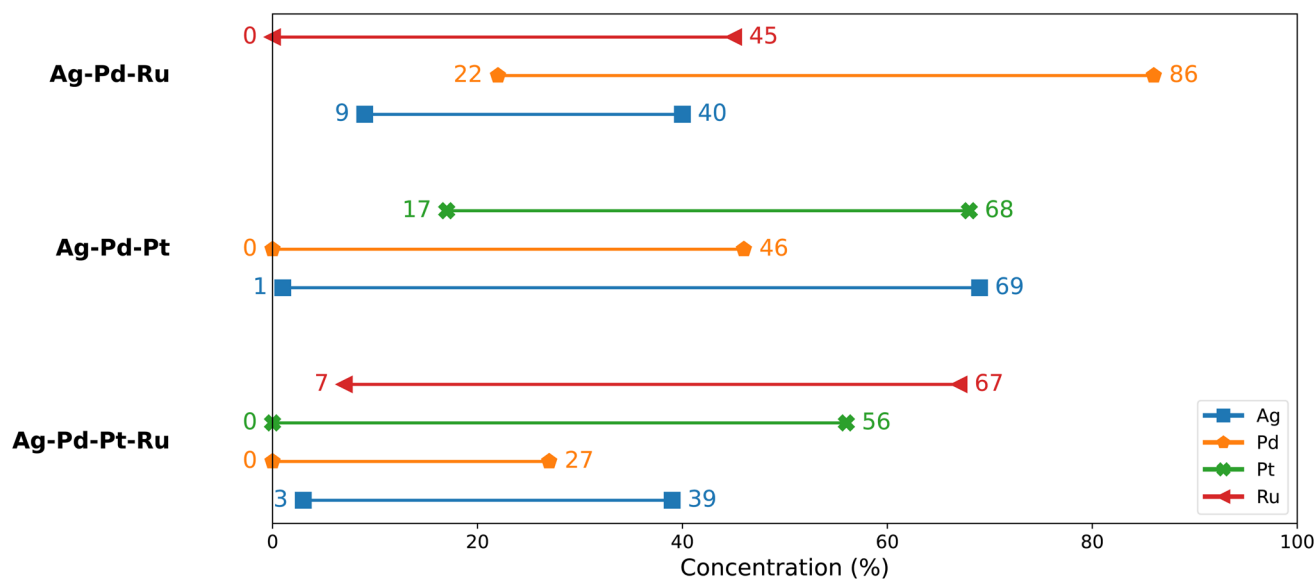
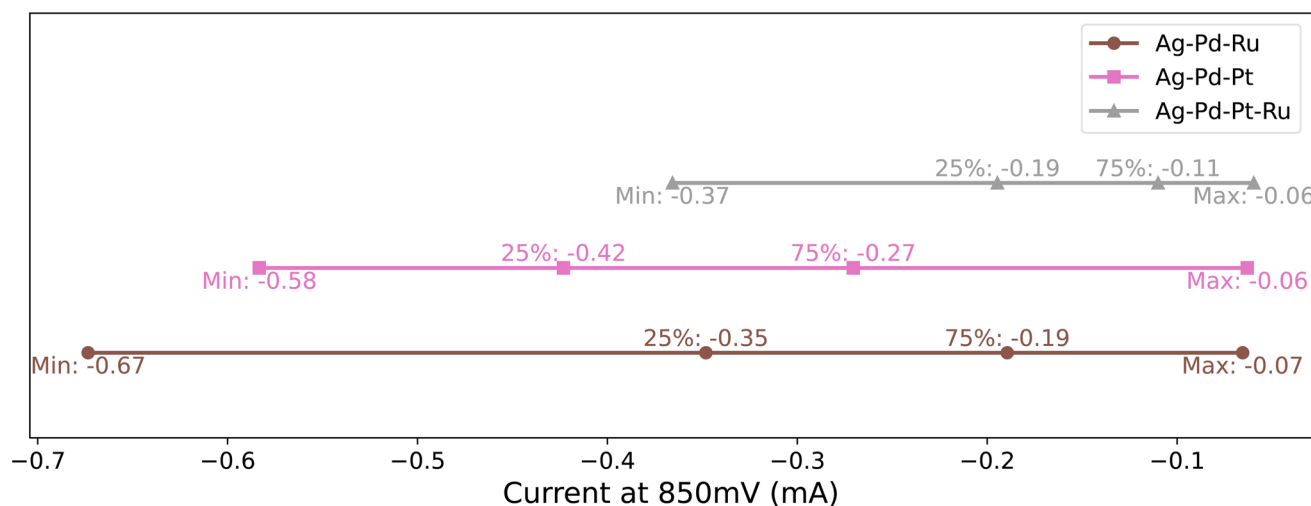Fig. 3 Compositional ranges of synthesised materials.

Fig. 4 Current density ranges of synthesised materials.

two systems, suggesting a comparative decrease in performance (Table 2).

The correlation coefficients present a complex picture. Pd's negative correlation (−0.502) is less pronounced than in the other systems, indicating its diminished influence in the presence of Pt, which shows a strong negative correlation (−0.771) with the current. This suggests that in this system, Pt plays a more critical role in enhancing performance than Pd. Ru and Ag show positive correlations, similar to the Ag–Pd–Ru system, suggesting their less favorable impact on performance (Fig. 9).

