Corrosion Science 273 (2026) 114276

Contents lists available at ScienceDirect

Corrosion Science

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

Different alloy, Different story? Impact of transfer learning on predictive models for in silico screening of Mg dissolution modulators

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

[a] Institute for Continuum and Material Mechanics, Hamburg University of Technology, Eißendorfer Straße, Hamburg, 21073, Germany
[b] Institute of Surface Science, Helmholtz-Zentrum Hereon, Max–Planck-Straße, Geesthacht, 21502, Germany
[c] Institute of Material Systems Modeling, Helmholtz-Zentrum Hereon, Max–Planck-Straße 1, Geesthacht, 21502, Germany
[d] German Center for Artificial Intelligence (DFKI), Germany
[e] Saarland University, Saarbrücken, Germany
[f] Institute for Interface Physics and Engineering, Hamburg University of Technology, Am Irrgarten 3-9, Hamburg, 21073, Germany

ARTICLE INFO

Dataset link: https://github.com/MatthiasHBusch/CrossAlloyTL

Keywords:
Magnesium corrosion
Inhibition efficiency
Transfer learning
Structure–activity relationships
Machine learning benchmarks

ABSTRACT

Ideally, datasets for training machine learning models consist of thousands of versatile data samples — a scale that is seldom met in materials research to date. Transfer learning (TL) is widely used to mitigate data scarcity in available training sets by leveraging knowledge from related domains, since models pretrained on auxiliary datasets may learn structure–property relationships that are transferable from one to another. In this work, the potential of TL for predicting the performance of small organic molecule dissolution modulators is probed using corrosion inhibition responses of eleven Mg-based materials taken from the previously published ExCorr database. A significant improvement in prediction accuracy is observed when transfer learning is applied. However, for the dataset configuration studied here, the apparent cross-alloy TL gains are shown not to arise from transferable structure–activity relationship (SAR) knowledge but rather from a learned alloy-context signal embedded in the data. Seven Mg-based materials—six commercial alloys plus the high-purity variant HPMg51—form a shared-SAR cluster with strongly correlated inhibitor rankings, such that molecules labelled on one material are readily predictable for others within the cluster. Within this shared-SAR cluster, extensive overlap reduces the effective training pool from approximately 900 records to only around 180 unique molecules, placing the task firmly in a small-data regime. In this setting, an in-context large language model (Gemini 3.1 Pro) is the only approach that successfully learns meaningful SAR, with its pretrained knowledge providing the transfer that the explicit source dataset fails to deliver.

1. Introduction

Magnesium(Mg)-based materials combine low density with high specific strength rendering them attractive for lightweight structural applications in automotive and aerospace structures [1,2] and, more recently, in biomedical engineering where it is used as base material for bioresorbable implants [3,4]. Beyond their use in lightweight structural applications, Mg-based materials are also gaining increasing attention in energy applications, particularly as anode materials in next-generation battery systems due to their high electrochemical activity [5–8]. However, this same electrochemical activity - beneficial or even functional in biomedical and energy-storage contexts — also drives the susceptibility of Mg to rapid corrosion in humid or saline environments, whereas structural applications would ideally require near-passive or corrosion-immune behaviour. It is obvious that one needs to gain control over the corrosion rate of the used Mg-based materials to unlock their full potential as lightweight engineering materials [9–11].

Organic corrosion inhibitors have emerged as a promising solution to gain control over the dissolution rate of Mg-based materials [12–14], but their chemical space is essentially infinite and cannot be explored based on experimental (high-throughput) techniques alone [14,15]. Naturally, predicting the corrosion inhibition performance of small organic molecules for metallic substrates based on structure–activity modelling has become a central challenge in computational corrosion science over the course of the last decades [14–16] and can be considered a digital shadow for state-of-the-art high-throughput testing approaches [17–24]. Predictive models for corrosion inhibitors are frequently described as quantitative structure–property relationships

* Corresponding authors.
E-mail addresses: matthias.busch@tuhh.de (M. Busch), roland.aydin@uni-saarland.de (R.C. Aydin), christian.feiler@hereon.de (C. Feiler).

https://doi.org/10.1016/j.corsci.2026.114276
Received 13 June 2026; Received in revised form 17 September 2026; Accepted 20 September 2026
Available online 26 September 2026
0010-938X/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Open access funding enabled and organized by Projekt DEAL.

(QSPR) in materials research. However, the endpoint of interest in corrosion inhibition is not a static molecular or materials property, but the functional ability of a compound to suppress an ongoing materials degradation process. Inhibition efficiency depends on dynamic interfacial phenomena such as adsorption, surface coverage, charge transfer, and modification of anodic and/or cathodic reaction rates. Moreover, small organic inhibitors may also act through a corrosion-promoter scavenging mechanism. By chelating Fe^{2+}/Fe^{3+} ions, which are known to accelerate Mg corrosion through cathodic activation and redox-mediated degradation pathways [25,26], the inhibitor decreases the availability of electrochemically active iron species that increases the corrosion activity of Mg-based materials by re-plating. Complexation of Fe^{2+}/Fe^{3+} ions thereby reduces catalytic cathodic activity and limits disruption of the passivating $Mg(OH)_2$-rich surface film. Thus, the corrosion-affecting compound does not only protect Mg by direct adsorption on the surface, but also by chemically deactivating aggressive trace-metal species in the interfacial electrolyte. Thus, the target quantity is more naturally interpreted as an activity, analogous to biological activity in drug-discovery quantitative structure–activity relationship (QSAR) studies where compounds are screened for their biological or pharmacological activity (e.g. enzyme inhibition, receptor binding, ion-channel modulation, cancer cell growth inhibition) [27–29]. Consequently, dissolution modulators and more specifically corrosion inhibitors may be considered as functional molecular agents for materials protection. Like bio-active molecules in drug discovery, they modify specific interactions of a substrate in a complex and dynamic environment. Therefore, corrosion-inhibition efficiency can be interpreted as an activity endpoint, supporting the use of QSAR-type terminology despite the fact that the term QSPR has historically been allocated in this research field.

Predictive modelling of small organic molecule dissolution modulators (compounds that either inhibit or accelerate the dissolution of Mg) is further complicated by the fact that the underlying training data sets are small [15,30,31] from a machine learning perspective and are usually confined to a small domain of applicability. As corrosion-inhibitor research has not yet reached the level of data availability typical of established big-data domains, transfer learning (TL) [32–34] may provide a particularly valuable modelling framework, enabling small experimental datasets to be leveraged more effectively by reusing information across (un)related molecules and other substrates.

Commercial magnesium alloys differ in composition and microstructure—from cast alloys to rare-earth-containing wrought alloys and high-purity Mg variants. Experimentalists routinely screen the same inhibitor libraries across alloy families [12,13] to understand the differences in corrosion mechanisms between the alloys. In turn, this provides an ideal basis for the evaluation of TL, which facilitates the development of one large prediction model across all alloys following the same principle: learning from (dis)similar QSAR across available datasets and transferring knowledge on how a set of inhibitors performs on one material to another. For magnesium specifically, however, the ExCorr screening data [13,14] have so far been used only to fit single-alloy inhibitor-efficiency models [17–19]; none treats the alloy itself as a variable, so cross-alloy transferability has not previously been quantified. A complementary and rapidly maturing route is to predict molecular properties with large language models [35–38], including a recent in-context study on Mg inhibitors [39], although their apparent performance raises a pretraining-corpus contamination concern that a blinded follow-up addresses directly [40].

The assumption that information can be transferred from one system to another is evaluated using the performance of a predictive machine-learning model as the probe in this work. Data samples of the three most heavily measured Mg-alloys in the ExCorr [13,14] electrochemical corrosion database (AZ31, AZ91, WE43) are used as the target for the transfer, and the impact of TL is assessed based on three elements. First, the prediction accuracy of different models is compared when the training set is restricted to the target alloy (Exact, 60 samples) versus when it is augmented with the remaining Mg-alloy data (Close transfer, ≈900 samples) or with all non-Mg corrosion data as well (Far transfer, ≈4000 samples) for several different molecular property regression methods, to probe whether cross-alloy TL improves the reliability of the prediction. Secondly, to distinguish the transfer of generalisable structure–activity information from a cross-alloy-but-same-molecule transfer, a molecule-disjoint variant of the standard protocol is introduced: for each test fold, every record whose canonical SMILES coincides with a test-fold molecule is removed from the source pool, to provide insight into what is actually being transferred. The two variants correspond to two distinct application scenarios. The first application case revolves around screening molecules that have not been measured on any material yet and the second focuses on predicting the performance of an already-measured molecule for another substrate. Moreover, a systematic Spearman rank-correlation analysis of inhibitor performance orderings across the eleven considered Mg-based materials is provided to interpret the TL and filtering results.

