Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MILES: a learnable multiscale materials language for corrosion science and beyond

Vipul Gupta
Helmholtz-Zentrum Hereon https://orcid.org/0000-0002-6311-4422
Tim Würger
Helmholtz-Zentrum Hereon
Johannes M.C. Mol
Delft University of Technology https://orcid.org/0000-0003-1810-5145
Sviatlana Lamaka
sviatlana.lamaka@hereon.de
Helmholtz-Zentrum Hereon https://orcid.org/0000-0002-0349-0899
Christian Feiler
Helmholtz-Zentrum Hereon https://orcid.org/0000-0003-4312-7629
David Winkler
La Trobe University https://orcid.org/0000-0002-7301-6076
Majid Laleh
University of Wollongong
Mikhail Zheludkevich
Helmholtz-Zentrum Hereon
Anthony Hughes
Deakin University

Research Article

Keywords: Multiscale material descriptors, text-based material representation, data standardization, machine learning, language modelling, FAIR data

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10871829/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

MILES: a learnable multiscale materials language for corrosion science and beyond

V. Gupta[1, †], T. Würger[2, †], J.M.C. Mol[3], S.V. Lamaka[2*], C. Feiler[2,4,5] D. A. Winkler[6,7,8], M. Laleh[9], M. L. Zheludkevich[2,5] and A.E. Hughes[10,11].

[1] *Institute of Materials Physics, Helmholtz-Zentrum Hereon, 21502 Geesthacht, Germany; vipul.gupta@hereon.de*
[2] *Institute of Surface Science, Helmholtz-Zentrum Hereon, 21502 Geesthacht, Germany*
[3] *Department of Materials Science and Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands; j.m.c.mol@tudelft.nl*
[4] *Institute for Interface Physics and Engineering, Hamburg University of Technology, Am Irrgarten 3-9, 21073 Hamburg, Germany*
[5] *Kiel Nano, Surface and Interface Science KiNSIS, Kiel University, 24118 Kiel, Germany*
[6] *Department of Biochemistry and Chemistry, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria, 3086, Australia; d.winkler@latrobe.edu.au*
[7] *Monash Institute of Pharmaceutical Sciences, Monash University, 392 Royal Parade, Parkville 3052, Australia; david.winkler@monash.edu*
[8] *School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, UK; paadw4@nottingham.ac.uk*
[9] *Photon Forge Additive Manufacturing Facility, John de Laeter Centre (JdLC), Curtin University, Perth, WA 6102, Australia*
[10] *CSIRO Mineral Resources, Bayview Avenue, Clayton, 3168, Australia; tony.hughes@csiro.au*
[11] *Institute for Frontier Materials, Deakin University, Waurn Ponds, VIC 3216, Australia*

† These authors contributed equally to this work.

*corresponding author: Sviatlana Lamaka Sviatlana.lamaka@hereon.de , Helmholtz-Zentrum Hereon.

Keywords: Multiscale material descriptors, text-based material representation, data standardization, machine learning, language modelling, FAIR data

Abstract

Machine learning (ML) models continue to accelerate materials design and discovery. They rely on atomic/molecular-scale descriptors that represent entities used for training at the corresponding scale. However, as scales increase, descriptors must also span scales, creating the “multiscale modelling dilemma". Here, we introduce **M**aterial **I**nput **L**ine **E**ntry **S**ystem (MILES) for encoding descriptors that span the micro-to-macro scales. MILES encapsulates a nested series of descriptors for a range of metrics from materials, characterization, performance, or mechanical data. MILES tokens begin with a two-letter code that designates a property followed by its value(s). Tokens are arranged in levels in a modular structure to capture information in a MILES string. We demonstrate a transformer-based encoder that, when trained on ExCorr corrosion inhibitor database-derived MILES strings, learns latent representations capturing material-, inhibitor-, and performance-dependent organization. A model trained on MILES predicts inhibitor performance on untested alloys. These results highlight MILES’s potential as a materials language for ML. We also describe the software tool `milescoder` that simplifies MILES encoding and decoding as well as demonstrate automated extraction of literature information for constructing MILES strings using `MILES-GPT`. We furthermore describe an ablation study showing improved latent-space organization when MILES strings are augmented with literature-mined information.

Introduction

Increasingly, artificial intelligence (AI), particularly the machine learning (ML) is making dramatic contributions to materials science, from accelerating design of new materials to the maintenance of industrial structures. Materials science benefits from leveraging the success of these powerful technologies in many other fields as diverse as economics and protein structure prediction [1, 2]. In materials science, AI and ML facilitate the design of new materials with superior properties for bespoke applications such as high entropy alloys [3-6], organic photovoltaic materials [7, 8], and microstructural design in Mg-Al alloys [9]. Documented applications to corrosion problems such as pipeline corrosion have increased exponentially between 2020 and 2024 [10].

The diverse successes in many branches of materials science have stimulated considerable interest in using these methods to find better corrosion inhibitors, the exemplar for this work, for a range of industrial metals and alloys [11, 12]. However, current applications of ML to corrosion (and indeed some other important areas of materials science) are still limited. One limitation is the use of a single metric (e.g. inhibition efficiency (%IE), inhibition power (IP) or symmetrized inhibition efficiency metrics) to describe corrosion inhibition performance [12-14]. These metrics are typically acquired on systems where other experimental parameters such as the metal composition and surface condition, electrolyte composition, concentration, temperature, oxygen concentrations, etc., are fixed. In such studies, the ML models are trained almost exclusively on the molecular features of the inhibitor molecules and typically do not include any information on the corrosion system more generally. While these obvious simplifications have been useful, they are clearly only valid for metal/corrosion environment systems on which they are trained and therefore cannot generalize to other metals, alloys, or environments. They also essentially integrate all aspects of the surface and many aspects of the electrolyte into a single performance metric. Additionally, many studies examined a small number of chemically similar inhibitor molecules that resulted in very limited ability to generalize to new inhibitors and were prone to chance correlations. More recent, chemically diverse studies employed hundreds of different inhibitor molecules [15, 16]. More realistic representations of inhibitor-surface environments require the inclusion of the effects of water (and potentially other electrolytes), ions, and metal/alloy compositions at the atomistic scale that capture the interactions between the inhibitor molecule and the surface and dissolved ions. While these interactions can be simulated by Monte Carlo (MC) and molecular dynamics (MD) calculations, they are very time- and resource-intensive. This resource issue expands markedly as the system being modelled increases in size and number of components. Moreover, the surface layer and the bulk are usually not homogeneous, rendering the development of a digital twin mimicking real conditions highly challenging.

Consequently, ML models of corrosion inhibitors have had only limited success because model predictions are only valid for the same metal or alloy and experimental conditions. Rolling up all the surface and electrolyte parameters into one metric is not realistic for training corrosion inhibitor models that can guide the design or optimization of better corrosion inhibitors for a variety of metals under real service conditions. Clearly, these other essential corrosion systems parameters need to be incorporated into the training of ML models to enable prediction of properties of inhibitors widely across metals, alloys, and service conditions. It is rational to conclude that establishing correlations between inhibition efficiency metrics and substrate microstructure, surface condition, and electrolyte as well as inhibitor chemistry should significantly improve the predictive, interpretative, and applicability/generalizability (across different metals) properties of ML models.

