Corrosion Science 255 (2025) 113080

Contents lists available at ScienceDirect

Corrosion Science

journal homepage: www.elsevier.com/locate/corsci

Large language models predicting the corrosion inhibition efficiency of magnesium dissolution modulators

Matthias Busch [a,*], Marius Tacke [c], Sviatlana V. Lamaka [b], Mikhail L. Zheludkevich [b], Kevin Linka [a], Christian J. Cyron [a,c], Christian Feiler [b,*], Roland C. Aydin [a,c,**]

[a] Institute for Continuum and Material Mechanics, Technical University of Hamburg, Eißendorfer Straße, Hamburg, 21073, Hamburg, Germany
[b] Institute of Surface Science, Helmholtz-Zentrum Hereon, Max–Planck-Straße, Geesthacht, 21502, Schleswig-Holstein, Germany
[c] Institute of Material Systems Modeling, Helmholtz-Zentrum Hereon, Max–Planck-Straße, Geesthacht, 21502, Schleswig-Holstein, Germany

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.14882459

Keywords:
Large language model
Machine learning
Material science
Corrosion
Magnesium
Prediction

ABSTRACT

Large language models (LLMs), such as GPT-4o, have shown promise in solving everyday tasks and addressing fundamental scientific challenges by leveraging extensive pre-trained knowledge. In this study, we investigate their potential to predict the efficiency of various organic compounds in inhibiting the corrosion of the magnesium alloy ZE41. Traditional approaches, such as Multilayer Perceptrons (MLPs), rely on non-contextual data, often necessitating large datasets and substantial effort per sample to achieve accurate predictions. These methods particularly struggle with small datasets as their training data and domain of applicability is limited to a small area of the available chemical space. LLMs can contextualize and interpret limited data points by drawing on their vast knowledge, including the chemical properties of molecules and their influence on corrosion processes in other materials (e.g. iron and aluminium). By prompting the model with a small dataset, LLMs can provide meaningful predictions without the need for extensive training. Our study demonstrates that LLMs can predict corrosion inhibition outcomes and outperform classical approaches, such as MLPs, having access to the identical number of training samples.

1. Introduction

In recent years, large language models (LLMs) have gained significant attention in both public discourse and academic research, evolving from their original purpose in natural language processing to demonstrating remarkable versatility across various domains [1,2]. Initially designed to handle tasks such as answering general questions, assisting in programming, and performing simple reasoning tasks, LLMs have since been applied to more complex challenges, including regression, prediction, and classification of numerical datasets.

To enhance the quality and accuracy of LLM-generated responses, researchers have developed strategies like prompt engineering and advanced techniques such as chain of thought (CoT [3]), tree of thought (ToT [4]), and skeleton of thought (SoT [5]). These methods enable LLMs to produce structured and well-reasoned outputs by allowing them to engage in more reflective and self-guided problem-solving processes. This growing recognition of LLMs as powerful tools has spurred their exploration in fields beyond traditional language processing, including chemistry.

The idea of using natural language processing (NLP) based machine learning models for chemistry tasks is among the more recent research topics. Tetko et al. [6], for example, used models based on the transformer architecture trained on SMILES strings [7] to predict synthesis steps for molecules. A step further goes the idea to use LLMs that are not trained for one purpose with a special vocabulary, but instead for general NLP tasks. The application of LLMs in chemistry, in particular, has emerged as a promising area of research, with efforts focused on how these models can effectively leverage chemical knowledge. Researchers have employed various approaches to enhance LLM performance in this domain. For example, Hatakeyama-Sato et al. [8] examined GPT-4's ability to process and apply chemical knowledge through prompt-based interactions, identifying both its potential and its limitations in reasoning and knowledge gaps. In addition, Jablonka et al. [1] reviewed techniques like fine-tuning GPT-2 and GPT-3.5 and connecting GPT-4 with external tools to improve its utility in chemistry-related tasks. Advancing this work, Bran et al. [9] developed ChemCrow, a tool that integrates simulations, internet access, and chemical APIs

* Corresponding authors.
** Corresponding author at: Institute for Continuum and Material Mechanics, Technical University of Hamburg, Eißendorfer Straße, Hamburg, 21073, Hamburg, Germany.
E-mail addresses: matthias.busch@tuhh.de (M. Busch), christian.feiler@hereon.de (C. Feiler), roland.aydin@hereon.de (R.C. Aydin).

https://doi.org/10.1016/j.corsci.2025.113080
Received 6 March 2025; Received in revised form 30 April 2025; Accepted 1 June 2025
Available online 20 June 2025

to extend LLM capabilities in molecular transformations and property predictions. Moreover, Jablonka et al. [10] demonstrated the potential of fine-tuned GPT-3 models in tasks such as classification, regression, and molecule design, while Zhang et al. [11] trained a domain-specific LLM, ChemLLM, on a large chemical database, showcasing the value of specialized models in chemical research.A very recent assessment of the capabilities of fine-tuned LLMs was done by Van Herck et al. [12]. This study provides an overview of the performance of fine-tuned open source LLMs on a variety of datasets from different fields in chemistry and material sciences. The authors showed that fine-tuning LLMs enables them to outperform classical machine learning approaches in various classification tasks.

Magnesium (Mg) and its alloys are increasingly used in various novel industrial applications due to their abundance, affordability, and versatility. However, the high chemical reactivity of magnesium requires domain-specific adjustments of its degradation behavior. In transportation [13,14], corrosion must be prevented to avoid critical material failure. In medical applications (e.g. temporary biodegradable stents or bone screws [15,16]), the degradation rate must be precisely controlled to match varying treatment or patient healing rates. For energy applications like Mg-air primary [17,18] and secondary batteries [19], a steady rate of Mg dissolution is essential to ensure constant energy output [20]. Several strategies, including alloying and surface coatings, have been developed to regulate the degradation of Mg-based materials [21,22]. A promising approach involves the use of small organic molecules, which have shown significant potential in controlling the dissolution properties of pure Mg and its alloys [23,24]. The vast chemical diversity of organic modulators is a major advantage, offering nearly limitless possibilities for tailored solutions. The number of available organic compounds is rapidly growing, with 120 million new compounds reported in the last decade alone. Estimates suggest that there could be as many as 10^{63} [25] organic compounds with useful properties, making the chemical space effectively infinite. Recent advancements in automation, robotics, and combinatorial chemistry are enabling the synthesis of larger and more diverse chemical libraries, and the integration of computer-assisted synthesis methods further accelerates the growth of available compounds [26].

However, even with the most sophisticated high-throughput techniques available today [27–29], researchers can only explore a tiny fraction of this space. Fortunately, quantitative structure–property or structure–activity relationships (QSPR, QSAR) techniques have emerged as tools to efficiently search larger regions of chemical space with significantly less time and effort, rendering them invaluable for shortlisting molecules with desirable properties for specific applications [30–37], but require extensive, reliable, chemically diverse, and balanced training datasets to achieve accurate and generalizable predictions.

Unfortunately, the available training data in the field of corrosion modulator research is quite limited from a machine learning point of view as a large part of chemicals that are labeled with an IE value are proprietary data and not available to the public domain. Fortunately, LLMs show great potential to mitigate the lack of available training data by utilizing their extensive base knowledge to understand contextual data and enhance model predictions. By drawing on a vast repository of scientific information, LLMs can help bridge gaps in existing datasets, supporting more accurate and reliable predictions of inhibition efficiency without the immediate need for additional experimental data.

This study compares and evaluates various prediction approaches for the corrosion inhibiting effect of small organic molecules for magnesium that utilize LLMs. The models include general models, such as o1 and Llama-3.1-405B, as well as two specialized LLMs, ChemLLM and ChemBERTa. These models are provided with exemplary data points and compared to baseline results from a basic neural network architecture that we have previously published. The baselines consist of an MLP from Schiessler et al. [38].