3.2 Results of method 1: GP model with elemental composition

Table 3 and Fig. 10(a and b) present the results of the application of Gaussian Process (GP) based solely on elemental

compositions. This approach demonstrates a baseline predictive capability with an overall correlation coefficient (r) of 0.85 and a coefficient of determination (R^2) of 0.08. The mean electrochemical current was measured at −0.16 mA cm^{-2} with a standard deviation of 0.07 mA cm^{-2}. The model's predictions deviate slightly, with a mean predicted current of −0.22 mA cm^{-2} and a comparable standard deviation of 0.07 mA cm^{-2}. This method demonstrates a Mean Absolute Error (MAE) of 0.06 mA cm^{-2} and a Root Mean Square Error (RMSE) of 0.07 mA cm^{-2}, indicating a moderate level of accuracy in the predictions.

3.3 Results of method 2: enhanced GP model with material vectors

The GP model's performance significantly improved using a word embedding-derived representation of materials as input
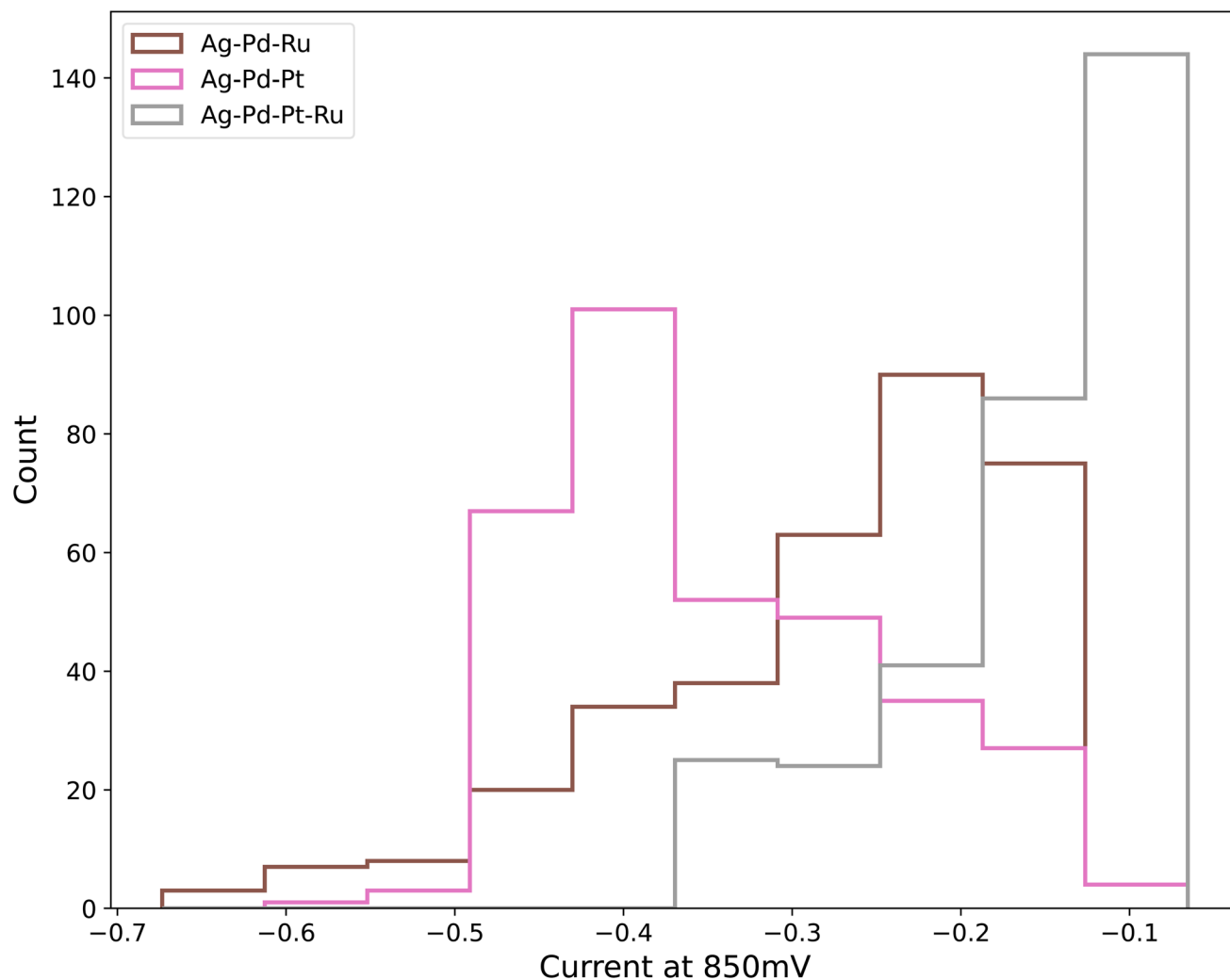
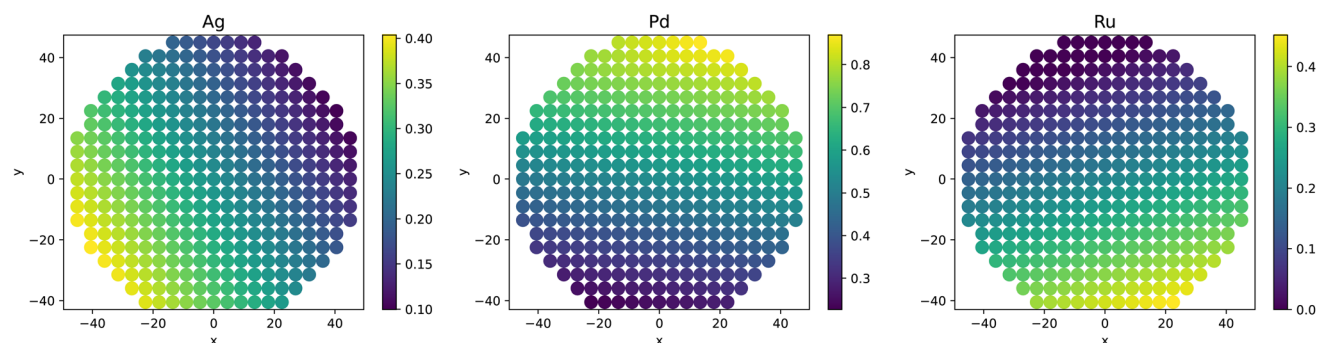Fig. 5 Stacked step histogram of current density across the samples.