2. Data and methods

2.1. Dataset

A Mg-alloy subset of the ExCorr electrochemical corrosion inhibitor database [14] is used as a data source for this study. The full dataset comprises 5167 IE measurements on 1224 canonical molecules across 11 Mg-based materials as well as for Al-, Fe-, Cu- and Zn-based substrates. AZ31, AZ91 and WE43 were chosen as target alloys and constitute the testing set for this study because they are among the most heavily measured Mg materials (≥75 measurements) and jointly span the commercial Al–Zn and rare-earth-containing magnesium alloy families.

Target-alloy protocol. For each target alloy the largest experimental group with homogeneous conditions was identified, and its first 75 samples—which correspond to 75 distinct molecules, so that the folds are molecule-disjoint within the target alloy—were partitioned into 5 disjoint test folds of 15 molecules each. For every fold the remaining 60 records form the target training set. The largest homogeneous groups comprise 80 (AZ31), 76 (AZ91) and 79 (WE43) records, so the truncation to the first 75 discards at most 5 records (≤6%) per alloy. A selection or ordering bias from this step (e.g. reflecting literature source or screening batch) is therefore confined to a negligible fraction of the data. This assumption is confirmed empirically in SI Sec. S14.

Transfer pools. The close pool (839 records, 178 unique canonical molecules) consists of all Mg-alloy measurements whose alloy is not in the target set. The far pool (3955 records, 1097 unique canonical molecules) consists of all corrosion measurements on non-Mg base materials (Al, Fe, Cu, Zn).

Filtered variants. For each fold, filtered variants of both pools were generated by removing every record whose canonical SMILES coincides with a test-fold molecule. On average this removes 96 records per fold from the close pool (because all 15 test molecules recur in the close pool, each appearing on roughly 5–8 other Mg materials) and 37 records from the far pool (3–10 of the 15 test molecules overlap, each appearing on roughly 7 non-Mg substrates). Per-alloy and per-fold overlap statistics are given in the supplementary material (Tables S1 and S2).

Sample counts. Per-alloy record counts and unique-molecule counts are tabulated in SI Sec. S1 (Table S1). The close-pool members that dominate the close transfer pool are AM50 (122 records, 104 molecules), ZE41 (123 records, 105 molecules), E21 (117 records, 102 molecules) and the high-purity and commercial-purity Mg group CPMg220/CPMg342/HPMg50/HPMg51 (combined 416 records). Mg-0.8Ca contributes 56 records to the close pool. In addition, it is analysed separately as the outside-cluster contrast in Section 3.5, where it serves as the target alloy. Because Mg-0.8Ca shares only a handful of molecules

with the cluster alloys (Section 3.5), its 56 records have a negligible effect on the main target-alloy results. The interpretation of why Mg-0.8Ca behaves so differently—a combination of its distinct Ca-containing microstructure and the different experimental boundary conditions under which it was measured [13,41]—is deferred to Section 3.5.

2.2. Transfer learning problem definition and terminology

In the standard transfer learning framework [32,33], a domain comprises both the feature space and the associated data distribution. In the present study, domains are defined by the material system, i.e. the alloy identity including its impurity state, together with the experimental protocol under which the inhibitor data were generated. The shared task across all domains is the regression of inhibition efficiency (IE) from a molecule and its exposure context. AZ31, AZ91, and WE43 serve as the target domains. Transfer is achieved by augmenting the target-alloy training data with measurements from other Mg-based materials (close transfer) or from non-Mg substrates (far transfer). Domain identity is supplied to the models as an input feature (Section 2.3). In this work two transfer learning scenarios are considered:

- **Scenario A — new-molecule screening** of molecules never measured on any material: recurrence of a test molecule in the source pool constitutes label leakage. Emulated by the molecule-disjoint (filtered) settings.
- **Scenario B — cross-alloy calibration** of a molecule already measured on another alloy: the same recurrence is the defining feature of the task, not leakage. Emulated by the unfiltered settings.

Throughout this work, the term leakage refers exclusively to Scenario A. In Scenario B the corresponding recurrence is instead described as molecule overlap and the resulting gain as alloy-context transfer. Consequently, the performance gap between filtered and unfiltered settings (Section 2.5) quantifies the extent to which the apparent transfer learning gain originates from cross-alloy calibration of previously measured molecules.

2.3. Structural encoding and context

All feature-based models use 2048-bit Morgan fingerprints [42,43] (radius 2, chirality-aware) computed with RDKit [44], concatenated with three experimental context features: HCl%, NaCl%, and molar concentration in mM. Under the close-transfer settings the model also receives a one-hot alloy encoding; under the far-transfer settings it additionally receives a one-hot base-material encoding. ChemProp is trained on raw SMILES strings with the same context features supplied as "`x_d`" extra descriptors.

2.4. Models

In the following section the potential of TL is probed for six different model archetypes and their performance is assessed based on their prediction accuracy.

- ***k*NN-Tanimoto** ($k=5$, distance-weighted mean): a parameter-free similarity look-up that predicts target IE as the weighted average of the k most Tanimoto-similar training molecules.
- **Random Forest** (RF, 200 trees; default scikit-learn [45] hyperparameters).
- **Gradient Boosting** (GBR, 200 trees, depth 5, learning rate 0.05, subsample 0.8).
- **Multi-Layer Perceptron** (MLP, 256–128–64 hidden units, ReLU activations, early stopping on a 15% validation split).
- **ChemProp D-MPNN** [46,47] (hidden dim 300, 5 FFN layers, 5 training epochs, batch size 64, Adam optimiser with the ChemProp v2 default Noam-type schedule: 2 warm-up epochs, initial/maximum/final learning rate $10^{-4}/10^{-3}/10^{-4}$).
- **Gemini 3.1 Pro** (`google/gemini-3.1-pro-preview` via the OpenRouter API, reasoning effort `minimal`). The model is given a chemist-role system prompt and an in-context "training set" consisting of the per-fold training records (IUPAC names, SMILES and labelled IE values) plus the experimental conditions, and is asked to predict the IE of one held-out molecule per call. The same per-fold molecule-disjoint filter is applied to its in-context examples in the filtered settings. Each prediction is averaged over 3 independent calls (against the 15–25 seeds of the other models, owing to API cost). The in-context protocol is adopted from an earlier study on LLMs for Mg-inhibitor prediction [39]. Because the model is accessed through a closed commercial API whose served weights may change over time, the exact model identifier, provider, access dates, decoding parameters, prompt templates and answer-parsing rules are documented in SI Sec. S6. Additionally, the raw request/response transcripts of every call are archived in the accompanying repository, ensuring that the reported results remain auditable even if the endpoint evolves.

Every (alloy, fold, setting, model) combination is averaged over 25 random seeds for RF/GBR/MLP/*k*NN-Tan, 15 seeds for ChemProp, and 3 independent API calls for Gemini 3.1 Pro, totalling $3\times5\times5\times(4\times25+15)=8625$ individual model fits and $3\times5\times5\times15\times3=3375$ Gemini calls for the main experiment. The Mg-0.8Ca contrast adds $4\times5\times(4\times25+15)=2300$ further fits. The primary metric reported is the coefficient of determination $R^2=1-\mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$ on the held-out test fold. Pearson correlation r—shown alongside R^2 in Fig. 1—and mean absolute error (MAE), with full per-model r and MAE breakdowns are reported in the SI (Fig. S5). R^2 is preferred for the main analysis because, unlike r, it penalises systematic bias in absolute IE—the precise failure mode the paper uses as a basis to assess the potential of the transfer. All quantities are aggregated as mean ± standard error across the 15 (alloy, fold) combinations per setting. Because each test fold contains only 15 molecules, the per-fold aggregation is complemented by pooled molecule-level statistics (Table S14); all central claims rest on setting-to-setting differences that are consistent across all 15 combinations, not on the absolute R^2 of any single fold.

2.5. Statistical comparison

The overlap delta Δ_{R^2} of a model is defined as $\Delta_{R^2}=\bar{R}^2_{\mathrm{unfilt}}-\bar{R}^2_{\mathrm{filt}}$ at the close-transfer setting (i.e. Close-unfiltered minus Close-filtered). In terms of Section 2.2, Δ_{R^2} measures how much of the apparent TL benefit is specific to Scenario B (cross-alloy calibration of already-measured molecules); from the perspective of Scenario A (new-molecule screening) the same quantity is a leakage delta. A large Δ_{R^2} indicates that removing overlapping molecules eliminates most of the apparent TL benefit. The significance of the exact-vs-close-filtered comparison with a paired Wilcoxon signed-rank test over the 15 folds is provided to compare different levels of transfer.