A study of the current extent of incorporation of non-molecular metrics in ML models reveals a mixed story, largely due to a paucity of publications. A review by Coelho et al. [17] examined 19 papers and found five reported models in which alloy composition was included and two employing additional metallurgical information. Yan et al. [18] included bulk composition of 18 low alloy steels in a study of marine corrosion using ML. Surprisingly, the alloy composition was not found to be an important factor in corrosion

performance [19]. Similarly, Liu et al. included corrosion system metrics such as redox potential, salinity, and pH, and other variables in an ML model and found the redox potential to be important [20]. Rojas et al. [21] identified the role of Ni in stress corrosion cracking (SCC) through microstructural properties such as heat treatment and mechanical properties as well chemical composition. Murthy applied ML to model SCC concluding that corrosion caused acceleration of SCC [22]. Wang et al. [23] provided a good overview of the synergistic combination of high-throughput testing (including alloy fabrication) for ML modelling. Zhi et al. [24] published a random forest ML modelling study of several low alloy steels (LAS) of varied alloy composition and environmental exposures and found that the latter (particularly pH) dominated performance. They also noted that lack of microstructural information limited their ability to report on this aspect in their study. Olawole et al. [25] assessed methods of determining the important surface parameter of metal work function in training ML models of corrosion. However, many reported studies do not provide any information about the metals under study [26-29]. Elmas et al. [29] used random forest models to predict corrosion susceptibility in offshore oil platforms but did not include metal microstructure structure. In its absence, their data included a range of metrics capturing the influence of time, environment, coatings, and location on offshore structure. The review by Anadebe et al. [30] clearly showed that, in industrial applications, the severity of environmental conditions, dosing, and sensor input are important metrics for ML.

Other important issues are inherent in corrosion that are also issues in many other areas of materials science. One is the information gap arising from the enormous differences in scale between the atomic/molecular level to the working materials component level (**Figure 1**). This is exacerbated by the large orders of magnitude differences in timescales, from rapid molecular movements and reactions to the extended timescales over which materials degrade. Clearly, finding an effective way to encode these scales and surface chemistry and microstructure, electrolyte, and ions would be desirable. Attempts to bridge these gaps have involved extending existing metrics used to describe molecules to larger temporal and size scales. The extension of the InChI system for encoding molecular structures as text strings to chemical reactions (RInChI) [31] or the nanomaterials (NInChI) [32] are two examples. However, these approaches quickly encounter problems when the scale is extended even to the size of nanoparticles [32] suggesting that an alternative approach is required. However, they provide encouragement that multilevel textual strings may be a productive path to pursue.

Corrosion, in common with other areas of materials science, also encounters issues around the quantity and type of data used to train ML models. One major problem is the need for expanded high throughput (HT) corrosion inhibition measurement methods [33] to generate sufficient experimental data for ML training and wider use of them. As stated, current methods commonly assess many candidate molecules on a single alloy. This only generates models for a single alloy type. For example, HT studies have been used to generate larger datasets for Mg alloys, aiming to find efficient corrosion inhibitors [34] and for electrolyte additives to support dissolution control of Mg anodes for Mg-air batteries [35]. These models employed calculations of local atomic environments using the smooth overlap of atomic positions (SOAP) to explore the structure-property landscape [35-37] or molecular descriptors to predict inhibition efficiencies [15, 38-40]. While the experimental data for nine different Mg alloys were available, the models were limited to one alloy in each study as the molecular descriptors and similarity values do not encode differences in alloy composition or microstructure. As a result, the same inhibitor could be associated with up to nine different labels without providing the models with the context needed to unravel the differences.

A shortcoming of the current approach, which is a focus of this paper, is described by Winkler et al.[16]. They reported the %IE of 100 inhibitors on two Al alloys, AA2024-T3 and AA7075-T6 and developed ML models predicting %IE using descriptors of the inhibitor molecules alone. The most relevant descriptors differed completely between the two alloys. This appears incongruous when the same inhibitors were used in the same electrolytes on two Al-alloys that were both over 90% α-Al. This strongly suggests that the different elemental composition and size, distribution and electrochemical activity of intermetallic particles in the two alloys significantly affected inhibitor effectiveness, and that inhibitor descriptors alone did not capture this.

Embedding metrics reflecting the underlying microstructural elements of the substrate and spatially distributed surface chemistry is important and a driver for this study.

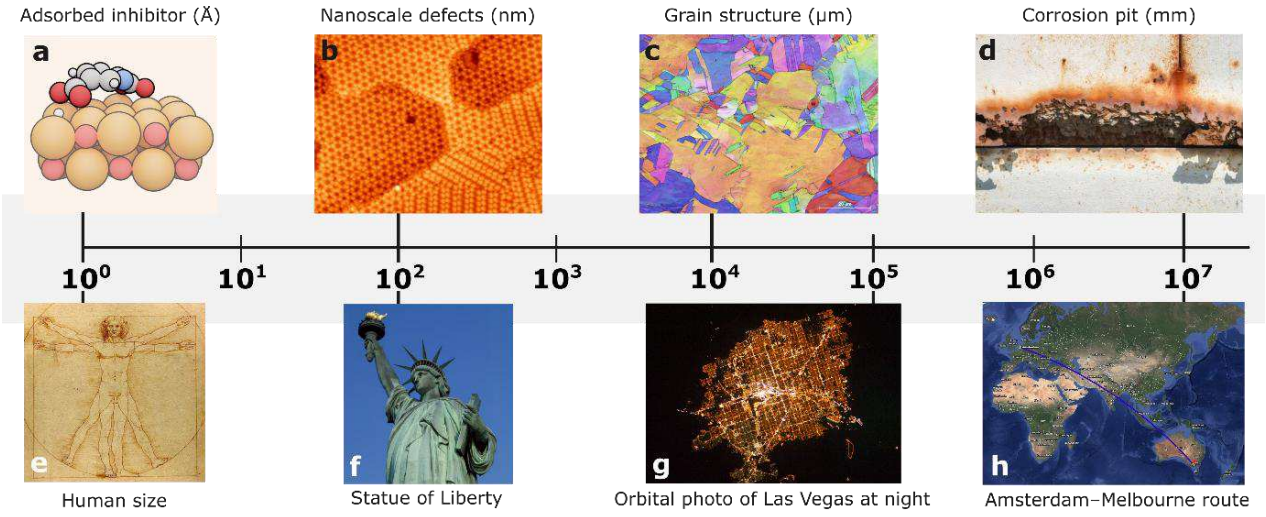


***Figure 1**: A look at the order of magnitude of scales involved in corrosion compared with the change of scales in everyday experience. In (a-d) the change in the order of magnitude of scales has a factor of ~ 10^7 from the scale of the inhibitor and surface features (a), over nanoscale defects (b) and grain structures (c) to corrosion pits (d). In (e-h) the equivalent changes in order of magnitude are going from human size (e), to the Statue of Liberty (93 m)[41] (f), the metropolitan area of Las Vegas (ca. 50 km across)[42] (g), and finally Amsterdam to Melbourne distance (ca. 17 000 km) (h).*

To address this, we describe a generalised, simple, text-based coding system for typical corrosion systems. It encodes the bulk and surface properties of metallic materials, inhibitors, solvents and ions, methods of corrosion inhibition measurement and the degradation of these materials in corrosive environments in a textual form that is suitable for training deep learning models. We aim to better characterize the entire metal-inhibitor corrosion system in a way that is useful for models to predict metal corrosion, thereby accelerating the discovery of better corrosion inhibitors. We have called the coding system the **M**aterial **I**nput **L**ine **E**ntry **S**ystem (MILES). In principle, all properties of the surface of the metal can be coded into a simple string format using a range of MILES tokens described below. MILES resembles the Simplified Molecular Input Line Entry System (SMILES) for encoding small organic molecules, but the philosophy is entirely different. It is closer in concept to the InChI system and its extensions that are used for encoding complex materials and reactions. MILES encodes material properties and experimental metadata across multiple time and length scales and can readily expand its performance metrics to cover the types and extent of corrosion problems and, importantly, materials applications more broadly. It could also provide a general machine-readable coding for experimental sections in papers, improving data availability, quality and utility. At the molecular level it is compatible with SMILES and InChI, especially where complex electrolyte constituents and inhibitors are encoded.