Fig. 6 Color-coded plot of compositional gradients in Ag−Pd−Ru system.

(Table 3 and Fig. 10(a and c)). Most notably, the overall R^2 increases to 0.65, indicating that the model accounts for a much larger proportion of the variance in the data. This suggests a significantly stronger relationship between the predictions and actual measurements when using material vectors. While the correlation coefficient (r) slightly decreases to 0.83, the

model's ability to capture the general trend of the dataset is markedly improved. This is evidenced by the mean predicted current of -0.15 mA cm^{-2}, which closely matches the actual mean current. Additionally, with a standard deviation of 0.05 mA cm^{-2}, the predictions are more precise compared to the composition-based representation. Finally, the MAE and RMSE
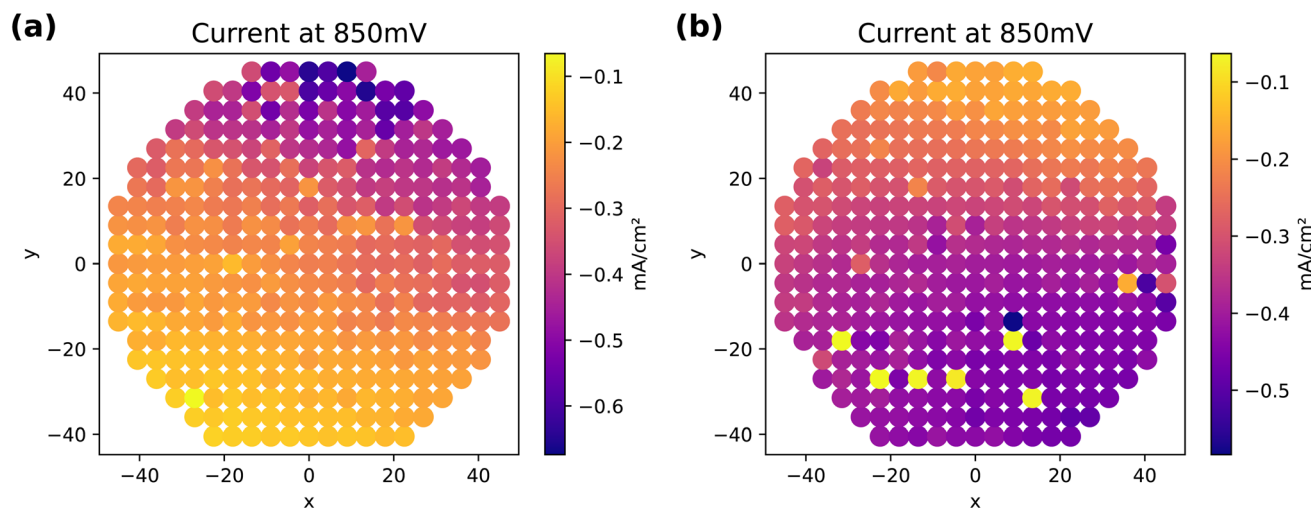
(a) Current at 850mV

(b) Current at 850mV

Fig. 7 Color-coded plot of current density gradients in: (a) Ag−Pd−Ru system and (b) Ag−Pd−Pt system.
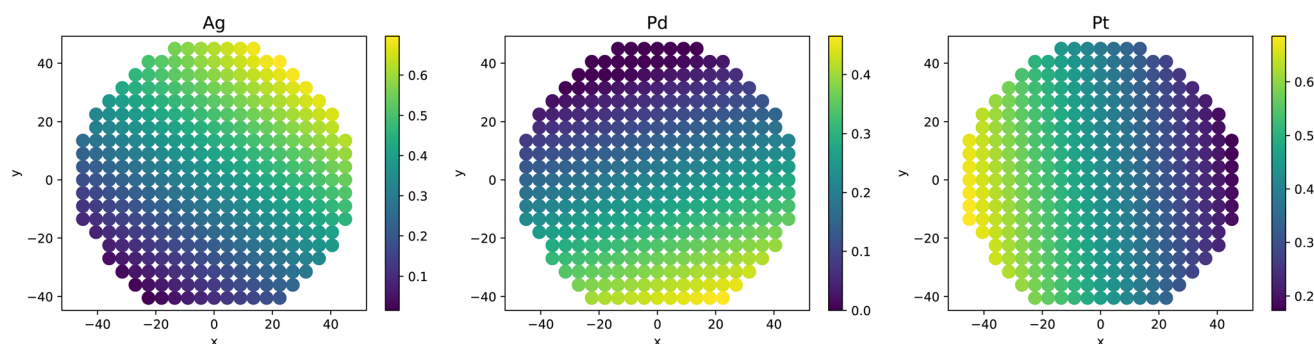
Fig. 8 Color-coded plot of compositional gradients in Ag−Pd−Pt dataset.