3. Results and discussion

3.1. Cross-alloy TL improves target-alloy prediction

Figs. 1 and 2 summarise the cross-alloy transfer learning (TL) experiments on ExCorr. Fig. 1 reports the test-set R^2 (left) and Pearson r (right) of six model families across the five transfer settings, averaged over the 15 (target alloy, fold) combinations. Fig. 2 pairs each unfiltered TL scope with its molecule-disjoint counterpart, in which the test molecules of the target alloy are removed from the source pool before training; the vertical annotations give the overlap delta Δ_{R^2} (Section 2.5).

Adding the close-Mg transfer pool unfiltered raises the mean R^2 (averaged over the five fitted models) from −0.13 (Exact) to +0.39 (Close-unfiltered)—a $\Delta R^2=+0.52$ shift that takes the predictors from

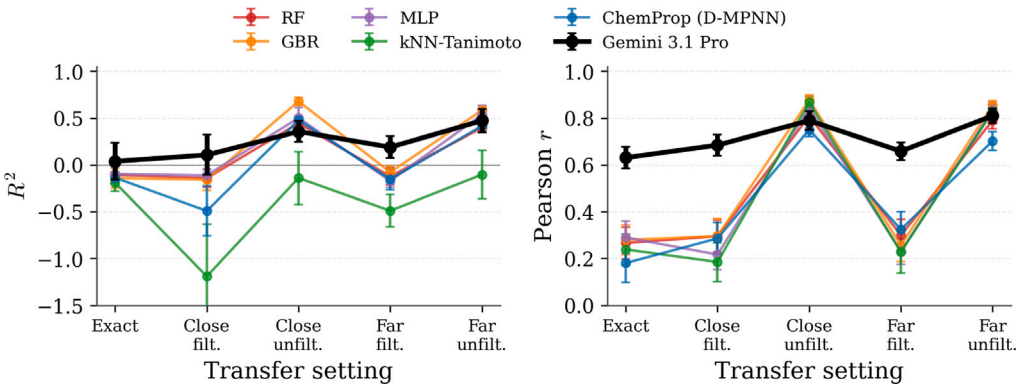


Fig. 1. Test-set R^2 (left) and Pearson r (right) across the five transfer settings. Mean ± standard error over the 15 (target alloy, fold) combinations. Models: RF (red), GBR (orange), MLP (purple), *k*NN-Tanimoto (green), ChemProp D-MPNN (blue), and Gemini 3.1 Pro in-context (black). R^2 (left panel) penalises miscalibration of the absolute IE scale, whereas Pearson r (right panel) reflects ranking only. The four calibrated regressors (RF, GBR, MLP, ChemProp) reach positive R^2 only in the unfiltered transfer settings and fall back to negative values when filtered; the parameter-free *k*NN-Tanimoto stays negative in R^2 throughout despite reaching a high Pearson r in the unfiltered settings—it ranks molecules well but does not calibrate absolute IE. Gemini 3.1 Pro stays positive in R^2 in every setting and attains the highest Pearson r in the filtered and Exact settings. The MAE companion to this figure is given in Fig. S5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

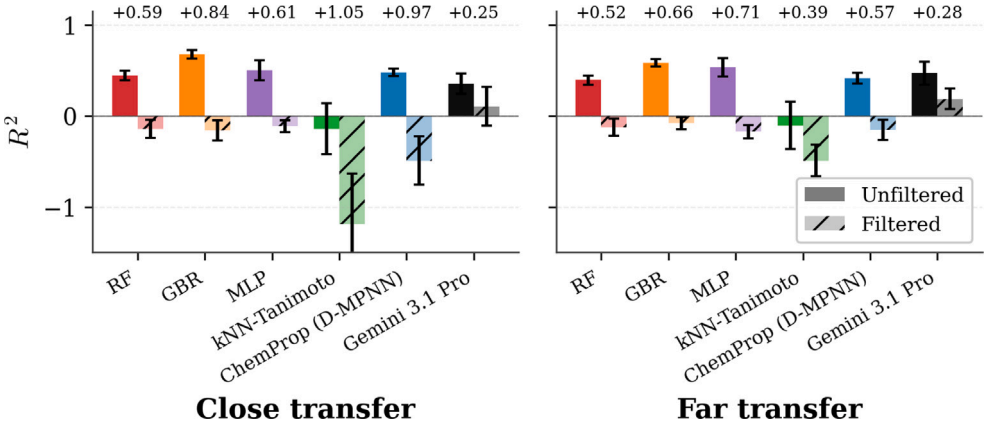


Fig. 2. Filtered-versus-unfiltered TL. Left: close-pool comparison; right: far-pool comparison. Solid bars show unfiltered TL (with molecule overlap between test set and source pool—Scenario B of Section 2.2); hatched bars show filtered TL (test molecules removed from source pool—Scenario A of Section 2.2); vertical annotations give the overlap delta Δ_{R^2}. Every fitted model family loses $\Delta_{R^2} = +0.59$ to $+1.05$ of variance explained on filtering. The parameter-free *k*NN-Tanimoto shows the largest Δ_{R^2}, as expected for a method that can only exploit direct molecular re-use.

worse-than-mean-IE to recovering roughly 40% of test-set IE variance, with the strongest model (GBR) reaching $R^2 = +0.68$. Adding the far pool provides essentially no further gain (mean $R^2 = +0.37$), consistent with diminishing returns from increasingly dissimilar source data. The gain is broadly architecture-agnostic among the four calibrated models. In the close-unfiltered setting, RF, GBR, MLP and ChemProp span the narrow band $R^2 = +0.44$–$+0.68$. The parameter-free *k*NN-Tanimoto look-up is the lone outlier ($R^2 = -0.14$) which is discussed in more detail below.

Filtering reverses these gains almost entirely. Removing the target's test molecules from the close pool—on average 96 records per fold, corresponding to all 15 test molecules recurring in the close pool—collapses the mean R^2 from +0.39 to −0.42. The overlap delta Δ_{R^2} averages +0.81 across the five fitted models, ranging from +0.59 for RF to +1.05 for *k*NN-Tanimoto. The same pattern holds for the far-pool comparison ($R^2 = +0.37 \rightarrow -0.20$, $\Delta_{R^2} \approx +0.57$), and filtered close-pool ranking performance is statistically indistinguishable from the Exact baseline (paired Wilcoxon on Pearson r; Table S12). Gemini 3.1 Pro marks an exception here and retains positive R^2-values in all settings, outperforming the other approaches on the exact and filtered settings significantly.

3.2. Decomposition of the TL gain

The five settings of Fig. 1 together decompose the apparent TL gain into two chemically interpretable components.

- **Exact** measures the small-data baseline—what each method extracts from ~60 target-alloy training samples without auxiliary information.
- **Exact → filtered transfer** (first step of Fig. 1) measures the inhibitor-SAR contribution: what the model gains by seeing labels for additional, non-overlapping molecules from neighbouring alloys. In this case this step is essentially zero.
- **Filtered → unfiltered transfer** (second step at each scope and detailed in Fig. 2) measures the alloy-context contribution: what the model gains by seeing labels for the same molecules on a different alloy. This step is the main contribution of the TL gain with $\Delta_{R^2} \approx +0.8$.

Two factors explain why the inhibitor-SAR step is flat. First, the close pool contributes few new molecules to the target's chemical space: the inhibitor libraries used across Mg-based materials in ExCorr overlap heavily, so the bulk of the 839 close-pool records are repeat measurements of molecules already present on the target alloy; the 839

close-pool records contain 178 unique molecules. With cluster-alloy rankings at Spearman r = 0.67–0.92 (Fig. 3), a labelled measurement on one cluster member carries essentially the same information as the same molecule measured on another cluster alloy—at least that is the information that the models extract. Second, the inhibitor SAR on Mg alloys is genuinely complex [17–19,39]: labels from the few new molecules in the close pool are too few to pin down predictive performance on a SAR that does not reduce to a small number of easily extrapolable descriptors. The flat SAR step is therefore partly a property of ExCorr specifically whereas a source pool with broader chemical diversity could in principle reveal genuine SAR transfer that ExCorr cannot expose.