We hypothesize that, with LLM-based prediction approaches, replacing the full set of input data for the MLP with only the molecule names and their SMILES strings [7] can lead to at least the same prediction accuracy on the test set. This would simplify building a feature prediction model since no additional input data would be needed. We further hypothesize that if the molecule names and SMILES strings are combined with the MLP input data, the prediction results may be further improved by leveraging the pre-trained knowledge of the LLMs. This approach is particularly important given that the dataset contains only 75 samples, and increasing its size requires significant effort. Moreover, we performed a number of sanity checks to eliminate the risk that the training corpus of the used LLMs contains the original manuscript that is used as a benchmark in this work.

2. Results and discussion

Naturally, the small size of the dataset with only 75 samples raises concerns about the generalizability of the findings. To test the generalizability of the LLM-based approaches, the best performing variants are applied to four more corrosion datasets, for which the results can be found in the supplementary information. Additionally, our hypothesis is that the approaches presented in this study mitigate this risk due to the fact that the LLMs they are based on have been pre-trained on very large and versatile training data and are not specifically built for this prediction task. In this context, we want to emphasize that the predictive models were developed as if the blind test set did not exist. By withholding the blind test set from all pre-processing and training steps for all models that are evaluated in this work it is ensured that the blind test set probes the generalizability of the models.

The prediction approaches that we will probe in the next paragraphs are categorized into three groups: non-LLM approaches used to generate reference results (baselines), approaches using LLMs that have been fine-tuned on chemical texts/data (specialized LLMs) and approaches that rely on the use of general LLMs without fine-tuning (general LLMs). The main text mainly focuses on the error metrics root mean squared error (RMSE), mean absolute error (MAE) and correlation (Pearson r) to compare different approaches. In the supplementary information, sample based comparisons can be found, such as box plots for prediction distributions.

2.1. Prediction accuracy of baseline models

Three non-LLM approaches were chosen as baselines. The first one is taken from a recent paper, Schiessler et al. [38], and the data that was used within it is used as basis to test our hypotheses. The other two baseline models are implemented using the Python library Chemprop [39,40], which builds a neural network with molecular graphs as input. The first Chemprop implementation relies solely on the molecular structure for prediction, while the second architecture additionally uses the descriptors that the MLP also receives as input.

In Fig. 1, the three metrics – root mean squared error (RMSE), mean absolute error (MAE), and correlation – are depicted for the three baseline approaches. It can be observed that the MLP performs similarly to the second Chemprop implementation, except in the correlation metric, where the Chemprop implementation with descriptors slightly outperforms the neural network-based approach from our recent work Schiessler et al. [38]. However, the first Chemprop implementation, which only uses molecular graphs as input data without the descriptors, performs significantly worse than the other approaches. Using the blind test set that was selected to compare different machine learning approaches in this study, it is apparent that a model using molecular graphs as input does not outperform a traditional neural network approach that relies on distinct molecular descriptors as input features. Adding these descriptors as additional information to the input vector of the graph-based model improves the correlation between predicted and experimental values. However, this slight bump in the correlation

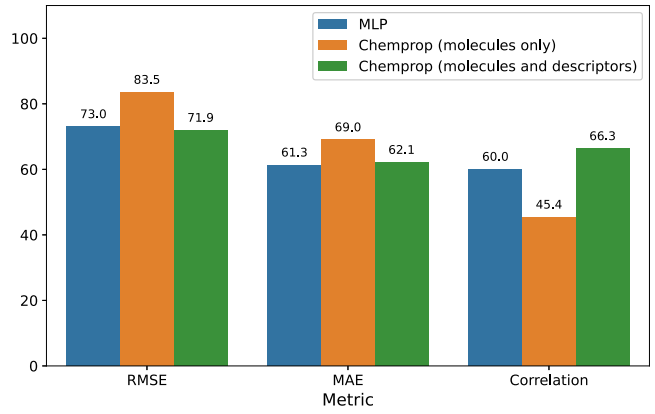

Fig. 1. Overview of the baseline results using different evaluation metrics. The MLP results are sourced from Schiessler et al. [38]. The correlation metric indicates the relationship between the predictions and the experimental values, where higher values signify better results (right). Conversely, lower values for the error metrics RMSE (center) and MAE (left) denote improved performance.

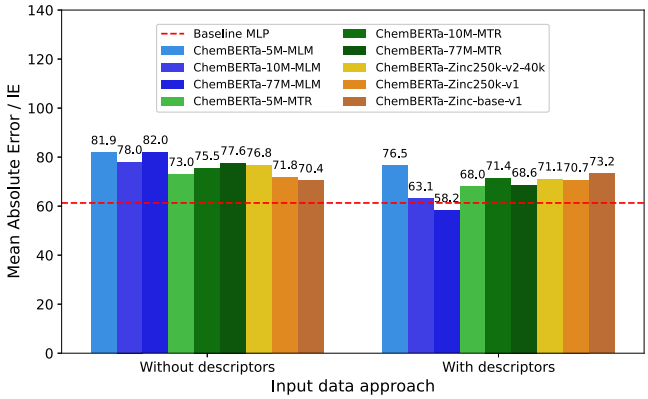

Fig. 2. Overview of the mean absolute errors (MAEs) for the predictions of the ChemBERTa approaches using varying sets of parameters. The red dashed line represents the MLP results from our previous study for direct comparison.

might not justify the additional effort of applying a graph-based neural network instead of directly using a classic MLP approach.

The reason for the lack of predictive power using the Chemprop approaches might be the large neural networks inside the model that are trained on a very small training set. One could potentially mitigate this by using transfer learning to build a feature extraction model for molecules and, in a second step, training a neural net on these features. These features could be more useful for the prediction task than the features extracted for the MLP. However, this is just an assumption and finding a dataset together with labels that result in a fitting feature extraction model can be difficult given that data in the domain of corrosion research is already scarce. Fortunately, in the LLMs that we will probe in the next chapter for their capability to predict corrosion inhibition efficiencies of small organic molecules, the transfer learning is automatically included because of their pre-trained knowledge.

2.2. Using domain-specific LLMs for predicting IEs

Since we are interested in predicting the impact of small organic molecules on the corrosion rate of ZE41, an intuitive choice is to probe the performance of LLMs that are tailored to the domain of chemistry. Here we chose ChemBERTa [41,42] and ChemLLM [11] to compete with our original neural network model.

ChemLLM is a larger model, prompted with variants containing 2 billion and 8 billion parameters. However, this model was unable to effectively analyze the training dataset or predict the corrosion inhibition efficiencies. The ChemLLM variants yielded near-zero or negative correlation and MAEs of 90–95, so we have omitted their plots from the main text. For reproducibility, all ChemLLM model code is included in our project repository (see Section "Code availability"). The results obtained from ChemLLM show that the prediction task is not trivial for language models. Knowledge about the molecules and the problem setting is not enough to make accurate predictions. For a prediction task, an LLM has to be able to analyze a large amount of data, find patterns in it and apply this knowledge to a test sample. ChemLLM seemingly does not have these abilities. The results of the second domain-specific LLM (ChemBERTa) are shown in Fig. 2. We fed ChemBERTa only the SMILES string and attached a regression head, trained on our data, that directly outputs predicted efficiencies. The names of the ChemBERTa variants indicate different training set sizes and modes. The parameter counts of the ChemBERTa variants range from 3.4 million (for the first 6 MLM and MTR variants) to 83 million (for the Zinc-250k versions). Their performance varied, and without the use of descriptors, they consistently produced significantly less accurate predictions than the baseline MLP from our previous work Schiessler et al. [38]. Incorporating descriptors into the input improved accuracy to a level comparable with the graph-based networks; only the "77M-MLM" variant slightly outperformed the baseline MLP in MAE. The other models were unable to achieve this level of accuracy. The results indicate that additional features (the ChemBERTa outputs) fed into an MLP (here the regression head) can decrease the prediction quality. This again raises the question of whether the effort required to compute molecular descriptors and feed them into an expensive LLM is justified, given that a simple neural network achieves similar accuracy. This serves as an important reminder that simply adding more training data is not a panacea—instead, judicious use of existing data is crucial for building robust models.