values decreased to 0.03 mA cm^{-2} and 0.04 mA cm^{-2}, respectively, further confirming the improved accuracy of the model using material vectors.

3.4 Results of method 3: standard vector method

The standard vector approach which uses weighted vector representations of material properties results in very promising improvements of the prediction (Table 3 and Fig. 10(a, d), S1 and S2†). Specific statistical metrics are not provided for this model such as R^2, MAE, or RMSE because the mode does not predict the current directly but a similarity measure which strongly correlates with the currents at 0.94. This value proves a significant correlation with the quaternary system's material performance, particularly in predicting lower electrochemical currents, that is predicting compositions with higher electrocatalytic performance, which are promising candidates for experimental assessment.

Fig. 11 shows all model predictions in comparison to the experimental data discarding outliers above a threshold of -0.075 mA cm^{-2} along a line across the CSML from the minimum to the maximum of the activity. The location of the measured data points are shown as gray background markers, the color-coded line represents the continuous interpolation of

current values across this direction. Fig. 11(b) shows the predictions from the three models along the interpolated measured data. It is notable that the GP model captures the non-linear behavior of the data more effectively while the standard vector method exhibits noticeable deviations w.r.t. the trend across the CSML.

4 Discussion

4.1 Interpretation of results

The outcomes of our study demonstrate that the choice of representation in computational models is critical for prediction performance. Model 1, GP based on elemental composition provides a reference prediction. However, its comparatively lower predictive accuracy (R^2 of 0.08) suggest complex (nonlinear) interplay of composition and catalytic performance in the quaternary system, where interactions between elements may not be fully captured using only a compositional representation.

Model 2, the GP model based on word-embedding based representations of materials, shows a significant improvement in predictive accuracy (R^2 of 0.65). We attribute this improvement to the latent knowledge captured through word
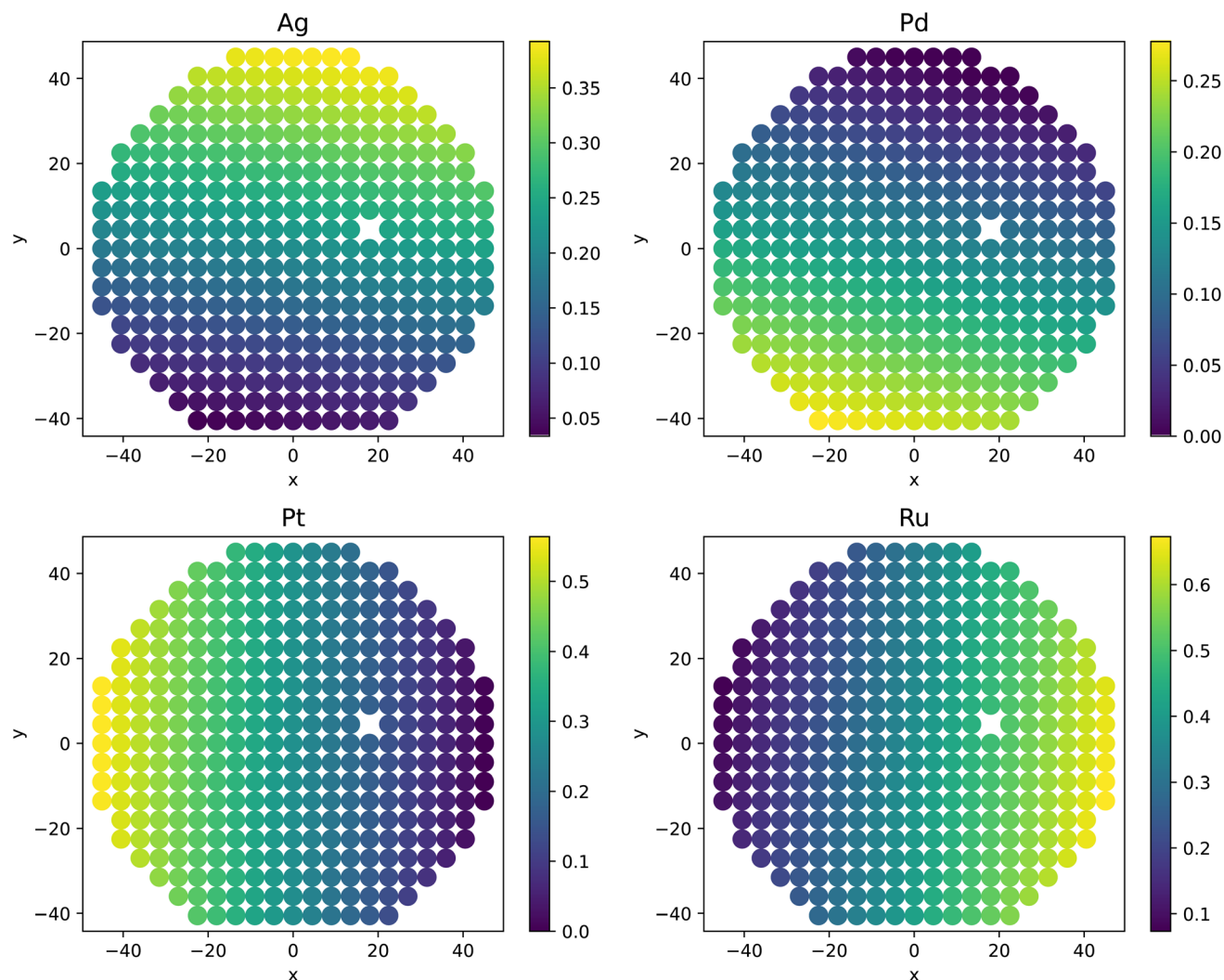
Fig. 9 Color-coded plot of compositional gradients in Ag–Pd–Pt–Ru system.