The alloy-context step is easy for the mirror reasons, though not as trivial as "any regressor can do it". The alloy-to-alloy effect on the IE of an already-known molecule is, at leading order, a monotone re-scaling captured by the pairwise Spearman structure of Fig. 3—a one-dimensional transformation that the four calibrated regressors (RF, GBR, MLP, ChemProp) all fit to a similar quality (R^2 = +0.44 to +0.68). The parameter-free kNN-Tanimoto look-up, by contrast, ranks well—reaching Pearson r = +0.87 in Close-unfiltered, indistinguishable from the calibrated models' band r = +0.75–+0.88 (Tables S4, S14)—but does not estimate the rescaling and ends up with negative R^2 in every TL setting (Fig. 1, Table S3): cross-alloy rescaling needs a parametric fit. The same molecular overlap that handicaps SAR transfer becomes an advantage here, providing many paired (same-molecule, different-alloy) observations from which the rescaling is estimable with little data. This is further supported by the observation that the ChemProp D-MPNN does not stand out among the calibrated regressors, but the reason is probably that the ExCorr training pool is simply too small for a graph neural network to extract additional structure beyond what the tabular regressors capture. The alloys are, after all, qualitatively similar rather than identical.

In other words, the large Δ_{R^2} of this step measures the gain from applying a rescale mapping to the same molecule under a different label, not the discovery of any new chemistry. This asymmetry—not the raw size of the source pool—is what this cross-alloy TL benchmark on ExCorr implicitly measures. Reporting molecule-disjoint metrics alongside the standard ones, as done in this work, makes this distinction visible and cleanly separates genuinely transferable chemistry from re-use of labelled reference compounds—a separation worth keeping in mind for any small-data cross-alloy or cross-substrate transfer learning evaluation. Gemini 3.1 Pro is the sole exception to this pattern: it retains positive R^2 in every setting, including the molecule-disjoint filtered ones (Fig. 1). The result is consistent with pretraining supplying transferable SAR knowledge beyond that available in the ExCorr training data. This observation is examined in greater detail in Section 3.6.

3.3. The SAR-similarity cluster explains both the gain and its filter-sensitivity

Fig. 3 shows the pairwise Spearman rank correlation of inhibitor rankings between the 11 Mg systems. Seven Mg materials—the commercial alloys AM50, AZ31, AZ91, E21, WE43 and ZE41 together with the high-purity variant HPMg51—form a tight cluster with pairwise r = 0.67–0.92 (each pair computed over 73–100 shared molecules; 95% bootstrap CI lower bound ≥0.52 for every pair, and the cluster is preserved when the top-decile-IE anchor molecules are excluded (see SI Sec. S4)). Within this cluster, a molecule's effectiveness ranking is largely conserved from one material to another. The three target alloys analysed in this work (outlined in black) all sit firmly inside this cluster.

This ranking conservation grounds both main results of Sections 3.1–3.2. Because the materials in the cluster share a common inhibitor ordering, a model trained on any one of them can predict the others almost by look-up: the labelled molecules recur across alloys, and the IE ranking they induce transfers nearly verbatim. Conversely, removing the recurring molecules from the source pool erases the apparent

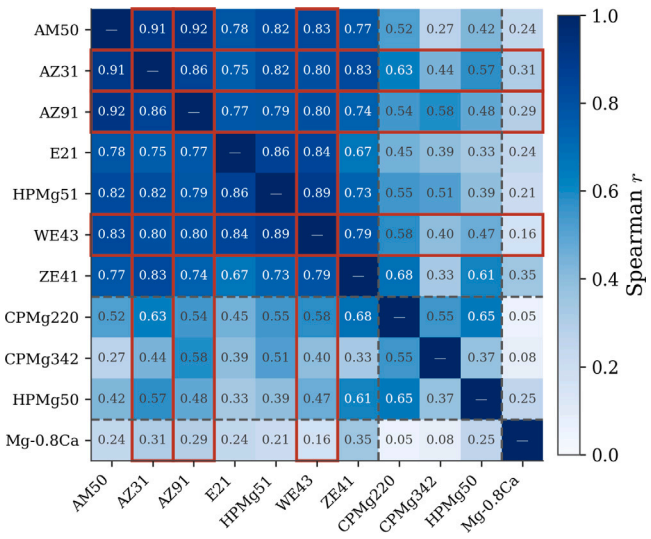


Fig. 3. SAR-similarity structure across Mg materials. Pairwise Spearman rank correlation of inhibitor rankings across the 11 Mg-based materials in the ExCorr database. The highest correlation ($r > 0.85$) was observed for the Al-containing alloys in the investigated dataset (AM50, AZ31, AZ91) whereas the three target alloys chosen for this study (AZ31, AZ91, WE43; outlined) are located within a tightly correlated cluster of seven Mg materials (six commercial alloys plus high-purity HPMg51; upper-left block) with pairwise rankings at r = 0.67–0.92. The high-Fe impurity variants (middle block) form a looser second cluster, and Mg-0.8Ca (bottom row) is the sole alloy outside any cluster, correlating with the others at $r \leq 0.35$. However, the pronounced drop in r for Mg-0.8Ca has to be interpreted with caution, as this dataset originates from a separate testing campaign with different experimental conditions while the other ten materials were investigated under the same boundary conditions.

benefit—what remains are unseen molecules whose IE values the model must predict without direct labelled neighbours. The cluster's existence is therefore responsible for both the large cross-alloy TL gain and its filter-sensitivity. We emphasise that the cluster structure describes correlated inhibitor rankings, not identical alloy corrosion behaviour in absolute terms: two cluster members agree on which molecules inhibit best, but not necessarily on the absolute IE values.

From a modelling perspective this means that any calibrated regressor—from a random forest to a 300-dimensional graph network—can exploit the shared ranking under unfiltered TL with similar variance-explained quality, while a parameter-free nearest-neighbour lookup ranks molecules well but cannot calibrate absolute IE. The task reduces to the near-monotone cross-alloy re-scaling of a neighbouring alloy's label for the same molecule. It also means that extrapolation beyond the molecules present in the joint Mg-alloy dataset is not genuinely supported by cross-alloy TL. A practical consequence is that, once a candidate inhibitor has been measured on one cluster member, repeating the measurement on a second cluster member adds little new information if performed at identical experimental conditions. Hence, screening budgets are better spent on measuring new molecules not yet present in the joint dataset, on alloys outside the cluster where the SAR diverges and independent measurements carry non-redundant information, or on re-measuring molecules under different conditions (e.g. composition and concentration of the background electrolyte, inhibitor concentration, electrolyte-volume-to-sample-surface-area ratio, initial pH, and alloy processing route and microstructure). From a materials-engineering perspective, the correlation matrix of Fig. 3 may serve as a practical decision tool. If the inhibitor rankings of a newly developed Mg alloy correlate strongly with those of the main cluster it is likely that the existing joint dataset may serve as basis for inhibitor pre-selection. Conversely, weak correlations would indicate an out-of-distribution alloy for which an independent screening campaign may be required.