Surfaces cannot be treated as extensions of simple atoms or molecules in the same sense that SMILES and InChI treat the coding of molecules. Clearly, the extensions of the InChI to RInChI or NInChI[32] are inadequate to describe a surface where electrochemical reactions span scales from nanometres to millimetres (**Figure 1**). For example, if the characteristic distance between an inhibitor molecule and a surface is ~ 2Å, this is seven orders of magnitude smaller than a small corrosion pit observed visually (2mm). This is illustrated in **Figure 1 (a – d)** where the interacting molecule in (a) is small compared to grain boundaries and dislocations in (b) and is a fraction of a pixel in the electron backscattered diffraction (EBSD) map in (c). We can often not fully comprehend the enormity of the change of scale. If this order of magnitude is transferred to human scale (1.7 m), then seven orders of magnitude is the distance between Amsterdam and Melbourne, ca. 17 000 000 m, **Figure 1 (e – h)**. So, there is a strong need for metrics that span multiple orders of magnitude between the molecular scale (a) and the pit (d).

To address this “scale-gap” MILES is multilayered. Its tokens, based on two-letter codes and numerical substrate values, capture both surface and environment information and a MILES string can also carry metadata on the experiments used to measure the property of interest (e.g., corrosion rate). Tokens are organised into independent modular levels, each optional, so levels can be included or omitted, depending on data availability and application. The flexibility of MILES and current MILES codes already makes it applicable to other areas of materials science to describe more general materials’ properties. We have developed a digital toolkit for encoding and decoding MILES strings, comprising the complete source code distributed as the `milescoder` Python package (https://pypi.org/project/milescoder) as well as web-based implementations of the encoder (https://miles.hereon.de/encoder) and decoder (https://miles.hereon.de/decoder) for direct use. We also demonstrate that the system allows the extraction of rich information (e.g., metal alloy compositions, processing conditions, bulk microstructures, surface layers and corrosion performance) from the literature and constructing MILES strings using `MILES-GPT` (https://miles.hereon.de/miles-gpt).We further explore the potential of MILES strings as a learnable materials language, showing that a transformer-based encoder can learn meaningful representations of material and performance-related information encoded at different levels of MILES, and that these representations become more structured as the strings are enriched.

Results

Representations of Surfaces for Machine Learning

We begin with a general overview of how surfaces in materials science are encoded by ML models and explore their relevance to corrosion science. Computational modelling of bulk materials properties by ML is generally (and historically) referred to as quantitative structure-property relationship (QSPR) models. Here, the relevant bulk properties (e.g., composition and structure) are used in QSPR models to describe how changes in these properties influence materials performance. On the other hand, there are many circumstances such as battery electrodes, catalysis and corrosion where the material properties of the surface are important. Corrosion is an example where processes are clearly dominated by the materials surfaces and the surrounding environment. Thus, corrosion overlaps only partially with other materials informatics spaces because surface structure and chemistry are only indirectly related to bulk composition and structure. Corrosion is closer to heterogeneous catalysis than most other fields. Indeed, Medford et al.[43], highlighted differences between catalyst performance metrics based on bulk material descriptors and those based on surface descriptors. They concluded that structure-property relationships should not be trained on bulk properties but rather on (structure + environment) properties (as depicted in **Figure 2**) as changes in structure and environment can influence reaction rates by several orders of magnitude. The environment in a corrosion context includes electrolyte and a changing surface structure through corrosion reactions.

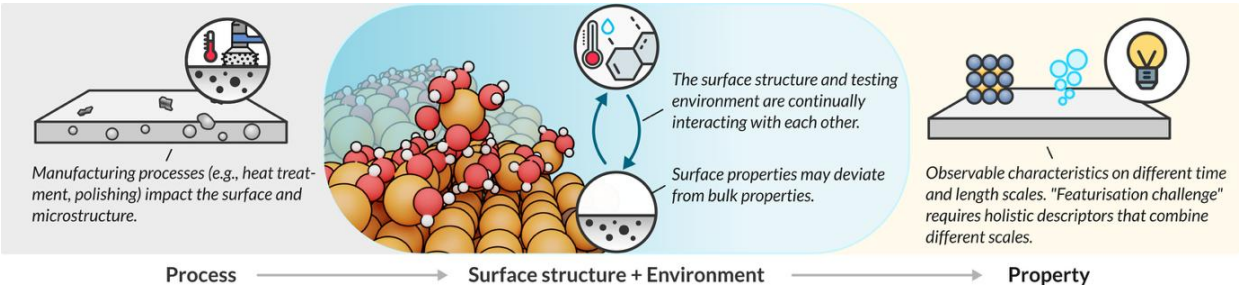


***Figure** 2: Process-(Structure & Environment)-Property relationship based on the concepts of Medford et al. [43]. In the catalysis area Medford argues that the “structure” part of process-structure-property relationships is a mixture of environment and surface. This type of approach is equally valid for a number of other fields including corrosion where the structure changes during the property assessment.*

While corrosion may resemble catalysis in terms of requiring metrics that describe surfaces, there are still significant differences, as detailed in the **Supplementary Information (SI), Table SI 1.** In corrosion, local

surface reactions also control the reactivity, and the bulk structure only relates to the subsurface susceptibility to corrosion. Once corrosion starts, new surfaces are generated, e.g. in pits or intergranular attack, that have compositional and electrochemical gradients separating them from the bulk electrolyte on the external surface.

Another important issue is the "featurisation challenge". In catalysis the most relevant features can be difficult to identify. Features control fundamental catalytic behaviour because they involve active site structures that may change under reaction conditions at the atomic and nanoscales. In corrosion, the challenge is different because surface reactions are locally rapid, and switch at early stages from random chemical/electrochemical reactions to electrochemically driven highly active sites. Such phenomena require metrics that capture irreversible changes with time and chemistry, calling for new descriptors applicable across time and length scales.

MILES Classification System and Metrics

The previous section emphasizes the complexity of surface conditions in corrosion, identifying the need for more comprehensive, relevant system descriptors or features for training ML models. **Figure 3** outlines a classification system for surfaces and provides a framework for discussing the development of MILES.