Table 3 Consolidated statistical analysis of actual *vs.* predicted electrochemical currents across different models

Metric	Gaussian process (GP)	GP with embeddings	Standard vector method
Mean (actual) (mA cm^{-2})	−0.16	−0.16	−0.16
Mean (predicted) (mA cm^{-2})	−0.22	−0.15	—
Standard deviation (actual) (mA cm^{-2})	0.07	0.07	0.07
Standard deviation (predicted) (mA cm^{-2})	0.07	0.05	—
Minimum (actual) (mA cm^{-2})	−0.37	−0.37	−0.37
Minimum (predicted) (mA cm^{-2})	−0.35	−0.07	—
Mean absolute error (MAE)	0.06	0.03	—
Root mean square error (RMSE)	0.07	0.04	—
Overall coefficient of determination (r^2)	0.08	0.65	—
Overall correlation (r)	0.85	0.83	0.80
Correlation (r) for current < −0.2 mA cm^{-2}	0.63	0.60	0.94

embedding representations of the compositions. It demonstrates that the complex interactions between materials beyond elemental composition can be captured in representations and effectively used for prediction.

Model 3, the standard vector approach, further exploits relationships of word embeddings by not directly predicting performance but instead focusing on the optimization of a similarity measure between materials vector representation and a standard
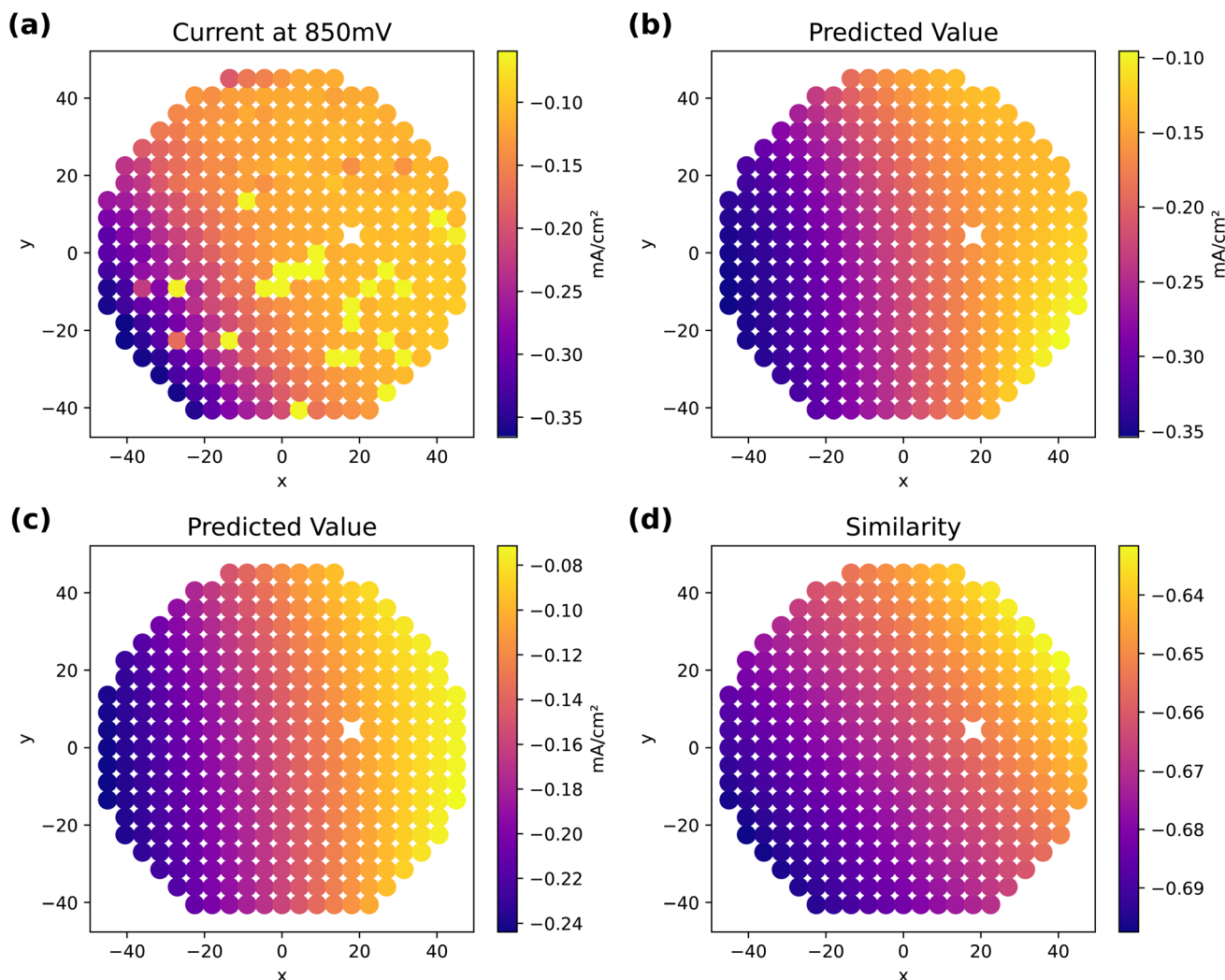
Fig. 10 Experimental results of Ag−Pd−Pt−Ru system (a) and prediction results using GP model (b), enhanced GP model with material vectors (c) and standard vector method (d).
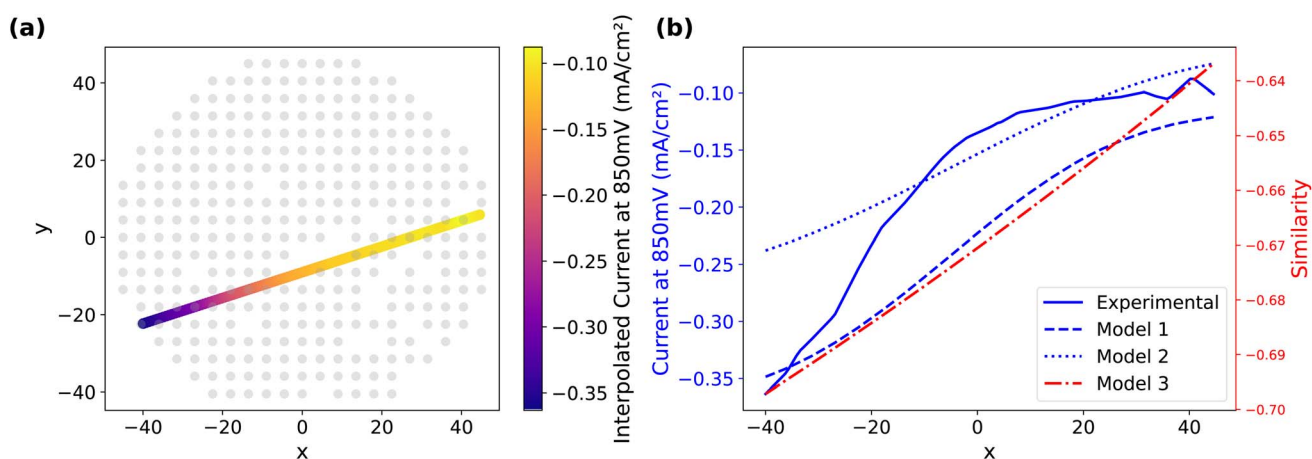

Fig. 11 Interpolation results across the whole dataset: (a) illustration of used line from the maximum and minimum current values with inter-polated results, (b) experimental data and predictions from all models along the direction indicated in (a).