2.3. Putting general LLMs to the test

In the next chapter, we will put a number of more general LLMs to the test as they created the impression that they have an accurate answer to all questions in recent years. We selected four OpenAI models (GPT-3.5, GPT-4o, o1, and o3-mini) and Meta's open LLM Llama-3.1-405B for our experiments. For each model, four different prompting strategies were explored (for details see method Section 3.2.4). Each configuration was tested in two variants: firstly, feeding the models only molecular names and SMILES strings; and secondly – similar to the preceding chapters – with the additional set of molecular descriptors.

The MAEs of the results are depicted in Figs. 3 and 4. Generally, o1 or o3-mini, being the newest models, return the best results in every configuration except for the configuration with descriptors and with pre-analysis. Interestingly, this approach returns the overall best predictions, where GPT-4o outperforms o1 and o3-mini.

Meta Llama-3.1-405B

Meta's Llama-3.1-405B shows the lowest prediction accuracy without descriptors across all prompting strategies, with correlation values (supplementary information) at most around 0, indicating no capability to predict the impact of compounds on the magnesium corrosion process from the molecular structure. However, adding descriptor data to the prompt improves Llama's prediction quality across all prompting strategies, outperforming GPT-3.5 in three of four approaches and nearly reaching the baseline MLP with the pre-analysis prompting strategy. Notably, it outperforms the baseline MLP with respect to correlation (Fig. 5).

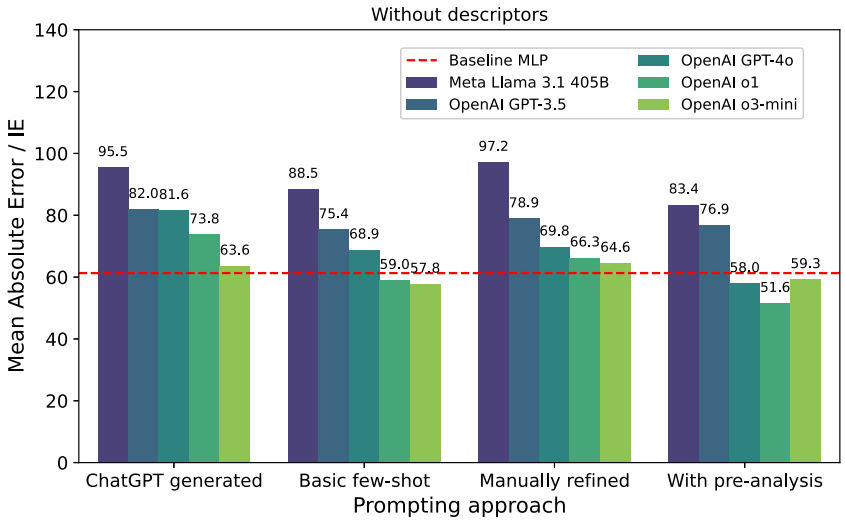

Fig. 3. Overview of the general LLM results: MAEs without the additional descriptor input. The red dashed line shows the MLP result for comparison. The prompting approaches denote one ChatGPT-written and three manually written prompting strategies. The basic few-shot approach does not include any helping steps, which the manually refined prompt does include. The pre-analysis denotes one prompt to preemptively analyze the training data before the prediction step.

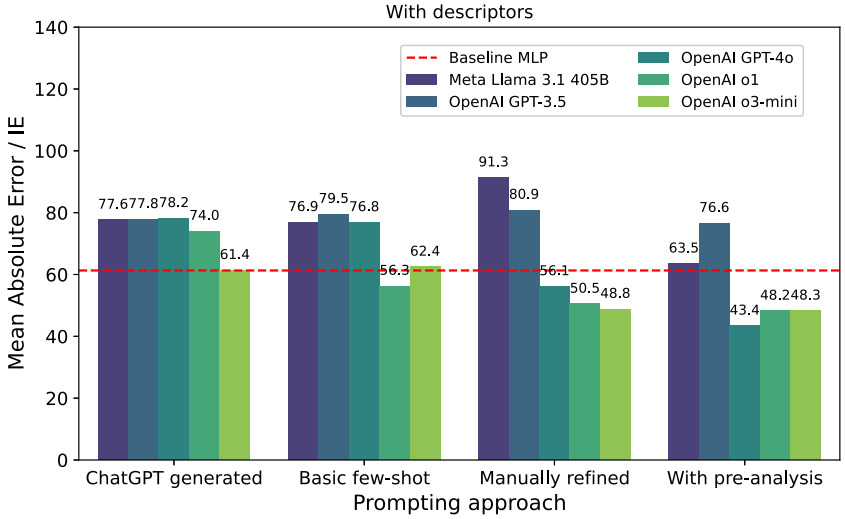

Fig. 4. Overview of the general LLM results: MAEs with the additional descriptor input. The red dashed line shows the MLP result for comparison. The prompting approaches denote one ChatGPT-written and three manually written prompting strategies. The basic few-shot approach does not include any helping steps, which the manually refined prompt does include. The pre-analysis denotes four prompts to preemptively analyze the training data before the prediction step.

GPT-3.5

GPT-3.5 consistently shows high MAE values and does not benefit from more sophisticated prompting approaches. Adding descriptors to the prompt does not change prediction quality, except for a slight increase in MAE. GPT-3.5 does not surpass the baseline in any metric with any approach. These results indicate that GPT-3.5 is not able to follow instructive prompts to be able to profit from more advanced prompting strategies.

GPT-4o

GPT-4o's performance varies considerably depending on prompting approach. Without descriptors, performance progresses from weakest with ChatGPT-generated prompts to strongest with the pre-analysis approach, which surpasses baseline results. With descriptors, results are mixed: ChatGPT-generated prompts show improvement but remain insufficient, while the basic few-shot approach surprisingly decreases in accuracy. The refined approach with descriptors exceeds baseline performance, and the pre-analysis approach yields the study's best MAE scores.

GPT-4o displays the highest sensitivity to prompting strategy, with pre-analysis prompting providing substantial improvements. This suggests limitations in context window utilization and data analysis capabilities without proper guidance, particularly for numerical data.

o1

o1 generally outperforms GPT-4o by 5–10 MAE IE in configurations without descriptors. With descriptors, performance differences vary more substantially. Unlike GPT-4o, o1 shows less dependency on prompting refinement, improving by only ~ 8 MAE IE from basic few-shot to pre-analysis approaches (compared to GPT-4o's ~ 34 MAE IE improvement). This indicates o1's superior capability to independently analyze molecular structures and formulate effective prediction workflows without extensive guidance.

While GPT-4o achieves the best overall MAE (43.4 vs. o1's 48.2) with optimized prompting, both models reach similar correlation scores (~ 86%). This suggests qualitatively similar predictions with different quantitative accuracy, possibly indicating GPT-4o's broader knowledge base versus o1's stronger reasoning abilities.