3.4. Spearman correlation coefficients in context with alloy properties

The correlation structure of Fig. 3 may reflect mechanistic differences between the different Mg-based materials. It should be emphasised that the subsequently discussed reasons for the uncovered correlations (or lack thereof) are hypotheses and that their assessment is based on the statistical evidence discussed in this work. Nevertheless, their interpretation is consistent with the observed ranking correlations, with the compositional and impurity metadata of Table S21, and with the established electrochemistry of Mg corrosion [5,9,10,25,26]. However, a deep electrochemical and meta data-resolved analysis is required to further substantiate the uncovered material-dependent behaviour and provide evidence for the individual mechanisms proposed below. AM50, AZ31 and AZ91 (Al-containing commercial Mg-alloys) exhibit very similar inhibitor ranking behaviour, reflected by the consistently high Spearman correlations among these alloys. This similarity is consistent with their shared alloying elements, microstructural characteristics and corrosion-control mechanisms. In Al-containing Mg alloys, iron impurities are largely incorporated into Al–Fe–Mn-type intermetallic phases, which are significantly less cathodically active than elemental Fe [9,10]. As a result, corrosion is not governed by noble Fe-driven micro-galvanic coupling but instead by processes such as surface-film formation, adsorption of organic species, and precipitation of Mg-containing corrosion products, as well as comparably weak galvanic coupling between Al-containing secondary phases and the Mg matrix. Consequently, small organic molecules tend to influence corrosion through generic surface interactions rather than through strong iron-complexation effects, leading to a stable and highly reproducible inhibitor ranking across this alloy group. WE43, ZE41 and E21 (rare-earth-containing Mg-alloys) cluster closely with the Al-containing alloys, albeit with slightly reduced correlation strength. In these materials, rare-earth elements and Zr modify impurity distribution in a manner analogous to aluminium. Iron is scavenged into RE- or Zr-containing phases, reducing the cathodic efficiency of Fe-bearing particles and suppressing Fe redeposition during corrosion [5,13]. As a consequence, the inhibitor-ranking behaviour is less governed by suppression of Fe-driven cathodic activity and more by how the organic molecules affect Mg matrix dissolution, interfacial adsorption, and the formation or destabilisation of Mg-containing corrosion products and electrolyte-derived precipitates. The hypothesis that organic molecules may act through mechanisms similar to those operating in Al-containing alloys is consistent with the observation that their relative performance remains highly conserved within this group, which would in turn explain the strong correlations observed. Although HPMg51 and HPMg50 have comparable bulk impurity levels, they occupy distinctly different positions in the correlation matrix, consistent with a critical role of impurity elemental composition and distribution within the Mg matrix, rather than total impurity content. HPMg51 correlates reasonably well with the main alloy cluster which is consistent with its Fe-containing impurity particles being chemically modified by significant incorporation of Al and Zr. These elements reduce the cathodic activity of Fe, suppress Fe^{2+}/Fe^{3+} re-deposition, and thereby limit impurity-driven galvanic corrosion [5]. As a result, corrosion of HPMg51 is governed largely by matrix- and surface-film-controlled processes, and small organic molecules rank similarly to how they do on AZ, AM, and RE-containing alloys. In contrast, the weak correlation of HPMg50 with the alloy group is consistent with its Fe impurities remaining electrochemically active. In this material, Fe-rich particles often contain insufficient Al or Zr to suppress their cathodic behaviour, leading to sustained micro-galvanic corrosion and dynamic Fe re-deposition. Under these conditions, corrosion becomes strongly controlled by Fe redox chemistry, and organic molecules that can suppress Fe activity or complex dissolved iron gain disproportionate importance. Such a shift in the hierarchy of inhibitor effectiveness away from the alloy-like pattern would explain the low correlations of HPMg50 with the other alloys despite similar nominal purity. HPMg50 can therefore be regarded as a transitional case between alloy-like Mg materials and commercial-purity Mg: although its nominal impurity level is low, the electrochemical activity of its Fe-containing particles causes inhibitor rankings to shift towards the Fe-controlled regime observed for CP Mg. Commercial purity Mg (CPMg220 and CPMg342) forms a clearly distinct group, positioned far from the alloy cluster in the correlation matrix. In these materials, corrosion is dominated by abundant Fe-rich impurity particles that act as highly efficient cathodes [5,11]. Continuous Fe re-deposition amplifies cathodic area and accelerates corrosion, making the system particularly sensitive to the iron-binding capability of organic molecules [25,26]. Consequently, inhibitors that are highly effective on CP Mg are often those that suppress iron re-deposition, while adsorption-type inhibitors that perform well on alloys may be far less effective. Such an impurity-controlled corrosion regime reshuffles inhibitor rankings relative to those observed for alloys and offers a consistent explanation for the limited correlation between CP Mg and alloyed systems.

3.5. Mg-0.8Ca as an out-of-distribution contrast

Mg-0.8Ca is the only alloy in ExCorr that lies outside the SAR-similarity cluster, with Spearman rank correlations of $r \leq 0.35$ against every other Mg material (see Fig. 3). Mg-0.8Ca's molecular overlap with the other alloys is also small (11–45 shared molecules per pair, versus 73–100 within the commercial cluster). Hence, individual Mg-0.8Ca pairwise correlations carry substantial sampling uncertainty (SI Sec. S4) and are only utilised to support the qualitative claim that Mg-0.8Ca lies uniformly outside the cluster. Consistency with this picture would require that cross-alloy TL cannot recover accurate predictions for Mg-0.8Ca under any setting.

This assumption is tested by rerunning the full experimental pipeline with Mg-0.8Ca as the target alloy. Because Mg-0.8Ca has only 56 records with 49 unique molecules, a 4-fold split is used rather than the 75/15 5-fold protocol of the main experiment. Fig. 4 shows the Mg-0.8Ca result alongside the main target-alloy result for direct comparison.

The Mg-0.8Ca Exact baseline is already mediocre—the best fitted model (ChemProp) reaches only $R^2 = -0.39$, and the mean across the four calibrated fitted models is $R^2 \approx -0.96$. Switching on cross-alloy transfer further degrades the model. In the Close-unfiltered setting the best Mg-0.8Ca model still has $R^2 = -3.46$ (ChemProp), the mean drops to $R^2 \approx -7.3$, and MLP and kNN-Tanimoto fall to $R^2 = -12$ and $R^2 < -50$ respectively. The same pattern holds in every transfer setting. The Exact-to-Close-unfiltered shift is therefore $\Delta_{R^2} \approx -3$ to -50 on Mg-0.8Ca—a degradation—versus $\Delta_{R^2} = +0.52$ for the cluster-resident target alloys (an improvement). The contrast is consistent with the main result: cross-alloy TL appears contingent on IE-ranking similarity to the target, and outside the cluster the source pool actively pulls predictions away from the target-alloy IE scale.

Mg–Ca alloys represent a fundamentally different corrosion class and therefore appear as extreme outliers in the correlation matrix. In Mg-0.8Ca, corrosion is governed primarily by the anodically active Mg_2Ca phase rather than by cathodic impurity particles [4]. Iron plays only a secondary role, while Ca^{2+} chemistry dominates the formation and stability of protective surface layers. Small organic molecules influence corrosion mainly by interacting with Ca^{2+} as well as Mg^{2+}, for example by complexation or precipitation, which can accelerate corrosion instead of inhibiting it. Because the controlling species and reaction pathways differ from those in Fe-controlled or matrix-controlled systems, the relative ranking of organic molecules in Mg-0.8Ca becomes effectively decoupled from that observed for all other materials, which might be one of the major reasons to result in a correlation that is essentially zero. However, the weak correlation of Mg-0.8Ca with the remaining Mg materials cannot be attributed unambiguously to alloy chemistry alone. The Mg-0.8Ca data was curated from a separate work [41], whereas most other Mg-alloy data were taken from a more

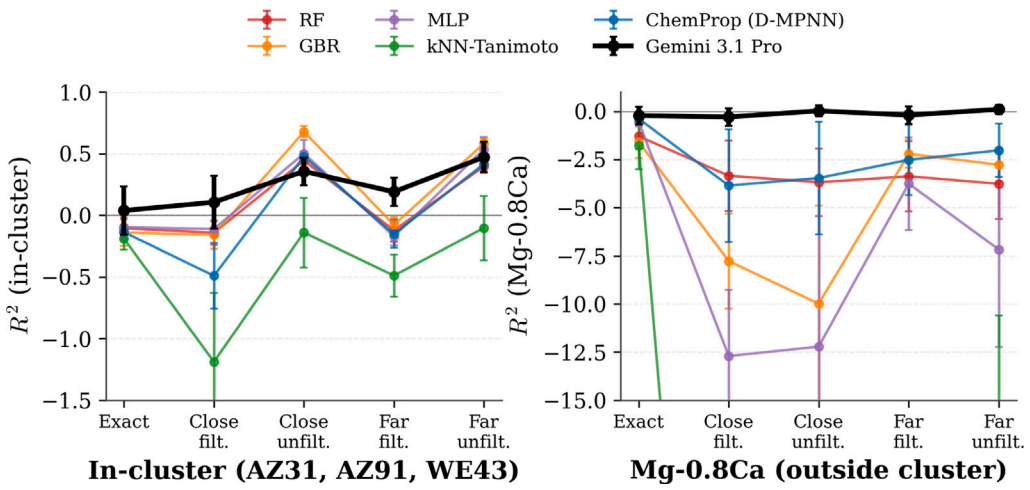


Fig. 4. Mg-0.8Ca contrast experiment. Both panels show R^2 but on different y-axis ranges (note the scale change between the panels). Left panel: main target-alloy mean (AZ31, AZ91, WE43), shares the y-range of Fig. 1. Right panel: Mg-0.8Ca—the one magnesium alloy outside the SAR cluster (see Fig. 3). For the in-cluster targets, the four calibrated fitted models reproduce the canonical TL pattern: Exact $R^2 \approx -0.13$ rises to Close-unfiltered $R^2 \approx +0.39$ and collapses again under filtering. For Mg-0.8Ca, Exact is only mildly negative ($R^2 \approx -0.4$ to -1.8) but cross-alloy transfer makes predictions worse, not better: Close- and Far-transfer R^2 fall to -2 to -12 for the four calibrated fitted models and to $R^2 < -50$ for kNN-Tanimoto (off-scale below the Mg-0.8Ca panel). For this out-of-distribution target, cross-alloy TL leads to a substantial deterioration of the predictive performance. Gemini 3.1 Pro (black) is the exception: on Mg-0.8Ca it stays near $R^2 \approx 0$ (-0.29 to $+0.11$) in every setting, never following the fitted models into strongly negative territory.