vector based on known correlations of certain terms with electrocatalytic performance and experimental data from the two ternary systems. The high correlation (r) of 0.94 for specific conditions indicates a success, emphasizing the method's capability to identify potential high-performing materials within a defined extrapolation space. Our approach highlights the potential of using latent knowledge from scientific literature about materials and their relationships and represents a new approach for the representation of materials in combination with experimentally measured data. Nevertheless, the word embedding-based material representation and the standard vector method offer greater flexibility. Unlike the GP model, which is fixed to the specific dataset, in particular its elements, the other approaches are applicable to other material compositions. Future work will explore non-linear combinations which likely improve the accuracy of the proposed standard vector approach.

4.2 Comparison with existing literature

Our findings resonate with and extend existing research in materials science, particularly the use of machine learning and vector-based representations for materials prediction.[32,33] Several studies have demonstrated the potential of machine learning models, especially those incorporating innovative data representations, to outperform traditional computational methods.[34] Our work aligns with these findings, showcasing the effectiveness of material vectors for capturing complex interactions. However, we introduce a unique focus on similarity measures combined with word embedding-derived representations of materials, a less explored approach within materials property predictions.

4.3 Advantages of the proposed methods

Word embedding-based representations are directly combined with experimental data to predict unknown, more complex composition-property spaces. By using latent knowledge encoded in word embeddings we counterbalance data scarcity typically prevalent in experimental discovery campaigns, thereby accelerating the discovery process.