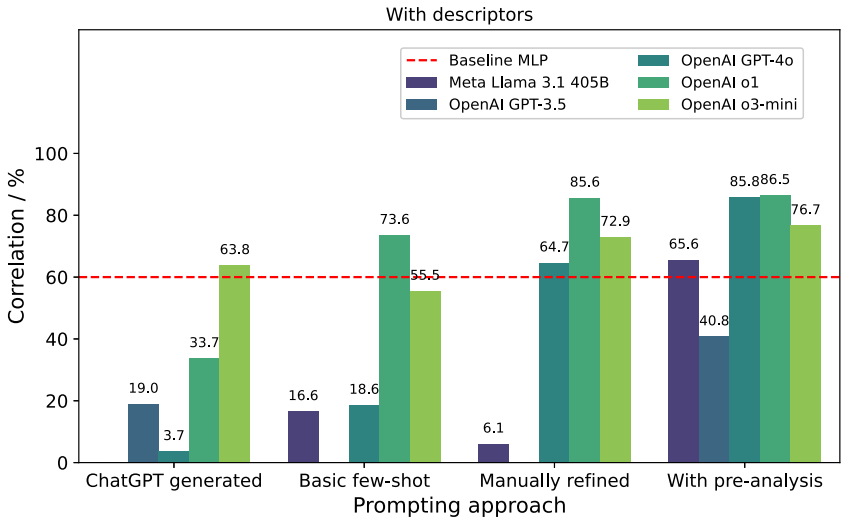

Fig. 5. Overview of the general LLM results: correlations with the additional descriptor input. The red dashed line shows the MLP result for comparison. The prompting approaches denote one ChatGPT-written and three manually written prompting strategies. The basic few-shot approach does not include any helping steps, which the manually refined prompt does include. The pre-analysis denotes four prompts to preemptively analyze the training data before the prediction step.

o3-mini

o3-mini shows remarkably consistent prediction accuracy across all prompting approaches without descriptors (MAE range 57–65). It outperforms other LLMs in three of four approaches but is still outperformed by o1 in the approach with pre-analysis. With descriptors, it performs similarly to o1. Most notably, o3-mini is the only model that approaches baseline MLP performance with ChatGPT-generated prompts, demonstrating the highest prompting-scheme independence. However, it does not match the peak performance of GPT-4o or o1, with correlation metrics approximately 10% lower than these models.

Comparative analysis of the prompting strategies

The evolution from GPT-4o to o1 to o3-mini shows progressively decreasing dependency on prompt optimization. As reasoning models, o1 and o3-mini's architectures focus on iterative answer improvement rather than raw parameter count, enabling more consistent task understanding regardless of prompt formulation. However, their reduced parameter count may limit access to specialized knowledge required for achieving optimal results in niche domains.

Within this study, prompts written by ChatGPT result in very inaccurate predictions, which are overly verbose and lack clear task structure—o3-mini is the exception, as it successfully interprets them. The basic few-shot approach performed better for most LLMs, especially o1, showing that a short, well-structured prompt leads to more accurate predictions than a long, non-optimized prompt with unnecessary tasks and information. Manually optimized prompts further improve the performance across LLMs, with GPT-4o benefiting most significantly (35 IE MAE reduction versus 14 IE for o3-mini). Additionally, prompt separation strategies substantially benefited GPT-4o but provided minimal improvement for the reasoning LLMs o1 and o3-mini.

2.4. Do they already know the answer to the question?

Their good prediction accuracy leads to the question of whether some of the LLMs might have been trained on the paper Schiessler et al. [38]. In this publication, a table with the names and labels of the blind test set samples is included and the publication date precedes the knowledge cutoff of OpenAI's newer GPT models. Therefore, the paper could be part of the training corpus of one or more of these models. However, no evidence was found to support this hypothesis. This was tested using greedy decoding to continue text excerpts from the publication. Furthermore, it is assumed, that if the LLMs were trained on the test set, the prediction error would be significantly lower.

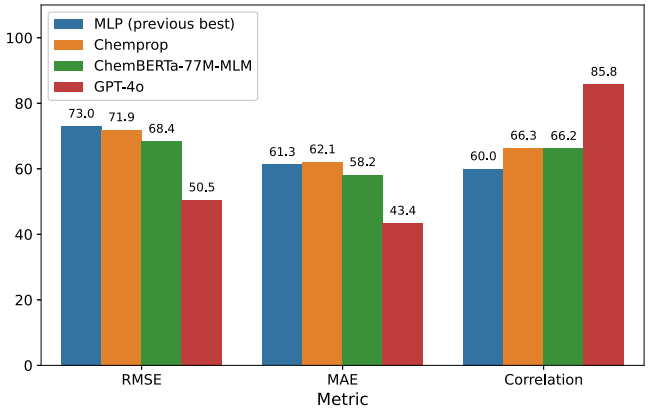

Fig. 6. Overview of the best results from each group using different evaluation metrics. The MLP results are sourced from Schiessler et al. [38]. The correlation metric indicates the relationship between the predictions and the experimental values, where higher values signify better results. Conversely, lower values for the error metrics RMSE (root mean squared error) and MAE (mean absolute error) denote improved model performance.

Notably, approaches using only names and SMILES strings should perform better, as the publication includes the names and IE values but lacks descriptors. However, unraveling the thought process of LLMs is one of the current limitations as there are no tools available yet that allow us to shed light on the black box.

2.5. Rise of the transformers?

Over the course of the preceding chapters, we compared multiple approaches for predicting the corrosion inhibiting effect of small organic molecule dissolution modulators for the Mg alloy ZE41. As it is easy to lose oneself in the myriad of tested variants, we will compare each group's best performing approach together with the chosen benchmark model from Schiessler et al. [38] (see Fig. 6).

As apparent from Fig. 6, neither the best performing chemical graph-based approach (RMSE = 71.9, MAE = 62.1, $r_{pearson}$ = 66.3%) nor the best performing chemistry domain-specific LLM ChemBERTa (RMSE = 68.4, MAE = 58.2, $r_{pearson}$ = 66.2%) significantly outperformed the MLP baseline model (RMSE = 73.0, MAE = 61.3, $r_{pearson}$ = 60.0%). It is noteworthy that the correlation between the second domain-specific

LLM (ChemLLM) and the experimental values was essentially zero. Finally, some of the approaches based on more recently published general LLMs (o1, o3-mini and GPT-4o) outperformed the MLP baseline significantly with GPT-4o being the best-performing model (RMSE = 50.5, MAE = 43.4, $r_{pearson}$ = 85.8%) whereas GPT-3.5 and Llama 3.1 405B did not outperform the benchmark MLP model by Schiessler et al. [38] under any prompting strategy.

2.6. Conclusion

This study demonstrates that using large language models (LLMs) like GPT-4o for prediction tasks on small datasets offers clear advantages (see Fig. 6). The predictions made by GPT-4o, especially when given additional contextual information, significantly outperform those from a traditional MLP as well as all other approaches that have been explored in this study. Several factors contribute to this result.

First, more recent LLMs (e.g. OpenAI's GPT-4o) already have substantial built-in knowledge, which makes them more efficient with training data. Similar to transfer learning, these later models do not need to learn basic scientific principles from scratch. LLMs understand molecules, their functional groups, and basic chemistry related to corrosion. This means the training samples are used only for the actual task—figuring out how the molecules affect corrosion.

Second, the ability to interpret input data in string format, like molecule names and their SMILES strings, improves the predictions. Even using only this data, some models perform better—e.g., o1 achieves an MAE of 51.6 compared to the MLP baseline's 61.3. When combined with the MLP descriptor input data, the accuracy improves further to an MAE of 43.4 for GPT-4o.