comprehensive screening campaign [13]. These two testing campaigns differ in several experimentally relevant parameters, including inhibitor concentration, NaCl concentration and electrolyte-volume-to-surface-area ratio. In particular, the inhibitors for the investigation of Mg-0.8Ca were tested at substantially lower and, in part, compound-specific concentrations whereas the majority of the modulators in the larger campaign was exposed to a fixed-concentration during experimental screening. In addition, the electrolyte-volume-to-surface-area ratio was approximately 10 for Mg-0.8Ca. This ratio was significantly smaller (1–2.8) for HPMg50 and the Al- and RE-containing alloy chips, whereas CPMg220 was tested at a ratio which was approximately one order of magnitude higher. These differences can influence both the apparent inhibition efficiencies and concomitantly their rank ordering. Indeed, the existing data directly demonstrate that the operating concentration alone can strongly affect the inhibitor performance: same-alloy, same-electrolyte rankings are conserved under moderate concentration changes (Spearman $r = 0.90$–0.93 for 20–32 mM vs. 50 mM) but collapse to $r \approx 0$ at 0.5–2 mM—the dilution factor separating the Mg-0.8Ca campaign from the main one (SI Sec. S11, Fig. S7). A concentration gap of this size is therefore by itself capable of destroying ranking similarity. The low Mg-0.8Ca correlations must consequently be read as the combined outcome of its distinct Ca-containing microstructure and the different boundary conditions. Disentangling the two would require re-measuring a subset of inhibitors under the main campaign's conditions—the single most valuable follow-up to this study. Mg-0.8Ca is thus an out-of-distribution target in the sense of Section 2.2, differing in both material system and protocol, rather than a general representative of Mg–Ca alloys. Importantly, the conclusion is unaffected by which factor dominates: In the present data, cross-alloy TL fails whenever differences in material chemistry or experimental protocol decorrelate the target's Spearman-ranked inhibitor performance from the source pool. Tracking the Spearman correlation coefficients in a multi-material dataset provides a robust starting point for a first educated guess on the probability for an effective transfer of domain knowledge from one material to others. The resulting matrix can be exploited as a map of dominant corrosion-controlling mechanisms rather than merely a statistical comparison of inhibitor performance. Like molecules sharing similar structural features, different metallic alloys will cluster when they share similar corrosion-controlling features. For Mg-based materials, one dominating aspect is the suppression of Fe activity after dissolution (e.g. by scavenging it from the electrolyte to prevent replating) and its distribution in the matrix. The decrease from alloyed Mg to high-purity Mg, then to commercial-purity Mg, and finally to Mg–Ca alloys reflects a systematic shift from matrix-controlled to Fe-controlled and ultimately to anodic-phase-controlled corrosion. The contrasting behaviour of HPMg51 and HPMg50 highlights that impurity distribution, not total impurity level, governs whether a material behaves in an alloy-like or impurity-dominated manner.

3.6. Gemini 3.1 Pro: in-context LLM prediction and a zero-shot contamination check

The Gemini 3.1 Pro in-context baseline reproduces the main SAR-cluster pattern of Section 3.1 qualitatively—higher R^2 in the unfiltered TL settings than in the filtered/Exact ones, and a clear filter-sensitivity ($\bar{R}^2_{\text{close,unfilt}} = +0.36$ vs. $\bar{R}^2_{\text{close,filt}} = +0.11$)—but the absolute level of its filtered/Exact performance is higher than for any of the five fitted models. Where RF, GBR, MLP, kNN-Tan and ChemProp collapse to negative R^2 once test molecules are removed from the training pool or omitted entirely (down to $R^2 = -1.2$ for kNN-Tanimoto), Gemini still reaches close-filtered $R^2 = +0.11$, far-filtered $R^2 = +0.19$ and Exact $R^2 = +0.04$ averaged across the three target alloys (Fig. 5; per-alloy breakdown in Fig. S1).

Such a gap raises an obvious concern: that Gemini has already encountered the ExCorr (molecule, IE) pairs during pretraining and is essentially recalling memorised labels rather than performing in-context structure–activity reasoning. This hypothesis is tested with a zero-shot variant in which the model receives the same chemist-role system prompt and the same experimental conditions for the held-out molecule, but no in-context training examples whatsoever (no Exact pool, no transfer pool). Each of the 75 molecules per target alloy is queried 3 times.

The result (Fig. 5, hatched bars, Table S13) is clear: zero-shot R^2 is negative on all three alloys ($R^2 \in [-0.66, -0.17]$), meaning the zero-shot predictor is worse than predicting the mean IE; Pearson r is +0.16 for AZ31, +0.22 for AZ91 and +0.15 for WE43; and MAE is 100%–147% IE, roughly 1.5× the in-context filtered-setting MAE. Were Gemini relying on memorised ExCorr labels, the zero-shot probe would inherit them and reach performance comparable to the filtered in-context setting. The in-context performance therefore reflects genuine use of

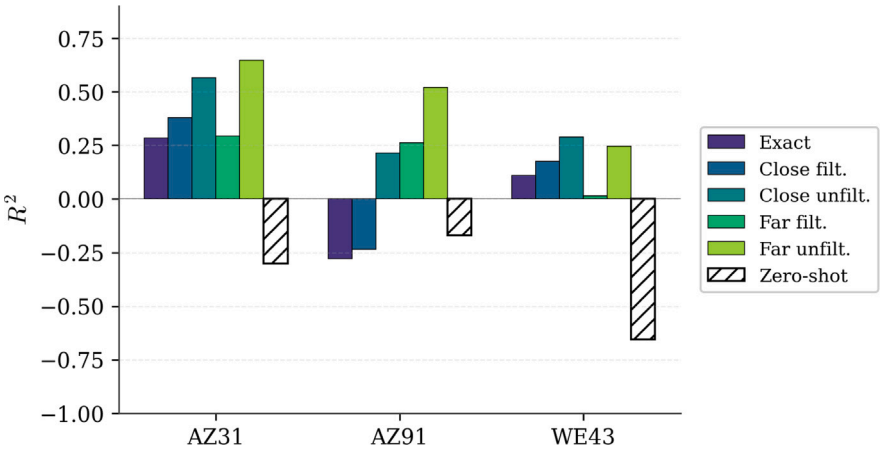


Fig. 5. Gemini 3.1 Pro: in-context settings vs. zero-shot. For each target alloy, coloured bars show the per-setting R^2 of the in-context Gemini predictions and the hatched white bars show the zero-shot baseline obtained without any in-context training records. Zero-shot R^2 is negative on every alloy ($R^2 \in [-0.66, -0.17]$, Table S13), with Pearson r nevertheless reaching +0.15–+0.22—a small ranking signal that is statistically significant when pooled across the three alloys (see SI). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the supplied training records, not pretraining-corpus contamination. Our companion study [40] provides a blinding methodology to distinguish memorisation-driven from knowledge-driven correlation. With the zero-shot correlation already at the edge of insignificance, there is no prior signal here whose origin would need to be disentangled, so that methodology is not required.

Three further controls bracket the mechanism (details in SI Secs. S13 and S15). Randomly permuting the IE labels among the in-context examples [48] collapses the Exact performance from $R^2 = +0.13$, $r = +0.67$ to $R^2 = -0.90$, $r = +0.10$—the zero-shot level—so the molecule–label association in the prompt, not its format or a recalled prior, drives the accuracy (Table S17). Re-serving the same examples in three random orders changes R^2 by at most 0.06: the predictions are robust to prompt order. Conversely, enriching the shallow models with ∼210 RDKit physicochemical descriptors does not close the gap (best shallow $r = 0.37$ vs. 0.69 in-context; Table S19). The LLM's small-data advantage is thus neither input richness nor label recall—leaving pretrained chemical knowledge exploited through genuine in-context learning as the operative mechanism. The consistently positive (but low) zero-shot Pearson r indicates a basic qualitative grasp of the Mg-inhibitor SAR that predates the in-context examples—a useful prior in ranking terms, but with negative R^2 it cannot account for the filtered-setting variance-explained gap.