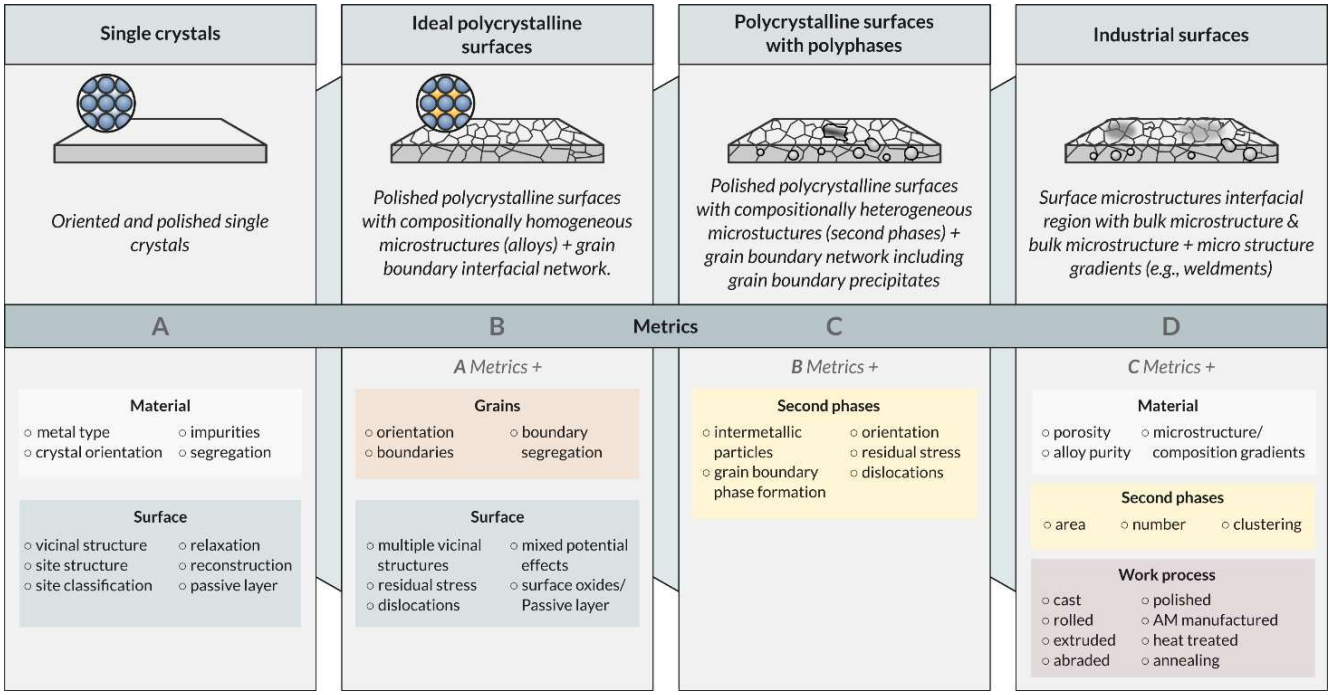


***Figure 3**: Example of the increasing complexity of surfaces in going from a pure single crystal surface to a polycrystalline material which is heavily processed and contains different surface microstructures and composition from the bulk. As well as not accounting for passive layers or surface oxidation, the figure also does not include the potentially numerous interactions with the surface by solvents, ions, inhibitors, etc.*

At the left-hand side of **Figure 3** (*A metrics*) describe single crystals and provide basic information about the structure of material phases. Crystal structures can be divided into seven different crystal groups, but here, we focus on sectioned surfaces denoted by the Miller and Bravais-Miller indices. Briefly different orientations have different fundamental properties such as minimal surface energy, oxidation and corrosion reactions. Surface properties clearly depend on metal type, metal–metal distances, electron configurations, number of exposed layers (high index planes can expose atoms below the surface) etc. Single crystal *alloys* have the additional complexity of how the alloying elements are distributed over sites and whether they segregate or aggregate.

Next (*B metrics*) are polycrystalline surfaces where alloying components are in solid solution. The surfaces of these materials contain individual crystals (grains) that are sectioned randomly (except where some structural texture is present). This results in a distribution of surface properties arising from random (stochastic) grain orientations and therefore reactivities. Notably, Esterhuizen et al. developed a general method for calculating adsorption energies on surfaces for use in ML model training [44]. Polycrystalline surfaces with polyphases (*C metrics*) require additional descriptions related to the structure and orientation of the alio-phases, interfacial characteristics, and electrochemistry between phases. Industrial processing of metals and alloys adds more complexity to the surface microstructure (*D metrics*). It can be seen in the columns at the bottom of **Figure 3** that the number of metrics accumulates as each level of complexity is added to an ideal crystal. These is addressed in detail below and in **SI**.

MILES Coding Principles

The complexity summarized in **Figure 3** suggests the need for better representations of surfaces and their interactions with exogenous materials. Given the massive increase in pattern recognition capabilities of deep learning methods like transformer-based LLMs, a text-based system akin to InChI and SMILES seems sensible. However, as noted above, InChI and SMILES approaches are too local (applicable to a simple or low-complexity molecule or material) to describe the spatial aspects of surfaces and their environments. Corrosion is not confined to local chemical reactions but is influenced by electrochemistry involving electrons generated tens of microns or even meters away. Nor can these approaches effectively represent the temporal aspects of corroding systems (corroding surfaces are in a state of flux, with materials being consumed and new ones being created). Measurements of corrosion rate or corrosion inhibitor efficiency are usually only taken at a single point in time. Moreover, surface electrochemistry depends on the local environment, which is itself modulated by electrochemical differences across the surface. These arise from surface potential variations caused by differing composition and crystal orientations of the base metal plus inclusions, oxides, grain boundaries, adsorption layers, and intermetallic particles (typically 0.1 – 10 microns in size).

MILES strings are divided into multiple levels (labelled "**L<*n*>**", where *n* is the level number starting at zero) separated by delimiters, **Figure 4**. Each level is an optional, self-contained module. When the MILES string is decoded, only the levels that are present are "catalogued", thus there is no requirement for all levels to be present and filled. Researchers can choose the amount of information to include in the MILES string. The independence and optionality of the layers increase the flexibility of the system because new MILES levels can be added at any time for specific applications, providing they conform to the master framework or ontology that defines what each level should contain. For example, simple MILES strings might encode basic information on the material and performance metrics provided by the experimental sections of research papers. More complicated strings could, in principle, also contain rich experimental metadata (**Figure 5**). In this case, each datum would require a MILES string, so a paper would generate an array of MILES strings where each entry describes an individual datum. These more comprehensive strings are data-rich, making them suitable for storing large amounts of experimental data in a condensed manner. To track future developments of MILES, a version number is included at the beginning of a MILES string. Newer versions should always be backwards-compatible to ensure that older versions are still usable.

The levels contain MILES tokens defined by two-capital letter characters. Different levels are delimited by the "/" symbol. These MILES tokens are then combined to form MILES strings. The currently implemented levels are as follows: **L0**=material, **L1**=metal, **L2**=alloy, **L3**=conventional manufacturing, **L4**=electrolyte, **L5**=additive manufacturing (current tokens are listed in **SI**). Note that the levels, discussed below in the context of metals and corrosion, can be modified to accommodate other materials contexts. Having fixed-length metric labels means that the MILES string can be decoded and extracted by text substring searches. The only other letter-based codes used in MILES are for the elements of the periodic table represented by an upper/lower case two letter code. Single-letter elements such as "O", "C", or "S" are encoded as "Oo", "Cc" or "Ss", to distinguish them from other MILES metric labels. Exceptions to this rule are elements in strings, e.g., SMILES and

compound names, which are treated explicitly. Mixed rare earths use the code “Me” (from Mischmetal). All double capital letter codes are followed by either another double-letter capital code, a number (<value>), a string or an upper/lower case code for elements. Molar concentrations, e.g., for compositions of an electrolyte, are enclosed in “[ ]” as per common usage. SMILES codes can be used to describe the chemical compounds in the electrolytes. The appearance of the next double capital letter code automatically terminates the definition of the respective MILES token.

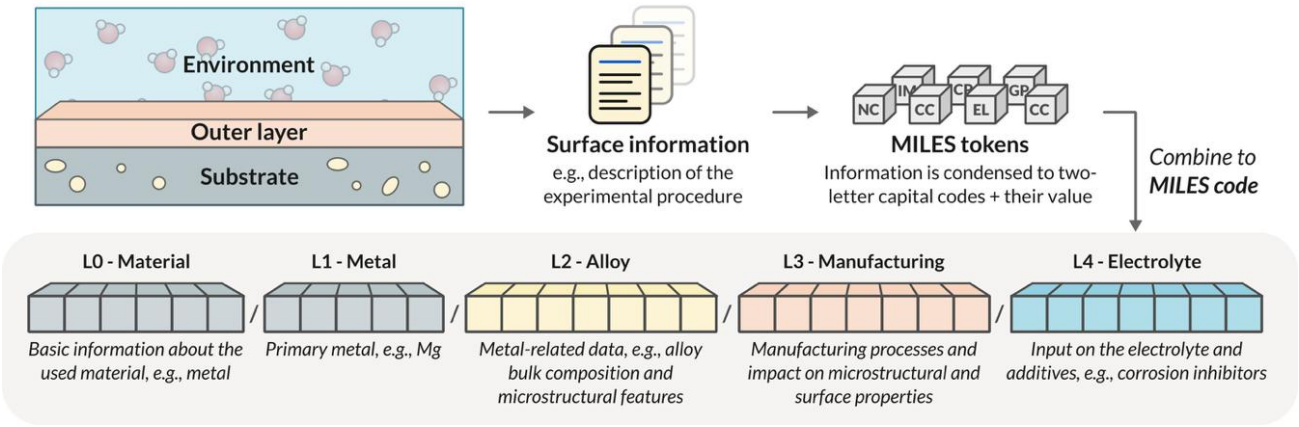


***Figure 4**: Flowchart of how information is converted to MILES string using corrosion-related example levels and the types of metrics they cover (current list, over 110 tokens, is in the **SI**). The flowchart aids human interpretation, but not all parts are required for MILES coding. Each level is independent and can be treated as a plug-in module. The MILES string levels are coloured according to the relevant site in the experiment (top left). In this case it is for a two layered system such as galvanised product, coating on a substrate or a deformed layer.*