Third, the ability to generalize from pre-existing knowledge and training data plays a key role. Despite likely recognizing the molecules in the dataset, chemistry-specialized models like ChemBERTa and ChemLLM show limited predictive capability for corrosion inhibition. This indicates that domain knowledge alone is insufficient—effective reasoning abilities are equally important. Newer general LLMs significantly outperform chemistry-specialized alternatives. Most models can improve with well-structured prompting strategies, but the oldest models struggled to follow these instructions. By contrast, the reasoning models show less dependency on the prompt, because they are able to instruct themselves well enough. o1 outperforms the baseline using an unoptimized prompting strategy. Similarly, splitting the prompt mainly improves non reasoning LLMs, probably because reasoning LLMs can handle much larger context windows without detailed instructions.

The generalizability of the general LLM-based prediction approaches shows mixed results correlating to the representativity of the training set in terms of chemical space overlap with the respective test set. The results clearly show generalizability to different datasets, but with partly diminished correlation. The results, further analysis on the generalizability and an analysis of the applicability domain can be found in the supplementary information (chapters 1 and 2). Furthermore, the differences in prediction consistency are remarkable, especially when comparing LLM-based approaches with the baseline MLP. An in-depth analysis of a selection of the distributions is provided in the supplementary information (chapter 3), which shows that the variance of GPT-4o's predictions is much more inconsistent than that of the MLP.

Overall, the results show the power of LLMs in predicting corrosion inhibition efficiency on small datasets. The ability to use pre-trained knowledge, incorporate additional data, and generalize from training examples makes them highly effective for this type of task. Building an accurate model to screen the chemical space of potential corrosion modulators is an important aspect of corrosion science. With the screening, potential candidates for laboratory tests can be identified. A more accurate screening model reduces the number of costly, time-consuming laboratory tests. Likewise, training on a smaller base dataset minimizes the number of compounds that must undergo real-world experiments. However, this study does not yet relate to real world applications, as all the data is laboratory-scale. That means, the behavior of a corrosion modulator might change significantly when putting it into a coating.

2.7. Outlook

Nonetheless, tailored models based on traditional approaches like the MLP used as benchmark in this study or other supervised learning approaches based on random forests [43] or kernel principal covariates regression will remain important tools in the next decades as they allow to generate reliable models despite being restricted to a smaller domain of the chemical compound space compared to LLMs with their inherent capability to generalize better across large domains of chemical space [44].

As evidenced by this work, large language models will drive the next paradigm shift in materials research—pushing the limits of prediction accuracy in quantitative structure–property relationship models and related tasks, especially when provided with additional context (e.g., molecule names, SMILES strings, and experimental details). The LLM-based approaches presented in this work will synergize with iterative experimental campaigns to rationally increase the foundation for in-context learning by adopting active learning or design of experimental strategies.

Moreover, the development in multiple directions—just stating classical models like OpenAI's GPT-4o, reasoning models like o1 and mixture of experts models like DeepSeek's V-3—renders further improvements likely. Another avenue is fine-tuning on domain-relevant texts (e.g., publications) and larger datasets to further improve prediction accuracy.

3. Methods

3.1. Data

The dataset used for this study is sourced from Schiessler et al. [38]. It comprises 75 sample molecules tested for their ability to modulate the corrosion rate of magnesium in NaCl solution. For a more detailed description of the experimental setup see our recent works Schiessler et al. [38,45]. The corrosion modulation is quantified in percentage based on the inhibition efficiency (IE) metric. A value of 100% denotes complete inhibition of corrosion, while 0% indicates no effect. Negative values, which are unbounded, represent acceleration of corrosion compared to a benchmark measurement in absence of any corrosion modulating agent. The inhibition efficiency (IE) serves as the label for all machine learning models discussed in this paper. This study employs the feature set that yielded the best results in Schiessler et al. [38], which consists of five features selected solely based on the training set (the first 60 samples) to augment the tested LLM architectures. Moreover, the used LLMs are fed with the isomeric SMILES string representation of the molecules used in the preceding study.

3.2. Prediction models

3.2.1. Selection of methods

The first method is the baseline multilayer perceptron (MLP). This method is selected because it is a previously published method [38] which also includes published results for the corrosion dataset primarily applied in this study. Hence the results can serve as a baseline for our methods. For a second baseline that can ingest both molecular structure and descriptors, we adopted the D-MPNN from Chemprop. It was decided to not use transfer learning, as every approach gets access to the same input data to show what each approach can achieve without access to additional data. In reality, scientists might not have access to a large transfer learning database, but a pre-trained LLM.

ChemBERTa and ChemLLM are two chemistry-specific LLMs meaning they were trained on chemistry related texts. However, they are rather small models and their reasoning capabilities are diminished compared to the general LLMs. While ChemLLM is not able to follow simple prompt instructions, the smaller and older ChemBERTa is not

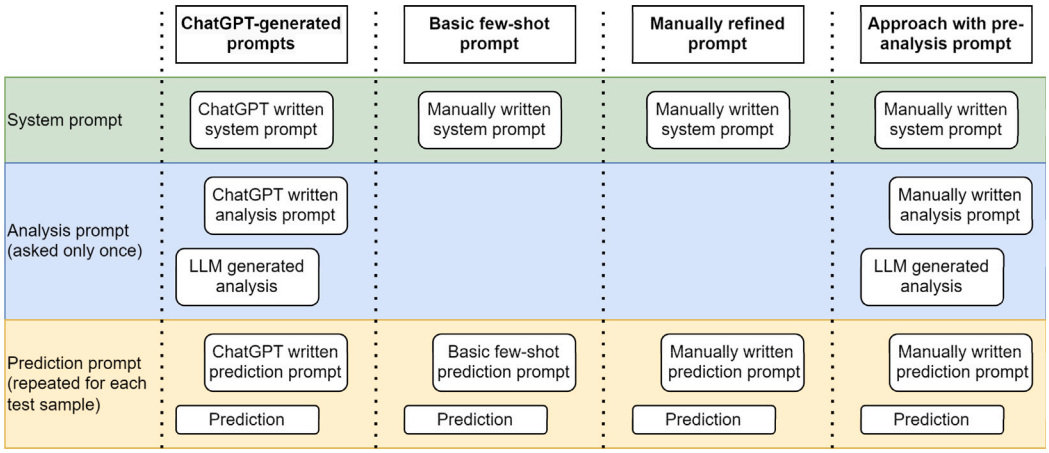

Fig. 7. Overview of the prompting strategies. With some exceptions, the prompting works like a chat starting at the top, where previous prompts and answers are also appended before the current prompt as chat history. This graphic shows the prompting schemes for the approaches without descriptors. When the descriptors are added to the input data, the pre-analysis prompt is expanded to a set of analyses with different focuses.

expected to perform better. Therefore, a different approach is used to incorporate its pre-trained knowledge via fine-tuning.

The main part of this study is the test of the prediction capabilities of general purpose LLMs without fine-tuning. It was decided to mainly use OpenAI's models, because they offer different variants of state-of-the-art technology. It is assumed that similar models from different providers achieve competitive results. As an open-model alternative, Meta's Llama-3.1-405B was chosen, which is the largest model of their series. The prompting approaches are divided into two unoptimized and two optimized approaches. The unoptimized approaches show what can be achieved without prompt engineering while the optimized prompts show how the result can be improved if fitted to the problem.

3.2.2. Baselines

The first baseline is derived from Schiessler et al. [38]. It is an MLP model that predicts the corrosion inhibition efficiency and is trained on molecular descriptors. The second and third baselines utilize a direct message passing neural network (D-MPNN) from Chemprop [39,40]. This architecture includes functionality that transforms SMILES strings into molecular graphs, which the D-MPNN is trained on. With this additional input, the model no longer relies solely on descriptors to encode molecular structure. However, it was found beneficial to include the five descriptors on which the MLP was trained. Consequently, two distinct approaches were adopted: one using only SMILES strings as input data, and another using both SMILES strings and the five descriptors.