Our standard vector approach introduces a novel approach by focusing on 'similarity' rather than direct prediction. Our method's success in identifying high-performing materials based on their similarity to an optimized standard vector highlights based on experimental data is a tool for material selection and discovery, especially in systems where direct performance data may is scarce or hard to predict because of yet-unknown correlations. In our approach, we combine reliable but expensive-to-obtain experimental data with the fuzzy but cheap-to-obtain correlations in word embeddings. Our 'standard vector' can be viewed as a electrocatalysis-specific sequence of materials features[35] for specific materials systems and is particularly useful in scenarios where data is scarce.

4.4 Limitations and challenges

While our methods demonstrate significant advancements, they are not without limitations. For one, the word embeddings depend on the corpus from which they are built. We have restricted ourselves to literature with open access licenses. More text data, *e.g.* from copyright-protected material, could in principle improve word embeddings. The past and current publishing routes, however, restrict usage of the knowledge in literature without special agreements with publishers. Second, we rely on comprehensive and accurately labeled (ideally experimental) datasets for training the models and finding the 'standard vector'. This, in general poses a challenge, particularly in material science, where experimental data can be scarce, incomplete, or inconsistent. Additionally, the complexity of the models, especially the standard vector approach, may introduce difficulties in interpretation and implementation, potentially limiting their accessibility for broader application.

Future research will focus on addressing these limitations, possibly through the development of more robust models that can handle more sparse or noisy data, and the exploration of methods to simplify model interpretations without sacrificing prediction accuracy.

4.5 Implications for future research

Our study highlights the usefulness of material vectors based on word embeddings and similarity measures for predicting material performance, paving the way for advancements in materials prediction for under-explored compositional spaces where partial high-quality data already exists. Here are specific directions for future research:

4.5.1 Integration with experimental approaches. Combining these computational methods with targeted experimental validation can lead to iteratively more refined models and accelerated materials discovery. Experiments can verify predictions, identify regions where models need improvement, and provide new data to further enhance predictive power to include elements for predictions of different properties.

4.5.2 Hybrid models. Combining our methods with other predictive techniques like *ab initio* simulations or machine learning algorithms[36] could create more robust hybrid, multimodal models. These models could leverage the strengths of different approaches, potentially addressing shortcomings and enhancing predictive accuracy across varied, multimodal datasets.

4.5.3 Complex material systems. The success shown in this study encourages applying these methods to other properties of complex material systems. These could include structural, energy storage, magnetic properties, *etc.*, *i.e.* any system where properties are mainly a function of composition and not of microstructure. In contrast to composition-based models as presented here, the word embedding-derived representations allow arbitrary choices of elemental combinations. We expect that the near future will allow to use more experimental data for refinement of 'standard vectors'. Provided more reliable data for specific composition-property relationships is be available, 'standard vectors' for specific use cases could be defined as references against which new compositions could be assessed. New compositions could then be judged w.r.t. (theoretical) suitability be useful for a specific use case. If several such standard vectors can be defined, new compositions could be assessed for their suitability for multi-functional purposes.

5 Conclusions

Our study has successfully demonstrated the potential of machine learning and vector analysis techniques in predicting materials performance in ternary and quaternary compositionally complex solid solutions based on parameter-free Gaussian Process (GP) and literature-derived materials representations. The use of a GP model with elemental composition established a baseline for predictive accuracy, achieving a coefficient of determination value (r^2) of 0.08. An improved version of the GP model based on material vectors as representations for the composition derived from literature mining marks a significant improvement, with an improved r^2 value of 0.65. However, the most notable advancement was achieved with our proposed similarity vector approach. This method, which relies on the construction and optimization of property vectors, demonstrates a remarkable correlation with experimental outcomes, evidenced by a correlation value of 0.94. The superior performance underscores the potential of word embedding-based methods to leverage knowledge and material correlations from existing literature.

Data availability

1. Implementation of MatNexus: The code for the implementation of MatNexus, which supports the findings of this study, can be found at on Zenodo at **https://doi.org/10.5281/zenodo.15406293**. 2. Model Predictions: Scripts for this paper are publicly accessible on Zenodo at **https://doi.org/10.5281/zenodo.15407349**. 3. Raw Experimental Data: The raw experimental data supporting this study are publicly available on Zenodo at **https://doi.org/10.5281/zenodo.13992986**. For additional information regarding dataset access or specific use, please contact the corresponding author.

Author contributions

Lei Zhang: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, visualization, experimentation, funding acquisition. Lars Banko: Synthesis and characterization of materials libraries, pre-processing of the dataset, editing Wolfgang Schuhmann: Electrochemical experimentation, supervision, writing – review. Alfred Ludwig: Conceptualization, writing – review & editing. Markus Stricker: Conceptualization, resources, supervision, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Notes and references