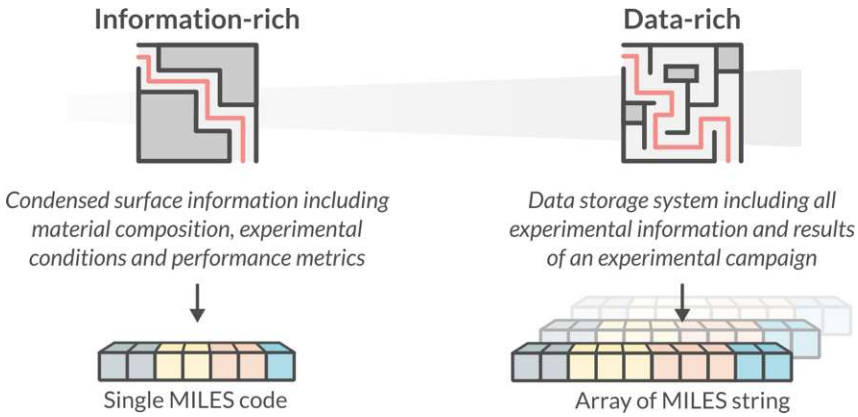


***Figure 5**: Scope of MILES coding. In the simplest case, one type of material is tested and coded. MILES string may include information from experimental sections of papers and may also include characterization results (e.g., XRD, EBSD, etc) and material performance. In complex situations, all results of a testing campaign (varying materials and experimental conditions) including experimental metadata could be encoded. In this instance a series of MILES strings would need to be generated (one for each datum). In this latter case the MILES strings would serve more like a data storage system. The colour coding represents the different levels, introduced in Figure 4.*

The types of data that can be incorporated into metal alloy level **L2** are indicated in **Figure 6**. A key metric is the alloy composition because this directly affects performance. For commercial alloys, this information is available in tables in textbooks, standards documents, or websites of metal composition. The commercial UNS alloy number could be used to directly obtain alloy composition data and incorporate it into the MILES string using the token ‘NC’, followed by the composition. In addition, information about second phases or intermetallic compounds can be included. For example, for the AA2024 (UNS A92024) aluminium alloy, this

additional information will include compositions for hardening precipitates (HP) and dispersoids (DI) in intermetallic particles (IM), as described in

Table 1.

***Table 1**: Typical alloy input tabular data encompassing the UNS number, composition in at.%, hardening precipitates, IM particles and dispersoids. This information can be available for automatic population of MILES strings.*

Alloy & UNS No	Nominal Composition String At.%	Hardening Precipitates	Intermetallics (n) = % area of IM type	Dispersoid
AA2024 A92024	Cu4.3Mg1.5Mn0.6Fe0.5Si0.5Zn0.25Ti0.15	Al2Cu Al2CuMg	Al2Cu(n) Al7Cu2Fe(n) Al19Fe4MnSi2(n) Al2CuMg (n) Al79Cu13Fe5Mn2Mg0.7Si (n) [45] Al67.7Cu3.36Fe14.29Mn6.87Mg0.43Si6.27(n) [46] Al77Cu10Fe7Mn5Si2(n) [45] Al76Cu6Fe7Mn5Si6(n) Al61Cu20Mg15(n) [47] Al51.6Cu24Mg23.5Zn0.75(n) [46] Al52.8Cu24.5Mg22.7(n) [46] Al70Cu27 (n) [47] Al63Cu33.8(n) [46]	Al20Mn3Cu2

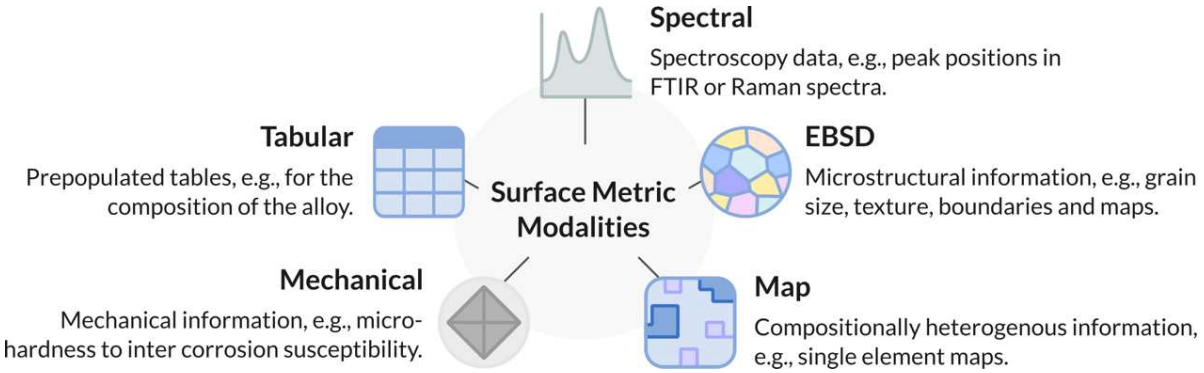


***Figure 6**: Examples of some different categories of data that could be encoded into the MILES strings for **L2**, comprising information about the metal alloy and microstructure. In this context we have explicitly separated Spectral, EBSD and Map information as examples of data that might be extracted using LLM Models. Thus, the examples are not intended to be a complete list but instead highlights the potential widespread nature of input data.*

Table 2 lists some examples of MILES tokens from different levels and information that can be encoded by them. The left-hand column gives the token name, the middle column provides the level, and the right-hand column defines the token and additional comments. The first token in the table (*EL*) denotes the composition of the electrolyte, where the chemical formula can be encoded using a SMILES format and a concentration. Multicomponent electrolytes are constructed by using *EL* repetitively via the encoder described below. The second and third entries relate to grain boundary phases and grain boundary lengths. The fourth and fifth are examples of tokens that share the same letters (*IM*) but are in different levels so are unambiguously unique. The first entry is for intermetallic compositions and the second is for immersion time. *NC* is the nominal composition of a material. *SM* indicates that the next token will be a SMILES string followed by a concentration for the compound. Finally, *LH* from **L5** level refers to the hatch distance used in laser powder bed fusion additive manufacturing. Note, that many of these tokens are not corrosion specific but relate to materials more generally, indicating that MILES can be deployed more extensively in materials science. Although each level (L<*n*>) has its own set of tokens, the two letters can be reused in other levels and have a different meaning because each level (delimited using "/") contains separate instructions and no ambiguity will arise. Ultimately, the encoding process generates an alphanumeric string with delimiters "/" and "number-letter" boundaries. Round "( )" and square "[ ]" brackets are also used to ensure there is no ambiguous coding. A

current list, comprising over 110 suggested tokens, is in the **SI, Table SI 2**. Additional tokens proposed through GitHub pull requests can be readily incorporate into `milescoder` and `MILES-GPT` tools.