3.2.3. Specialized LLMs

ChemBERTa [41,42] is based on the RoBERTa architecture and has approximately 3 to 100 million parameters, making it relatively small compared to other state-of-the-art large language models (LLMs). It lacks proper reasoning and reliable answering, making it less suitable for prediction tasks. Therefore, in this work, ChemBERTa was modified with a regression head to allow for the input of a molecule via a SMILES string and to utilize its pre-trained chemical knowledge while providing reliable predictions. Fine-tuning ChemBERTa with a special head layer is also suggested and implemented by Chitrananda et al. [41]. To further improve prediction accuracy, the regression head is modified to allow additional inputs for descriptors.

The input prompt is the SMILES string. Various variants were tested, but they did not lead to significant improvements on the training set. One variant worth mentioning involved slightly modifying the prompt to "SMILES: smiles_string" to explicitly indicate that the input is a SMILES string. Most results improved slightly; however, the only result that outperformed the baseline did not improve (as shown in the supplementary Figure 6).

ChemLLM is prompted in a few shot prompting mode to predict instead of being trained. The details are presented in the next Section 3.2.4. However, it is not able to follow prompt instructions effectively. For ChemLLM, only the basic few-shot prompting strategy was executed repeatedly, but produced uncorrelated predictions. Other strategies were attempted but rarely resulted in valid predictions. The post-processing of ChemLLM was adjusted to read the first number in the response as the prediction. This was done because, if a prediction was returned, it appeared in the first sentence. An example is: "I would conclude that −22". Most answers that did not include a prediction in the first sentence did not include it anywhere in the answer.

3.2.4. General LLMs

The general LLMs are prompted via Microsoft's Azure Cloud Services API to predict corrosion inhibition efficiency. The API provides access to various LLMs and also allows sending prompts with a custom chat history which is essential for more advanced prompting strategies. There is no training or fine-tuning process involved in these prediction models. The LLMs do not get access to tools like online search or Python libraries. The predictions they do purely come from their semantic reasoning process. An overview of the different strategies is provided in Fig. 7. A selection of prompts is shown in the supplementary Section 4 (Fully available in code repository). All prompting strategies include the training set of samples and are thus a variant of few-shot prompting. The manually written system prompt (used for all but the ChatGPT-generated prompt) includes a command to think step-by-step and is structured to enable better reasoning with chain-of-thought thinking. Four different prompting strategies are engineered to show differences in the LLMs' prediction capabilities:

- **ChatGPT-generated prompts:** a set of prompts structurally similar to the manually refined prompt with pre-analysis, but written by ChatGPT
- **Basic few-shot prompt:** a prompt showing the training set and one test sample with no further instructions other than to predict the label
- **Manually refined prompt:** a prompt similar to the basic few-shot prompt but with a custom set of pre-defined thinking steps to execute, performing an analysis and a prediction
- **Approach with pre-analysis prompt:** the manually refined prompt, split into one or multiple analysis prompts and a prediction prompt; not only split but also includes additional analysis steps

The first two prompting strategies serve as unoptimized baselines for comparison with the other two strategies, which were manually optimized on the training set. The "manually refined prompt" strategy features a single prompt that explicitly defines the processing steps to predict a corrosion value. By analyzing the training set and test sample before the prediction, an attempt is made to enhance the prediction quality. The prompt tells the LLM to analyze the molecular structure of the test sample and assess the influence of the findings on the inhibition efficiency by including the training set into the analysis. This prompt and the following prompts were manually refined for GPT-4o on the training set. The last prompting strategy features one (without descriptors) and four (with descriptors) analysis prompts. These prompts focus solely on analyzing each sample of the training set from different perspectives and expanding the dataset with additional information. Main points of the analysis are the expansion of the molecules by their functional groups, the reduction of the descriptor data to only 2–3 descriptors per molecule, and a pattern search. The prediction prompt then receives the output of the analyses as chat history and focuses on the analysis of the test sample and the prediction.

The post-processing of the LLM answers is done by reading in the last number found in the answer. The prediction is done one sample at a time, and in all prompting strategies apart from the basic few-shot prompting, there is an explicit command to write down the prediction as the last part of the answer. This approach works reliably if the LLM can follow the tasks. If it cannot follow the prompt tasks, the prediction is also not expected to be accurate. Problems arise for multiple LLMs, that are not able to follow the given pattern or reach a prediction value. For ChemLLM and GPT-3.5 this is generally the case, while for Llama-3.1-405B and GPT-4o this happens sometimes. Especially the ChatGPT-generated prompts seem to be problematic as they do not explicitly state that the prediction should be the last thing to write.

3.3. Reproducibility

To ensure consistency and reproducibility of the results, each prediction process is repeated multiple times, and the mean of the predictions is taken as the final prediction. Different approaches are repeated a varying number of times: the general LLM approaches are all repeated 20 times, while the ChemBERTa and Chemprop approaches are repeated 100 times. The baseline MLP approach from Schiessler et al. [38] was repeated 1000 times. Although the seed functionality available for some OpenAI models was tested, it unfortunately did not result in reproducible outcomes. A remark on the repetitions of the general LLM-based approaches: The whole approach is repeated, resulting in newly generated analyses each time.

The configurations and parameters for the general LLMs used in this study are shown in Table 1: Except o1 and o3-mini, the temperature is set to 0.7, the top_p value to 0.95 and the maximum number of tokens is set to 4000. For o1 and o3-mini, temperature and top_p cannot be set and the maximum number of tokens is set to 25000 (o1) and 100000 (o3-mini). Additionally, with o3-mini, a reasoning effort parameter was introduced, which was set to high. These adjustments are considered fair, as the longer context window and strong reasoning capabilities are strengths of the reasoning models o1 and o3-mini and also required because of the reasoning process. All LLMs are accessed using Microsoft Azure Cloud Services. The API version for Llama is set to "2024-05-01-preview", while the versions for the other general LLMs apart from o3-mini are set to "2024-10-01-preview". o1 is accessed as a preview model and is subject to change. o3-mini is accessed with the API version of "2024-01-01-preview".

ChemLLM is configured similarly to the general LLMs, with a temperature of 0.7, a top_p value of 0.95, and a maximum of 4000 tokens. It is utilized using the transformer library from Huggingface and executed on the cluster at TUHH. ChemBERTa is employed with default parameters, also via Huggingface.

Table 1
Parametrizations of the LLMs used in this study.

LLM	Temp.	Top_p	Max tokens	Platform
Llama-3.1-405B	0.7	0.95	4000	Microsoft Azure
GPT-3.5	0.7	0.95	4000	Microsoft Azure
GPT-4o	0.7	0.95	4000	Microsoft Azure
o1	–	–	25 000	Microsoft Azure
o3-mini (high)	–	–	100 000	Microsoft Azure
ChemLLM	0.7	0.95	4000	Huggingface
ChemBERTa	Default	Default	Default	Huggingface

Code availability

The code is available in the repository https://doi.org/10.5281/zenodo.14882459. The code is fully available there and can be executed if access to required computational resources exists. The Chemprop and ChemBERTa examples should also run without access to large computational resources.

CRediT authorship contribution statement

Matthias Busch: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Marius Tacke:** Writing – review & editing. **Sviatlana V. Lamaka:** Writing – review & editing, Funding acquisition. **Mikhail L. Zheludkevich:** Writing – review & editing. **Kevin Linka:** Writing – review & editing. **Christian J. Cyron:** Writing – review & editing, Funding acquisition. **Christian Feiler:** Writing – original draft, Visualization, Supervision, Conceptualization. **Roland C. Aydin:** Writing – review & editing, Supervision, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used o3-mini and ChatGPT in order to review the manuscript for errors and possible improvements. This includes rephrasing of parts of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This publication was funded via project 535656357 from Deutsche Forschungsgemeinschaft (DFG), Germany: https://gepris.dfg.de/gepris/projekt/535656357.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the support of the Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF), Germany.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.corsci.2025.113080.