The mechanism behind Gemini's dominance in the filtered and Exact settings is then straightforward: once molecule-level leakage is removed, every fitted model is effectively training on ∼60 target-alloy records, a regime in which conventional small-data ML pipelines on ExCorr have previously been shown to underperform in-context LLMs [39]. Pretraining has, in effect, supplied TL that the explicit Mg-alloy source pool turned out not to provide. This TL probably happens on a different level and covers rather general molecular properties and chemical knowledge and leads to a more efficient training sample usage.

The detailed cross-model comparison in the supplementary material confirms the same ordering: across all filtered/Exact settings the in-context LLM is the only method that retains useful predictive power, while the fitted baselines close the gap only in the unfiltered settings where the explicit transfer pool already supplies the labelled-molecule information.

3.7. Limitations and future work

The zero-shot probe cannot prove whether the ExCorr dataset is absent from Gemini's pretraining corpus, only that pretraining-corpus recall of labels is not driving the in-context performance: Gemini has some prior knowledge of Mg-corrosion chemistry, but at a low level consistent with the zero-shot Pearson r of 0.15–0.22 and R^2 values of −0.17 to −0.66. Should future models exhibit a substantial zero-shot correlation, the blinding methodology of our companion study [40] allows distinguishing memorisation from genuine chemical knowledge. However, for the present setting the near-absent zero-shot signal makes that distinction unnecessary.

The null SAR-transfer result found in this study is partly a property of ExCorr itself. Because the close pool consists almost entirely of molecules already present on the target alloy, the filtered setting removes most of the pool and leaves too few new molecules to reveal whether genuine SAR transfer is possible. A broader source pool, with inhibitor libraries deliberately designed to overlap less between alloys, could separate SAR transfer from labelled-molecule re-use more cleanly.

The Mg-0.8Ca contrast deliberately combines differences in both alloy chemistry and experimental protocol (Section 3.5), making it a particularly stringent test case for cross-alloy transfer learning. Unlike the cluster-resident alloys, Mg-0.8Ca occupies an out-of-distribution region of the dataset with respect to both materials characteristics and testing conditions. The analysis of SI Sec. S11 shows that the concentration difference between the two screening campaigns alone is sufficient to eliminate ranking similarity. Same-alloy, same-electrolyte rankings remain highly correlated under moderate concentration changes but collapse at the dilution factor separating the Mg-0.8Ca campaign from the main one. Consequently, the outside-cluster position of Mg-0.8Ca can most likely not be attributed solely to its Ca-containing microstructure and the observed drop in correlation is more appropriately interpreted as the combined consequence of alloy chemistry and experimental boundary conditions. Disentangling the two effects would require remeasuring a subset of reference inhibitors on Mg-0.8Ca under the conditions of the main screening campaign. Such an experiment could potentially isolate the alloy-chemistry contribution and represents the most informative follow-up to the present study. Importantly, the central transfer learning conclusion does not depend on which factor dominates since in the present dataset cross-alloy TL fails whenever the inhibitor ranking of the target domain becomes decorrelated from that of the source domains. This observation seems to be independent of the physical origin of that decorrelation and substantiates why Mg-0.8Ca serves as a deliberately challenging stress test of the transfer learning framework rather than as a general representative of Mg–Ca alloys.

4. Conclusion

Cross-alloy TL on ExCorr is technically effective but practically limited: the only information that transfers is a near-monotone IE re-scaling for molecules already measured on a correlated alloy. The inhibitor-SAR component, which is what would make TL useful for screening genuinely new candidates, contributes only insignificantly. Given ExCorr's heavy cross-alloy molecular overlap, this null result does not rule out SAR transfer in principle but shows that the current experimental database cannot expose it.

Adding Mg-alloy source data significantly improves prediction accuracy, but the gain vanishes entirely under molecule-disjoint filtering for every fitted model. Only Gemini 3.1 Pro retains predictive power in the filtered, ∼60-molecule regime, with pretraining supplying the transfer the explicit source pool does not [39].

The inhibitor-SAR contribution is zero: non-overlapping source molecules do not improve prediction. The entire correlation gain stems from alloy-context transfer, i.e. a monotone IE re-scaling for molecules already tested on a cluster-neighbour alloy.

Seven Mg-based materials (six commercial alloys and the high-purity variant HPMg51) share a common inhibitor ranking (pairwise Spearman $r \geq 0.67$)—they agree on which molecules inhibit best, though not necessarily on absolute IE values. This ranking similarity is what enables the alloy-context re-scaling. Mg-0.8Ca, the sole outlier (fitted-model $R^2 \leq -2$ for transfer settings), is consistent with cross-alloy TL being bound to the direct similarity of molecule labels between alloys.

Within the SAR-similarity cluster, alloy similarity itself is therefore not the bottleneck for cross-alloy TL—the chemical diversity of the experimental molecule set is. Future Mg-inhibitor screening campaigns on cluster-resident alloys should prioritise testing chemically distinct molecules across alloys rather than re-measuring previously tested molecules on each new cluster member: the latter only re-encodes information already in the joint dataset, whereas the former would let cross-alloy TL deliver the SAR-transfer potential the present database cannot expose.

Finally, the present study highlights an important limitation of current QSAR/QSPR approaches for dissolution modulator discovery which is not limited to Mg-based materials. The models commonly rely primarily on descriptors derived from the chemical structure of the tested compounds. While such descriptors are essential and work well for studies using data of a single metal or one of its alloys, the present results show that a generalisable model that is predictive of inhibitor performance cannot be derived based on molecular properties alone. The results indicate (and this does not come as a surprise) that the response depends strongly on the investigated material, its impurity state, microstructure as well as on the experimental boundary conditions under which the inhibition performance of the small organic molecules is assessed. This became particularly apparent for Mg-0.8Ca, where the reduced correlation with the remaining Mg materials cannot be assigned unambiguously to the distinct Ca-containing alloy chemistry or to differences in the testing protocol alone, but most likely reflects a combination of both. A further complication is that corrosion inhibition data in the literature are reported using different performance metrics, such as inhibition efficiency, inhibition power [49] or symmetrised inhibition efficiency [50] to name a few concepts. These metrics are not always directly interchangeable, because conversion between them requires knowledge of the underlying uninhibited baseline experiment, reference corrosion rate or reference hydrogen-evolution/weight-loss value, which is often not reported. Consequently, even nominally similar inhibitor performance values may encode different experimental assumptions, further limiting the direct comparability of datasets. As a follow-up to a recent review on the past decades of corrosion-inhibitor research [14], a unified string-based representation, conceptually related to SMILES, is currently being developed to encode inhibitor, material and environment information in a single machine-readable format. This representation is intended to enable the training of next-generation QSAR/QSPR models by combining relevant features of the tested modulators with rich metadata from experimental protocols and tested substrate to create the foundation for more comprehensive models.

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. **Christian J. Cyron:** Writing – review & editing, Funding acquisition. **Roland C. Aydin:** Writing – review & editing, Supervision, Conceptualization. **Christian Feiler:** Writing – original draft, Visualization, Supervision, Conceptualization.

Code availability

The source code for all experiments, including LLM prompt templates, MPNN training scripts, and analysis pipelines, is available at https://github.com/MatthiasHBusch/CrossAlloyTL.

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

During the preparation of this work the authors used Claude Code 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

Funding via project 535656357 by Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged: 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.

Acknowledgements

M.B., S.V.L., M.L.Z., C.J.C., R.C.A. and C.F. acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG; project 535656357). We gratefully acknowledge the support of the Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF).

Appendix A. Supplementary data

The supplementary information file `supplement.pdf` provides per-alloy record counts and per-fold test/pool molecule-overlap statistics (Tables S1–S2), full per-model per-setting metrics (R^2, Pearson r and MAE; Tables S3–S5), the per-target-alloy R^2 breakdown (Fig. S1, Table S6), the cluster-robustness analysis with bootstrap CIs and anchor-exclusion partial correlations (Table S7, Fig. S2), pairwise IE-scatter examples (Figs. S3–S4), the complete Mg-0.8Ca contrast results for the fitted models (Table S8) and for Gemini 3.1 Pro (Table S9), feature/hyperparameter details (Tables S10–S11), the paired Wilcoxon comparison of the Exact and Close-filtered settings (Table S12), the Gemini 3.1 Pro zero-shot contamination check (Table S13), the Pearson r and MAE companion figure to main text Fig. 1 (Fig. S5), pooled Pearson-r p-values (Table S14), the covariate-shift analysis of the

transfer pools (Sec. S10, Fig. S6), the rank-stability analysis under inhibitor-concentration changes (Sec. S11, Fig. S7, Table S15), the ChemProp hyperparameter-sensitivity grid (Sec. S12), the in-context label-shuffle and prompt-order controls (Sec. S13, Table S17), the fold- and sample-selection robustness checks (Sec. S14, Table S18), the descriptor-enriched shallow baselines (Sec. S15, Table S19), the continuous alloy-composition encoding variants (Sec. S16, Table S20), and the measured materials characteristics and boundary conditions (Sec. S17, Table S21).