***Table 2**: Examples of MILES tokens extracted from different levels*

Token (Metric)	Level	Type(s)	Unit	Definition
ELCN"name"SM"smiles"[<molarity>]	**L4**	string string number	– – mol/L	Electrolyte composition (*EL*) as a combination of compound names (*CN*), SMILES strings (*SM*) and concentrations given in molarity. Note: *IT*<medium> can be used for in-vitro electrolytes.
GP<element, content>[ ]	**L2**	string number	– at.%	Grain boundary phase decorating the grain boundaries, including the phase stoichiometry / composition.
GX(L2/IM_n) <value> GX(IM_1/IM_2) <value>	**L2**	number	m/cm^2	Grain boundary length that one phase shares with a second phase expressed in units of m/cm^2. This type of information is available from EBSD. Here **L2** is the alloy and IM_1, IM_2 are two intermetallic phases in contact with it and each other. These subscripted IM's are the tokens for intermetallic particle compositions stored in an array but input via the IM MILES code.
IM<element, content>[ ][ ]	**L2**	string number	– at.%	Intermetallic composition as an array of compositions.
IM<time>	**L4**	number	hours	Immersion time.
NC<element, content>[ ][ ]	**L2**	string number	– at.%	Nominal composition expressed as a series of elements followed by the content in at.%, e.g., Mg94.0Zn6.0. The element/number junction is the delimiter.
SM<SMILES>[<molarity>]	**L4**		–	SMILES string, e.g., for the components of the electrolyte followed by the molarity.
LH<spacing>	**L5**	number	µm	Hatch spacing for the scanning pattern of the laser in laser powder bed fusion (LPBF) additive manufacturing (AM).

MILES String Examples

MILES coding is exemplified by the case study below. A wrought aluminium alloy AA2024-T3 sheet that has been rolled, with a surface deformed layer 1.0 µm thick. It contains Al_7Cu_2Fe, Al_2Cu, and Al_2CuMg intermetallic particles with a clustering index of 3 and 30% (111) and 20% (110) preferred grain orientation. It has a average grain size of 5.5 µm, a grain boundary depleted zone of 2 nm, a boundary length of 6.7 m/cm^2, 36% low angle grain boundaries, and Al_2Cu grain boundary precipitates. The surface exhibits shingling. It has been exposed to a 0.1M NaCl electrolyte. The electrolyte contains a 2-mercaptobenzothiazole corrosion inhibitor (SMILES: C1=CC=C2C(=C1)NC(=S)S2) at a concentration of 0.03M that exhibits an %IE of 70 and pitting density of 10 pits/cm^2. This example corresponds to the following MILES string:

```
0.1A/L0ME/L1Al/L2ID"AA2024-T3"IMAl70.0Cu20.0Fe10.0IMAl66.7Cu33.3IMAl50.0C
u25.0Mg25.0TC"T3"BL6.7GZ2.0GS5.5CO30.0(111)CO20.0(110)LA36.0IC3.0/L3DE1.0
SHWR/L4PR10.0ELCN"NaCl"[0.1]CN"2-mercaptobenzothiazole"SM"C1=CC=C2C(=C1)N
C(=S)S2"[0.03]IE70.0
```

Alternatively, none of the clustering information, EBSD, oxide information, or pitting data may be known. So, a data-poor example for the same experiment may resemble:

```
0.1A/L0ME/L1Al/L2ID"AA2024-T3"IMAl70.0Cu20.0Fe10.0IMAl66.7Cu33.3
IMAl50.0Cu25.0Mg25.0TC"T3"/L3SHWR/L4ELCN"NaCl"[0.1]CN"2-
mercaptobenzothiazole"SM"C1=CC=C2C(=C1)NC(=S)S2"[0.03]IE70.0
```

The first, information-rich MILES string uses 239 characters encoding information about the surface, electrolyte, and material performance in solution; the second one uses only 180 characters, but this still captures a significant amount of information about the surface, which subsequently can be decoded and used to train ML models.

The MILES string captures the final microstructure, the one exposed to the electrolyte, without requiring its processing history, though some manufacturing details can be included. Two further examples, for stainless steel and magnesium alloys, are given in **SI**. Importantly, most of the encoded information concerns material properties, highlighting how MILES could extend areas of materials performance beyond corrosion, with other performance metrics replacing corrosion rate and inhibition efficiency.

Learning Textual Representations of Materials Using MILES

Since the first direct use of textual representations of molecules for ML models in the mid-2010s, the number of examples where SMILES and InChI have been used to generate models of molecular properties using deep learning algorithms and LLMs has been increasing rapidly. Much of this work on using LLMs with textual representation of molecules is very recent [48-50] and many exist as preprints. SMILES is widely used in the cheminformatics field and there are more publications using this textual representation of molecules, or variants such as BIGSMILES for polymers [51-57]. LLMs employing InChI for encoding molecules or reactions have also been reported [54, 58, 59]. While the examples are for small discrete molecules or polymers, they illustrate the potential of LLMs to use textual molecular information to generate useful, predictive, quantitative models.

Information-rich descriptors such as MILES provide the necessary foundation for LLMs to achieve their full potential in materials-related fields. An LLM can capture the obvious and latent connections in the encoded materials data, leading to improved prediction accuracy compared to models trained on simpler descriptors and/or simpler model architectures. The complex interplay of different length scales in MILES promises to reveal new insights in materials behaviour, for example in corrosion. To demonstrate this, we constructed a corpus of MILES strings from the ExCorr database [12], which contains information on substrate materials, corrosion inhibitor molecules (including SMILES), experimental conditions, and corrosion performance. After quality-control filtering to remove entries with incomplete substrate or alloy information and invalid IE values, 5,330 valid MILES strings were retained. An example of a constructed MILES string is shown below:

```
0.1A/L0ME/L1Al/L2ID"pure"/L3EM"potentiodynamic polarization"/
L4ELCN"HCl"[2.0]CN"(2S;5R;6R)-6-[[(2R)-2-amino-2-phenylacetyl]amino]-3;3-
dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylicacid
(Ampicillin)"SM"CC1(C)S[C@@H]2[C@H](NC(=O)[C@H](N)c3ccccc3)C(=O)N2[C@H]1C
(=O)O"IE88
```

Probing the statistical learnability of MILES

Lossless formats hold the same information, but they need not be equally easy to model. From the 5,815 ExCorr-derived strings, we kept the 5,330 records that had valid substrates, alloys, inhibitors with SMILES, and IE values (if available). These records were written as 1) MILES, 2) JSON with MILES codes as keys, and 3) JSON with descriptive keys. To verify that no information was lost, we converted every JSON record back into MILES, which reproduced the identical string in all cases. A character-level n-gram model (Witten-Bell interpolation, 12-character context) was trained on each format with the same three 90/10 train/test splits, and the incremental LZMA-compressed size of the test records served as an independent check. We found that JSON records were 1.6 – 2.1 times longer, and MILES had the shortest held-out code length with both methods (see **Table SI 3**): the n-gram model encoded 83% and 93% of test records more efficiently in MILES than in the two JSON formats ($p < 10^{-48}$ in each split). MILES also needed the fewest tokens with three tokenizers and compressed best with six compression algorithms. Because the information is identical, these differences arise from the compact MILES syntax, which avoids the keys, quotes and brackets that JSON repeats around

every value. These results establish that the syntax of MILES affects the statistical structure of the encoded strings. However, they do not establish its utility in machine learning. We therefore next investigated whether a language model trained on this representation captures meaningful relationships among materials systems. Specifically, the MILES strings were used to train a transformer-based encoder (seed: 42) using a masked language modelling (MLM) objective. Implementation details are provided in the Methods section. The latent representations generated using the model were subsequently used for downstream analyses, as shown in **Figure 7**.