Data availability

The data is available in the repository https://doi.org/10.5281/zenodo.14882459. The file is called "Old_Descriptors.csv" and contains the same data as published with Schiessler et al. [38]. The molecule names have been standardized with PubChem and a SMILES string column has been added ("isomeric_smiles").

References

[1] 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. Lála, T. Liu, S. Ma, S. Majumdar, G.W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M.C. Ramos, B. Ranković, S.G. Rodriques, J.N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B.E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G.A. Zia, A. Scourtas, K.J. Schmidt, I. Foster, A.D. White, B. Blaiszik, 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, Digit. Discov. 2 (2023) 1233–1250, http://dx.doi.org/10.1039/D3DD00113J.

[2] A. Plaat, A. Wong, S. Verberne, J. Broekens, N. van Stein, T. Back, Reasoning with large language models, a survey, 2024, arXiv:2407.11511. URL https://arxiv.org/abs/2407.11511.

[3] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P.S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H.D. Costa, S. Gupta, M.L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, P. Resnik, The prompt report: A systematic survey of prompting techniques, 2024, arXiv:2406.06608.

[4] S. Yao, D. Yu, J. Zhao, I. Shafran, T.L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, 2023, arXiv:2305.10601.

[5] X. Ning, Z. Lin, Z. Zhou, Z. Wang, H. Yang, Y. Wang, Skeleton-of-thought: Large language models can do parallel decoding, 2023, arXiv:2307.15337.

[6] I.V. Tetko, P. Karpov, R. Van Deursen, G. Godin, State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis, Nat. Commun. 11 (1) (2020) 5575, http://dx.doi.org/10.1038/s41467-020-19266-y.

[7] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36, http://dx.doi.org/10.1021/ci00057a005.

[8] K. Hatakeyama-Sato, N. Yamane, Y. Igarashi, Y. Nabae, T. Hayakawa, Prompt engineering of GPT-4 for chemical research: what can/cannot be done? Sci. Technol. Adv. Mater. Methods 3 (1) (2023) http://dx.doi.org/10.1080/27660400.2023.2260300.

[9] M.A. Bran, S. Cox, O. Schilter, C. Baldassari, A.D. White, P. Schwaller, Augmenting large language models with chemistry tools, Nat. Mach. Intell. 6 (5) (2024) 525–535, http://dx.doi.org/10.1038/s42256-024-00832-8.

[10] K.M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, Leveraging large language models for predictive chemistry, Nat. Mach. Intell. 6 (2) (2024) 161–169, http://dx.doi.org/10.1038/s42256-023-00788-1.

[11] D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, W. Ouyang, D. Zhou, S. Zhang, M. Su, H.-S. Zhong, Y. Li, ChemLLM: A chemical large language model, 2024, arXiv:2402.06852.

[12] J. Van Herck, M.V. Gil, K.M. Jablonka, A. Abrudan, A.S. Anker, M. Asgari, B. Blaiszik, A. Buffo, L. Choudhury, C. Corminboeuf, H. Daglar, A.M. Elahi, I.T. Foster, S. Garcia, M. Garvin, G. Godin, L.L. Good, J. Gu, N. Xiao Hu, X. Jin, T. Junkers, S. Keskin, T.P.J. Knowles, R. Laplaza, M. Lessona, S. Majumdar, H. Mashhadimoslem, R.D. McIntosh, S.M. Moosavi, B. Mouriño, F. Nerli, C. Pevida, N. Poudineh, M. Rajabi-Kochi, K.L. Saar, F. Hooriabad Saboor, M. Sagharichiha, K.J. Schmidt, J. Shi, E. Simone, D. Svatunek, M. Taddei, I. Tetko, D. Tolnai, S. Vahdatifar, J. Whitmer, D.C.F. Wieland, R. Willumeit-Römer, A. Züttel, B. Smit, Assessment of fine-tuned large language models for real-world chemistry and material science applications, Chem. Sci. 16 (2025) 670–684, http://dx.doi.org/10.1039/D4SC04401K.

[13] A. Dziubińska, A. Gontarz, M.a. Dziubiński, M. Barszcz, The forming of magnesium alloy forgings for aircraft and automotive applications, Adv. Sci. Technol. Res. J. 10 (31) (2016) 158–168, http://dx.doi.org/10.12913/22998624/64003.

[14] W.J. Joost, P.E. Krajewski, Towards magnesium alloys for high-volume automotive applications, Scr. Mater. 128 (2017) 107–112, http://dx.doi.org/10.1016/j.scriptamat.2016.07.035, URL https://www.sciencedirect.com/science/article/pii/S1359646216303621.

[15] A. Santos-Coquillat, M. Esteban-Lucia, E. Martinez-Campos, M. Mohedano, R. Arrabal, C. Blawert, M. Zheludkevich, E. Matykina, PEO coatings design for Mg-Ca alloy for cardiovascular stent and bone regeneration applications, Mater. Sci. Eng. C 105 (2019) 110026, http://dx.doi.org/10.1016/j.msec.2019.110026, URL https://www.sciencedirect.com/science/article/pii/S0928493119306320.

[16] F. Witte, N. Hort, C. Vogt, S. Cohen, K.U. Kainer, R. Willumeit, F. Feyerabend, Degradable biomaterials based on magnesium corrosion, Curr. Opin. Solid State Mater. Sci. 12 (5) (2008) 63–72, http://dx.doi.org/10.1016/j.cossms.2009.04.001, URL https://www.sciencedirect.com/science/article/pii/S1359028609000357.

[17] D. Höche, S.V. Lamaka, B. Vaghefinazari, T. Braun, R.P. Petrauskas, M. Fichtner, M.L. Zheludkevich, Performance boost for primary magnesium cells using iron complexing agents as electrolyte additives, Sci. Rep. (2018) http://dx.doi.org/10.1038/s41598-018-25789-8.

[18] M. Deng, L. Wang, B. Vaghefinazari, W. Xu, C. Feiler, S.V. Lamaka, D. Höche, M.L. Zheludkevich, D. Snihirova, High-energy and durable aqueous magnesium batteries: Recent advances and perspectives, Energy Storage Mater. 43 (2021) 238–247, http://dx.doi.org/10.1016/j.ensm.2021.09.008, URL https://www.sciencedirect.com/science/article/pii/S2405829721004268.

[19] J. Muldoon, C.B. Bucur, T. Gregory, Quest for nonaqueous multivalent secondary batteries: Magnesium and beyond, Chem. Rev. 114 (23) (2014) 11683–11720, http://dx.doi.org/10.1021/cr500049y, arXiv:https://doi.org/10.1021/cr500049y, PMID: 25343313.

[20] Z. Ma, D.R. MacFarlane, M. Kar, Mg cathode materials and electrolytes for rechargeable Mg batteries: A review, Batter. Supercaps 2 (2) (2019) 115–127, http://dx.doi.org/10.1002/batt.201800102, URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/batt.201800102, arXiv:https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/batt.201800102.

[21] J. Gray, B. Luan, Protective coatings on magnesium and its alloys — a critical review, J. Alloys Compd. 336 (1) (2002) 88–113, http://dx.doi.org/10.1016/S0925-8388(01)01899-0, URL https://www.sciencedirect.com/science/article/pii/S0925838801018990.

[22] C. Blawert, W. Dietzel, E. Ghali, G. Song, Anodizing treatments for magnesium alloys and their effect on corrosion resistance in various environments, Adv. Eng. Mater. 8 (6) (2006) 511–533, http://dx.doi.org/10.1002/adem.200500257, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adem.200500257, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/adem.200500257.