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.corsci.2026.114276.

Data availability

The experimental corrosion dataset and all model predictions are available at https://github.com/MatthiasHBusch/CrossAlloyTL.

References

[1] 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.

[2] A. Dziubińska, A. Gontarz, M.-y.-s.-ł. 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.

[3] 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 (2008) 63–72, http://dx.doi.org/10.1016/j.cossms.2009.04.001.

[4] 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.

[5] 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.

[6] 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.

[7] Z. Ma, D.R. MacFarlane, M. Kar, Mg cathode materials and electrolytes for rechargeable Mg batteries: A review, Batter. & Supercaps 2 (2019) 115–127, http://dx.doi.org/10.1002/batt.201800102.

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

[9] G. Song, A. Atrens, Understanding magnesium corrosion—A framework for improved alloy performance, Adv. Eng. Mater. 5 (12) (2003) 837–858, http://dx.doi.org/10.1002/adem.200310405.

[10] A. Atrens, G. Song, M. Liu, Z. Shi, F. Cao, M.S. Dargusch, Review of recent developments in the field of magnesium corrosion, Adv. Eng. Mater. 17 (4) (2015) 400–453, http://dx.doi.org/10.1002/adem.201400434.

[11] J. Hanawalt, C. Nelson, J. Peloubet, Corrosion studies of magnesium and its alloys, Trans. the AIME 147 (1942) 273–299.

[12] 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) 8489, http://dx.doi.org/10.3390/ma15238489.

[13] 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.

[14] 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.

[15] 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.

[16] D.A. Winkler, Predicting the performance of organic corrosion inhibitors, Metals 7 (2017) 553, http://dx.doi.org/10.3390/met7120553.

[17] 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.

[18] E.J. Schiessler, T. Würger, S.V. Lamaka, R.H. Meißner, C.J. Cyron, M.L. Zheludkevich, 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.

[19] E.J. Schiessler, T. Würger, B. Vaghefinazari, S.V. Lamaka, R.H. Meißner, C.J. Cyron, M.L. Zheludkevich, 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.

[20] 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.

[21] M. Fernandez, M. Breedon, I.S. Cole, A.S. Barnard, Modeling corrosion inhibition efficacy of small organic molecules as non-toxic chromate alternatives using CoMSA, Chemosphere 160 (2016) 80–88, http://dx.doi.org/10.1016/j.chemosphere.2016.06.044.

[22] 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.

[23] 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 (2020) 5624–5635, http://dx.doi.org/10.1021/acs.jpcc.9b09538.

[24] D. Winkler, M. Breedon, A. Hughes, F. Burden, A. Barnard, T. 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.

[25] D. Höche, C. Blawert, S.V. Lamaka, N. Scharnagl, C. Mendis, M.L. Zheludkevich, The effect of iron re-deposition on the corrosion of impurity-containing magnesium, Phys. Chem. Chem. Phys. 18 (2) (2016) 1279–1291, http://dx.doi.org/10.1039/C5CP05577F.

[26] S. Lamaka, D. Höche, R. Petrauskas, C. Blawert, M. Zheludkevich, A new concept for corrosion inhibition of magnesium: Suppression of iron re-deposition, Electrochem. Commun. 62 (2016) 5–8, http://dx.doi.org/10.1016/j.elecom.2015.10.023.

[27] D.J. Diller, D.W. Hobbs, Understanding hERG inhibition with QSAR models based on a one-dimensional molecular representation, J. Comput. Aided Mol. Des. 21 (7) (2007) 379–393, http://dx.doi.org/10.1007/s10822-007-9122-2.

[28] G. Falcón-Cano, A. Morales-Helguera, H. Lambert, M.-Á. Cabrera-Pérez, C. Molina, hERG toxicity prediction in early drug discovery using extreme gradient boosting and isometric stratified ensemble mapping, Sci. Rep. 15 (2025) 15585, http://dx.doi.org/10.1038/s41598-025-99766-3.

[29] S. Faramarzi, A. Bassan, K.P. Cross, X. Yang, G.J. Myatt, D.A. Volpe, L. Stavitskaya, Novel (Q)SAR models for prediction of reversible and time-dependent inhibition of cytochrome P450 enzymes, Front. Pharmacol. 15 (2024) 1451164, http://dx.doi.org/10.3389/fphar.2024.1451164.

[30] H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, R. Yoshida, Predicting materials properties with little data using shotgun transfer learning, ACS Central Sci. 5 (2019) 1717–1730, http://dx.doi.org/10.1021/acscentsci.9b00804.

[31] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, J. Pei, Transfer learning for drug discovery, J. Med. Chem. 63 (2020) 8683–8694, http://dx.doi.org/10.1021/acs.jmedchem.9b02147.

[32] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359, http://dx.doi.org/10.1109/TKDE.2009.191.

[33] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3 (9) (2016) http://dx.doi.org/10.1186/s40537-016-0043-6.

[34] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? in: Advances in Neural Information Processing Systems, Vol. 27, 2014, pp. 3320–3328.

[35] A.D. White, The future of chemistry is language, Nat. Rev. Chem. 7 (2023) 457–458, http://dx.doi.org/10.1038/s41570-023-00502-0.

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

[37] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N.V. Chawla, O. Wiest, X. Zhang, What can large language models do in chemistry? A comprehensive benchmark on eight tasks, in: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2023, arXiv:2305.18365.

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

[39] M. Busch, M. Tacke, S.V. Lamaka, M.L. Zheludkevich, K. Linka, C.J. Cyron, C. Feiler, R.C. Aydin, Large language models predicting the corrosion inhibition efficiency of magnesium dissolution modulators, Corros. Sci. 255 (2025) 113080, http://dx.doi.org/10.1016/j.corsci.2025.113080.

[40] M. Busch, M. Tacke, S.V. Lamaka, M.L. Zheludkevich, C.J. Cyron, C. Feiler, R.C. Aydin, In-context molecular property prediction with LLMs: A blinding study on memorization and knowledge conflicts, 2026, ArXiv Preprint. arXiv:2603.25857.

[41] D. Mei, S.V. Lamaka, C. Feiler, M.L. Zheludkevich, The effect of small-molecule bio-relevant organic components at low concentration on the corrosion of commercially pure Mg and Mg-0.8Ca alloy: An overall perspective, Corros. Sci. 153 (2019) 258–271, http://dx.doi.org/10.1016/j.corsci.2019.03.039.

[42] H.L. Morgan, The generation of a unique machine description for chemical structures — a technique developed at chemical abstracts service, J. Chem. Doc. 5 (1965) 107–113, http://dx.doi.org/10.1021/c160017a018.

[43] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754, http://dx.doi.org/10.1021/ci100050t.

[44] G. Landrum, RDKit: Open-source cheminformatics, 2024, https://www.rdkit.org, Version 2024.03.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[46] 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 (2019) 3370–3388, http://dx.doi.org/10.1021/acs.jcim.9b00237.

[47] 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 (2024) 9–17, http://dx.doi.org/10.1021/acs.jcim.3c01250.

[48] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Rethinking the role of demonstrations: What makes in-context learning work? in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2022, pp. 11048–11064, http://dx.doi.org/10.18653/v1/2022.emnlp-main.759.

[49] A. Kokalj, M. Lozinšek, B. Kapun, P. Taheri, S. Neupane, P. Losada-Pérez, C. Xie, S. Stavber, D. Crespo, F.U. Renner, A. Mol, I. Milošev, Simplistic correlations between molecular electronic properties and inhibition efficiencies: Do they really exist? Corros. Sci. 179 (2021) 108856, http://dx.doi.org/10.1016/j.corsci.2020.108856.

[50] A. Kokalj, C. Xie, I. Milošev, D. Crespo, How relevant are molecular electronic parameters for predicting corrosion inhibition efficiency: imidazoles as corrosion inhibitors of Cu/Zr materials in NaCl solution, Corros. Sci. 193 (2021) 109900, http://dx.doi.org/10.1016/j.corsci.2021.109900.