Latent representation captures substrate- and alloy-dependent structure

We first examined the organization of the latent space at the substrate level (**L1**) using principal component analysis (PCA). Projection of the 96-dimensional latent representations onto the first two principal components showed that PC1 and PC2 accounted for 34.4% and 24.9% of the variance, respectively (see **Figure 7a**, left plot). Although substrate identity was not used as a supervised target, the learned representations exhibited substrate-dependent organisation. Fe-based systems formed a distinct region, with a dense cluster well separated from the other substrates and a smaller group extending toward the centre of the latent space. Mg-based systems occupied a separate region characterised by high PC2 values, with a dense cluster centred at PC1 ≈ -5 and PC2 ≈ 6, and several smaller groups extending around. Al-based systems were concentrated at negative PC1 values and displayed substantial internal structure, forming multiple overlapping clusters across a broad range of PC2 values. Cu-based systems occupied a more localised region within the negative PC1 domain and partially overlapped with Al-based systems. Zn-based systems were comparatively sparse because of their limited representation in the dataset (37 MILES strings) and appeared within the broader Al/Cu region. This qualitative observation is supported quantitatively: the latent space achieved a substrate silhouette score of 0.425. The silhouette score measures, for each string, its cosine similarity to its own substrate cluster versus the nearest substrate cluster, then averages across all strings; values range from -1 to 1, with higher values indicating clearer separation context (see the **SI** for the detailed quantitative analysis). This confirms that the latent representations were meaningfully organised by substrate.

We next performed UMAP analysis (**Figure 7a**, right plot). The UMAP projection provides a complementary view of the same latent space by emphasising local neighbourhood relationships, which makes it a useful probe of the organization captured at the **L2** (alloy) level. Although alloy labels were not provided during training, several alloy classes form compact and well-defined groups, particularly within the Fe- and Mg-based families. Complete separation is not observed across all alloy classes, with several groups exhibiting substantial overlap; nevertheless, the emergence of distinct local clusters indicates that the latent representations largely capture information associated with alloy identity.

Latent representation captures substrate/alloy-dependent inhibitor behaviour

We examined the latent-space positions of identical inhibitors across different alloys to investigate how metallurgical context influences inhibitor representation. We restricted the analysis to inhibitors used on at least three distinct alloys, yielding 1,710 MILES strings from 141 unique inhibitors. Inhibitors were ranked first by the number of distinct substrate families covered, then by alloy count. The two inhibitors with the broadest substrate and alloy coverage are shown in **Figure 7b**. These were 2H-benzotriazole (BTA; 15 alloys; IE: -1372 to 99.9%) and sodium octadecanoate (sodium stearate; 14 alloys; IE: -236 to 97%). Additional inhibitor plots are provided in the **SI**. The mean pairwise centroid dispersion in PCA space is reported in each panel title for reference but was not used for selection. Across plots, the latent-space position of the same compound shifted substantially depending on the substrate/alloy, and these shifts further tracked changes in inhibition performance. Observations associated with corrosive or weak inhibition clustered separately from those associated with strong behaviour (alloys with the highest and lowest IE are annotated). It can be argued that the observed spread reflects random variation; we therefore computed semantic displacement across substrate pairs: the mean of 0.115 (95% Confidence Interval: 0.105 – 0.124) far exceeds the permutation null

mean of 0.006 ($p < 0.001$, based on 1,000 permutations), confirming that these shifts are statistically significant (see the **SI** for full quantitative analysis). This shows that the latent representations capture differences in inhibitor behaviour (**L4**) across substrate/alloy contexts (**L1/L2**), rather than representing the inhibitor as a fixed molecular identity.

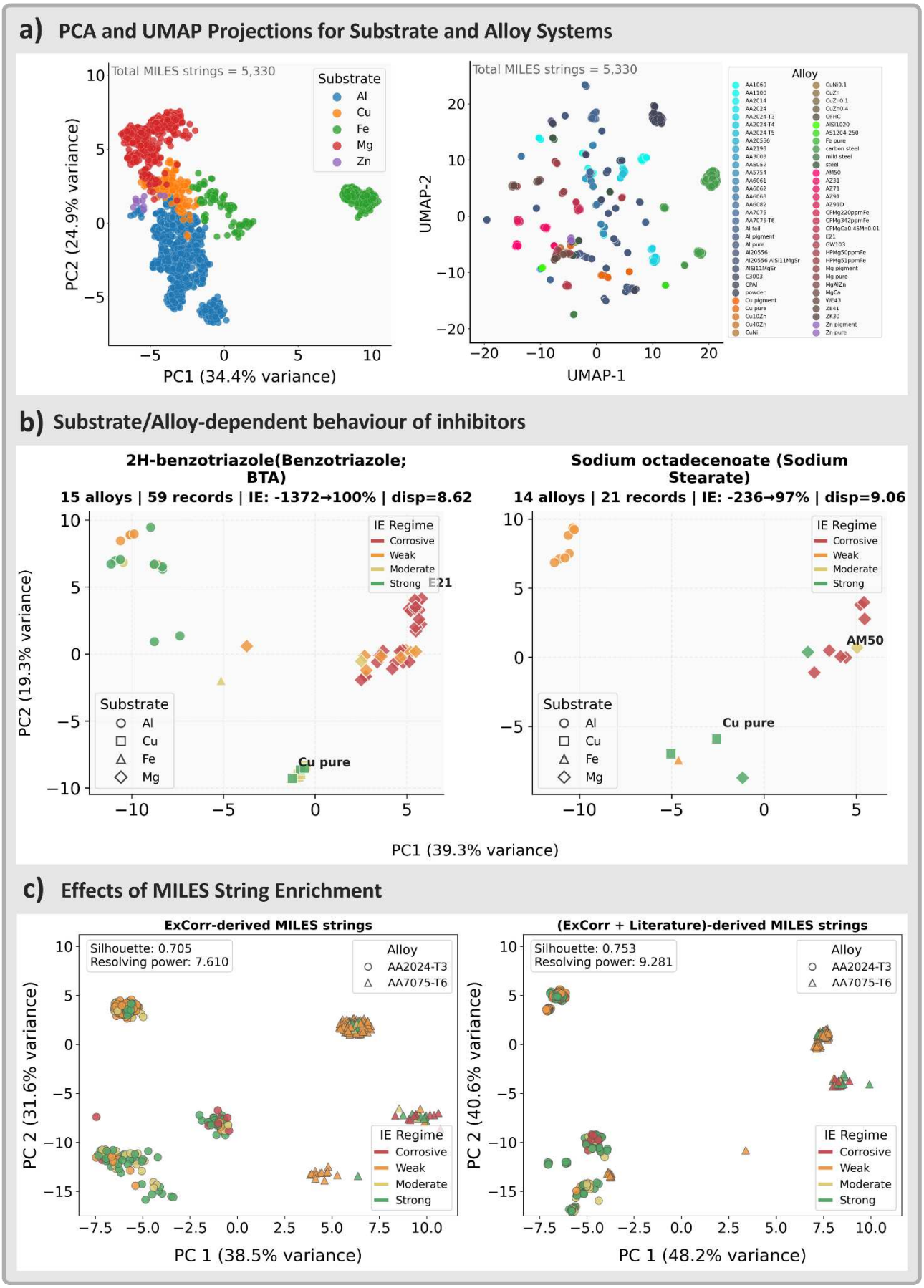


***Figure 7**: Self-supervised learning demonstrates that MILES has a potential to be a materials language with learnable structure. **(a)** PCA and UMAP of latent representations generated by a transformer-based encoder model trained on 5,330 MILES strings with MLM objective. Without explicit supervision, latent representations organise primarily according to substrate chemistry, with major*

substrate classes occupying distinct regions of latent space, indicating that the model captures underlying materials-related information directly from the MILES strings. UMAP of the same latent representations coloured by alloy identity. Closely related alloys cluster together while preserving global relationships, demonstrating that alloy-specific characteristics are encoded within the learned latent representations. ***(b)*** *Projections of inhibitor latent representations illustrate that the latent space captures substrate/alloy-dependent inhibition behaviour. The same inhibitor adopts distinct latent representations across different substrate/alloy combinations, while the resulting clusters are further separated according to IE. The latent representations are projected onto the PCA space derived from the filtered subset. Point shapes denote the substrates and are coloured according to IE regime: corrosive (IE < 0%), weak (0 – 45%), moderate (45 – 75%), and strong inhibitors (≥75 – 100%).* ***(c)*** *Effect of enriching the MILES strings, specifically for the two alloys AA2024-T3 and AA7075-T6, on latent representation quality. The latent representations learned from the ExCorr corpus exhibit weaker separation of alloy regimes than those learned from the ExCorr corpus augmented using literature mining. Enrichment produces more compact and better-resolved clusters, reflected by higher silhouette score and resolving power, indicating that increased chemical context at the* ***L2*** *level and additional tokens at the* ***L3/L4*** *levels enable the MLM to learn more discriminative representations.*