[23] S. Lamaka, B. Vaghefinazari, D. Mei, R. Petrauskas, D. Höche, M. Zheludkevich, Comprehensive screening of Mg corrosion inhibitors, Corros. Sci. 128 (2017) 224–240, http://dx.doi.org/10.1016/j.corsci.2017.07.011, URL https://www.sciencedirect.com/science/article/pii/S0010938X17303931.

[24] B. Vaghefinazari, E. Wierzbicka, P. Visser, R. Posner, R. Arrabal, E. Matykina, M. Mohedano, C. Blawert, M.L. Zheludkevich, S.V. Lamaka, Chromate-free corrosion protection strategies for magnesium alloys—A review: Part III—Corrosion inhibitors and combining them with other protection strategies, Materials 15 (23) (2022) http://dx.doi.org/10.3390/ma15238489, URL https://www.mdpi.com/1996-1944/15/23/8489.

[25] P. Kirkpatrick, C. Ellis, Chemical space, Nature (2004) http://dx.doi.org/10.1038/432823a.

[26] D.A. Winkler, A.E. Hughes, C. Özkan, A. Mol, T. Würger, C. Feiler, D. Zhang, S.V. Lamaka, Impact of inhibition mechanisms, automation, and computational models on the discovery of organic corrosion inhibitors, Prog. Mater. Sci. 149 (2025) 101392, http://dx.doi.org/10.1016/j.pmatsci.2024.101392, URL https://www.sciencedirect.com/science/article/pii/S0079642524001610.

[27] P. White, G. Smith, T. Harvey, P. Corrigan, M. Glenn, D. Lau, S. Hardin, J. Mardel, T. Markley, T. Muster, N. Sherman, S. Garcia, J. Mol, A. Hughes, A new high-throughput method for corrosion testing, Corros. Sci. 58 (2012) 327–331, http://dx.doi.org/10.1016/j.corsci.2012.01.016, URL https://www.sciencedirect.com/science/article/pii/S0010938X12000509.

[28] T. Muster, A. Hughes, S. Furman, T. Harvey, N. Sherman, S. Hardin, P. Corrigan, D. Lau, F. Scholes, P. White, M. Glenn, J. Mardel, S. Garcia, J. Mol, A rapid screening multi-electrode method for the evaluation of corrosion inhibitors, Electrochim. Acta 54 (12) (2009) 3402–3411, http://dx.doi.org/10.1016/j.electacta.2008.12.051, URL https://www.sciencedirect.com/science/article/pii/S0013468608014473.

[29] M. Meeusen, L. Zardet, A.M. Homborg, M. Lekka, F. Andreatta, L. Fedrizzi, B. Boelen, H. Terryn, J.M.C. Mol, A complementary electrochemical approach for time-resolved evaluation of corrosion inhibitor performance, J. Electrochem. Soc. 166 (11) (2019) C3220, http://dx.doi.org/10.1149/2.0271911jes.

[30] D.A. Winkler, Predicting the performance of organic corrosion inhibitors, Metals 7 (12) (2017) http://dx.doi.org/10.3390/met7120553, URL https://www.mdpi.com/2075-4701/7/12/553.

[31] M. Fernandez, M. Breedon, I.S. Cole, A.S. Barnard, Modeling corrosion inhibition efficacy of small organic molecules as non-toxic chromate alternatives using comparative molecular surface analysis (CoMSA), Chemosphere 160 (2016) 80–88, http://dx.doi.org/10.1016/j.chemosphere.2016.06.044, URL https://www.sciencedirect.com/science/article/pii/S0045653516308025.

[32] F.F. Chen, M. Breedon, P. White, C. Chu, D. Mallick, S. Thomas, E. Sapper, I. Cole, Correlation between molecular features and electrochemical properties using an artificial neural network, Mater. Des. 112 (2016) 410–418, http://dx.doi.org/10.1016/j.matdes.2016.09.084, URL https://www.sciencedirect.com/science/article/pii/S0264127516312655.

[33] D.A. Winkler, M. Breedon, A.E. Hughes, F.R. Burden, A.S. Barnard, T.G. Harvey, I. Cole, Towards chromate-free corrosion inhibitors: structure–property models for organic alternatives, Green Chem. 16 (2014) 3349–3357, http://dx.doi.org/10.1039/C3GC42540A.

[34] M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature (2018) http://dx.doi.org/10.1038/nature25978.

[35] L.B. Coelho, D. Zhang, Y.V. Ingelgem, D. Steckelmacher, A. Nowé, H. Terryn, Reviewing machine learning of corrosion prediction in a data-oriented perspective, Npj Mater. Degrad. (2022) http://dx.doi.org/10.1038/s41529-022-00218-4.

[36] T. Würger, L. Wang, D. Snihirova, M. Deng, S.V. Lamaka, D.A. Winkler, D. Höche, M.L. Zheludkevich, R.H. Meißner, C. Feiler, Data-driven selection of electrolyte additives for aqueous magnesium batteries, J. Mater. Chem. A 10 (2022) 21672–21682, http://dx.doi.org/10.1039/D2TA04538A.

[37] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 48 (9) (2008) 1733–1746, http://dx.doi.org/10.1021/ci800151m, PMID: 18729318.

[38] E.J. Schiessler, T. Würger, B. Vaghefinazari, S.V. Lamaka, R.H. Meißner, C.J. Cyron, Z.M. L., C. Feiler, R.C. Aydin, Searching the chemical space for effective magnesium dissolution modulators: a deep learning approach using sparse features, Npj Mater. Degrad. (2023) http://dx.doi.org/10.1038/s41529-023-00391-0.

[39] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing learned molecular representations for property prediction, J. Chem. Inf. Model. 59 (8) (2019) 3370–3388, http://dx.doi.org/10.1021/acs.jcim.9b00237, arXiv:https://doi.org/10.1021/acs.jcim.9b00237, PMID: 31361484.

[40] E. Heid, K.P. Greenman, Y. Chung, S.-C. Li, D.E. Graff, F.H. Vermeire, H. Wu, W.H. Green, C.J. McGill, Chemprop: A machine learning package for chemical property prediction, J. Chem. Inf. Model. 64 (1) (2024) 9–17, http://dx.doi.org/10.1021/acs.jcim.3c01250, arXiv:https://doi.org/10.1021/acs.jcim.3c01250, PMID: 38147829.

[41] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020, CoRR, arXiv:2010.09885.

[42] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa-2: Towards chemical foundation models, 2022, arXiv:2209.01712.

[43] T.L.P. Galvão, G. Novell-Leruth, A. Kuznetsova, J. Tedim, J.R.B. Gomes, Elucidating structure–property relationships in aluminum alloy corrosion inhibitors by machine learning, J. Phys. Chem. C 124 (10) (2020) 5624–5635, http://dx.doi.org/10.1021/acs.jpcc.9b09538, arXiv:https://doi.org/10.1021/acs.jpcc.9b09538.

[44] T. Würger, D. Mei, B. Vaghefinazari, D.A. Winkler, S.V. Lamaka, M.L. Zheludkevich, R.H. Meißner, C. Feiler, Exploring structure-property relationships in magnesium dissolution modulators, Npj Mater. Degrad. 5 (1) (2021) 2, http://dx.doi.org/10.1038/s41529-020-00148-z.

[45] E.J. Schiessler, T. Würger, S.V. Lamaka, R.H. Meißner, C.J. Cyron, Z.M. L., C. Feiler, R.C. Aydin, Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning models, Npj Comput. Mater. (2021) http://dx.doi.org/10.1038/s41524-021-00658-7.