1 Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman and M. S. Strano, *Nat. Nanotechnol.*, 2012, **7**, 699–712.
2 C. Suryanarayana, *Prog. Mater. Sci.*, 2001, **46**, 1–184.
3 T. A. Batchelor, J. K. Pedersen, S. H. Winther, I. E. Castelli, K. W. Jacobsen and J. Rossmeisl, *Joule*, 2019, **3**, 834–845.
4 E. D. Leshchenko, M. Ghasemi, V. G. Dubrovskii and J. Johansson, *CrystEngComm*, 2018, **20**, 1649–1655.
5 S. Back, A. Aspuru-Guzik, M. Ceriotti, G. Gryn'ova, B. Grzybowski, G. H. Gu, J. Hein, K. Hippalgaonkar, R. Hormazabal, Y. Jung, S. Kim, W. Y. Kim, S. M. Moosavi, J. Noh, C. Park, J. Schrier, P. Schwaller, K. Tsuda, T. Vegge, O. A. von Lilienfeld and A. Walsh, *Digital Discovery*, 2024, **3**, 23–33.
6 Z.-W. Zhao, M. del Cueto and A. Troisi, *Digital Discovery*, 2022, **1**, 266–276.
7 D. Persaud, L. Ward and J. Hattrick-Simpers, *Digital Discovery*, 2024, **3**(2), 281–286.
8 S. Durdy, M. W. Gaultois, V. V. Gusev, D. Bollegala and M. J. Rosseinsky, *Digital Discovery*, 2022, **1**, 763–778.
9 S. Lu, B. Montz, T. Emrick and A. Jayaraman, *Digital Discovery*, 2022, **1**, 816–833.
10 S. JOHANSEN, *Econometrica*, 1991, **59**, 1551–1580.
11 H.-M. Lu, J.-S. Chen and W.-C. Liao, *IEEE Trans. Knowl. Data Eng.*, 2021, **33**, 2669–2679.
12 W. Menke and R. Creel, *Surv. Geophys.*, 2021, **42**, 473–503.
13 E. Einarsson, O. Wodo, P. C. Nalam, S. R. Broderick, K. G. Reyes, E. Bruce Pitman and K. Rajan, *MRS Adv.*, 2020, **5**, 293–303.
14 B. Bayerlein, T. Hanke, T. Muth, J. Riedel, M. Schilling, C. Schweizer, B. Skrotzki, A. Todor, B. Moreno Torres, J. F. Unger, C. Völker and J. Olbricht, *Adv. Eng. Mater.*, 2022, **24**, 2101176.
15 J. Damewood, J. Karaguesian, J. R. Lunger, A. R. Tan, M. Xie, J. Peng and R. Gómez-Bombarelli, *Annu. Rev. Mater. Res.*, 2023, **53**, 399–426.
16 F. Thelen, L. Banko, R. Zehl, S. Baha and A. Ludwig, *Digital Discovery*, 2023, **2**, 1612–1619.
17 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lala, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mourino, B. Pelkie, M. Pieler, M. C. Ramos, B. Rankovic, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Voelker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik, *Digital Discovery*, 2023, **2**, 1233–1250.

18 L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. Persson, G. Ceder and A. Jain, *J. Chem. Inf. Model.*, 2019, **59**, 3692–3702.

19 T. Gupta, M. Zaki and N. M. A. Krishnan, Mausam, *npj Comput. Mater.*, 2022, **8**, 102.

20 L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster and A. Jain, *Comput. Mater. Sci.*, 2018, **152**, 60–69.

21 A. Subramanian, K. P. Greenman, A. Gervaix, T. Yang and R. Gomez-Bombarelli, *Digital Discovery*, 2023, **2**, 1006–1015.

22 G. K. P. Dathar, W. A. Shelton and Y. Xu, *J. Phys. Chem. Lett.*, 2012, **3**, 891–895.

23 J. K. Pedersen, C. M. Clausen, O. A. Krysiak, B. Xiao, T. A. A. Batchelor, T. Loeffler, V. A. Mints, L. Banko, M. Arenz, A. Savan, W. Schuhmann, A. Ludwig and J. Rossmeisl, *Angew. Chem., Int. Ed.*, 2021, **60**, 24144–24152.

24 G. Bampos, S. Tsatsos, G. Kyriakou and S. Bebelis, *J. Electroanal. Chem.*, 2023, **928**, 117008.

25 C. M. Clausen, O. A. Krysiak, L. Banko, J. K. Pedersen, W. Schuhmann, A. Ludwig and J. Rossmeisl, *Angew. Chem., Int. Ed.*, 2023, **62**, e202307187.

26 Y. Zhang and X. Xu, *Phys. Lett. A: Gen. At. Solid State Phys.*, 2020, **384**, 126500.

27 J. Wang and M. M. Molla Jafari, *Int. J. Chem. Eng.*, 2022, **2022**, 8264297.

28 L. Zhang and M. Stricker, *SoftwareX*, 2024, **26**, 101654.

29 V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder and A. Jain, *Nature*, 2019, **571**, 95–98.

30 L. van der Maaten, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, Hilton Clearwater Beach Resort*, Clearwater Beach, Florida USA, 2009, pp. 384–391.

31 E. Antolini, *Energy Environ. Sci.*, 2009, **2**, 915–931.

32 G. Huang, Y. Guo, Y. Chen and Z. Nie, *Materials*, 2023, **16**, 5977.

33 Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu and J. Hu, *Comput. Mater. Sci.*, 2020, **171**, 109203.

34 X.-l. Tian, S.-w. Song, F. Chen, X.-j. Qi, Y. Wang and Q.-h. Zhang, *Energ. Mater. Front.*, 2022, **3**, 177–186.

35 K. Rajan, *Annu. Rev. Mater. Res.*, 2015, **45**, 153–169.

36 C. M. Clausen, J. Rossmeisl and Z. W. Ulissi, *J. Phys. Chem. C*, 2024, **128**(27), 11190–11195.