To test whether the alloy-, substrate-, and inhibitor-dependent organizations are learned from MILES rather than imposed by the architecture, we encoded the same strings with an identical encoder but with random weights. We found that training doubles the separation of substrates (silhouette 0.42 vs 0.19) and of alloys (0.58 vs 0.30). We also measured how far an inhibitor moves in latent space when we change its substrate, relative to how far it moves when we replace it by another inhibitor on the same alloy. In the trained encoder this ratio is 15.9, against 2.0 in the untrained one. These differences hold across three trained and five untrained encoders (see **Table SI 4**).

Richer MILES strings improve latent space organization: an ablation study

To investigate how the information content of MILES strings affects model learning, we performed an ablation study comparing the existing model trained on ExCorr-derived MILES strings, which are usually sparse in nature, with a model trained on enriched MILES strings. In total, 661 enriched strings were generated by augmenting the ExCorr-derived MILES strings with additional information extracted from the literature using `MILES-GPT` for the AA2024-T3 and AA7075-T6 alloy systems. The additional information was encoded by new tokens at the **L2**, **L3**, and **L4** levels, where available in the literature. An example of this enrichment is shown below.

MILES String: (ExCorr):

```
0.1A/L0ME/L1Al/L2ID"AA2024-T3"ECA2024T3/L3/L4ELCN"NaCl"[0.05]CN"3H-1;3-
benzothiazole-2-thione(2-mercaptobenzothiazole; MBT)
"SM"C1=CC=C2C(=C1)NC(=S)S2"[0.002]IE99.53
```

MILES String: (ExCorr + Literature Mining):

```
0.1A/L0ME/L1Al/L2ID"AA2024-T3"EWCu4.35Cr0.10Fe0.50Mg1.50Mn0.60
Si0.50Ti0.15Zn0.25Al92.05NCCu4.35Cr0.10Fe0.50Mg1.50Mn0.60Si0.50Ti0.15Zn0.
25Al92.05TC"T3"/L3PO"diamond"EM"potentiodynamic polarization"/L4ELCN
"NaCl"[0.05]CN"3H-1;3-benzothiazole-2-thione(2-mercaptobenzothiazole;
MBT)"SM"C1=CC=C2C(=C1)NC(=S)S2"[0.002]IM48.0SE3.4IE99.53
```

Figure 7c compares PCA projections of the corresponding latent representations. Marker geometry denotes alloy (circles for AA2024-T3; triangles for AA7075-T6), while colour indicates IE regime as before. In the sparse MILES model, the two alloys separate along PC1 (silhouette score = 0.705), but AA2024-T3 representations are distributed across several sub-clusters rather than forming a single compact alloy-specific region. IE organisation is partial: weak cases are concentrated in the upper AA2024-T3 cluster, whereas strong, moderate and corrosive cases are intermixed in the lower clusters. The resolving power is 7.61, and latent displacement between alloy-specific representations shows no significant relationship with changes in IE (Spearman's ρ = +0.170, p = 0.147; n = 74 shared inhibitors). Resolving power is the ratio of between-alloy centroid separation to within-alloy spread, where values above one indicate the model distinguishes the two alloys more than it scatters them internally.

Enriching the MILES strings modestly increases alloy separation by reorganising the latent space. AA2024-T3 and AA7075-T6 representations separate predominantly along PC1, producing well-defined alloy-specific regions with minimal overlap. The silhouette score improved to 0.753 and resolving power to 9.28. However, the IE organization seen in the sparse model is largely retained: strong inhibitors preferentially occupy the lower AA2024-T3 clusters and weak inhibitors the upper cluster, whereas AA7075-T6 is dominated by weak inhibitors (85%), apart from a small cluster (n = 25) of mainly strong and corrosive cases. The correlation between latent displacement and alloy-dependent changes in IE increased numerically with enrichment (Spearman's $\rho = +0.170 \rightarrow +0.211$) but remained statistically non-significant in both models ($p = 0.147$ and $p = 0.071$). Within-alloy dispersion decreased significantly by 18% ($0.263 \rightarrow 0.216$; paired Wilcoxon signed-rank test, $p < 0.001$; 544 of 661 strings closer to their alloy centroid), while the mean per-inhibitor separation between alloy-specific latent representations increased 3.1-fold ($0.628 \rightarrow 1.931$). Together, these results show that richer MILES encoding increases alloy separation and compactness in the latent space. As more tokens are incorporated into the strings, the latent space becomes more compact and separates material environments and performance more effectively.

Using MILES strings for context-aware prediction

A corrosion inhibitor does not have one efficiency; it has one per treated material at fixed experimental conditions. For instance, in our corpus, Benzotriazole ranges from -77 % on high-purity Mg, where it accelerates corrosion, to 100 % on Cu (symmetrised IE averaged per alloy). A representation of the molecule alone, such as its SMILES, assigns the same value to every alloy and cannot express this dependence. MILES writes the material and the experimental conditions into the same string as the molecule. To test whether this makes inhibitor performance learnable across materials, we asked the question a corrosion scientist faces in practice: how will a known inhibitor perform on an alloy on which it has never been tested? We selected all 141 inhibitors tested on at least three alloys (1,123 inhibitor-alloy pairs), held out each pair together with all its strings in a ten-fold split over pairs, repeated three times, and trained a random forest on the remaining corpus (see Method section for implementation details). We gave the same model two inputs: the molecule alone (RDKit descriptors of the SMILES), or the full MILES string (the same descriptors together with the rest of the string read by the MLM encoder, which was pretrained on all strings with the IE values removed). For illustration, we show Benzotriazole, one of the two inhibitors of **Figure 7b**, which is tested on the most substrates (4) and the most alloys (15). Trained on MILES, the model reproduced how its efficiency changes from one alloy to another, including the change of sign from corrosion acceleration on Mg alloys to protection of Al, Fe, and Cu alloys ($\rho = 0.88$; **Figure 8a**). Across all 141 inhibitors, the predicted and measured trends correlated positively for 91 % of inhibitors (median $\rho = 0.67$; **Figure 8b**), and the model identified the better-protected of two alloys in 77 % of cases; a molecule-only representation cannot distinguish alloys (50 % by construction). To put the prediction error in context, we compared it with two simple references on the same pairs. Predicting each pair from the alloy's average over other inhibitors ordered the alloys almost as well (74 %; median $\rho = 0.63$, positive for 89 % of inhibitors) but gave a mean absolute error of 39 IE points, and the model given only the molecule gave 34; with MILES the error fell to 30. MILES thus allowed a single model to combine the inhibitor with its material context: it ordered alloys as well as the alloy average while lowering the prediction error below both references. The largest errors remain at the extremes, where the predictions are compressed towards the mean, and most pairs (86 %) involve Mg alloys, for which systematic multi-alloy data exist.

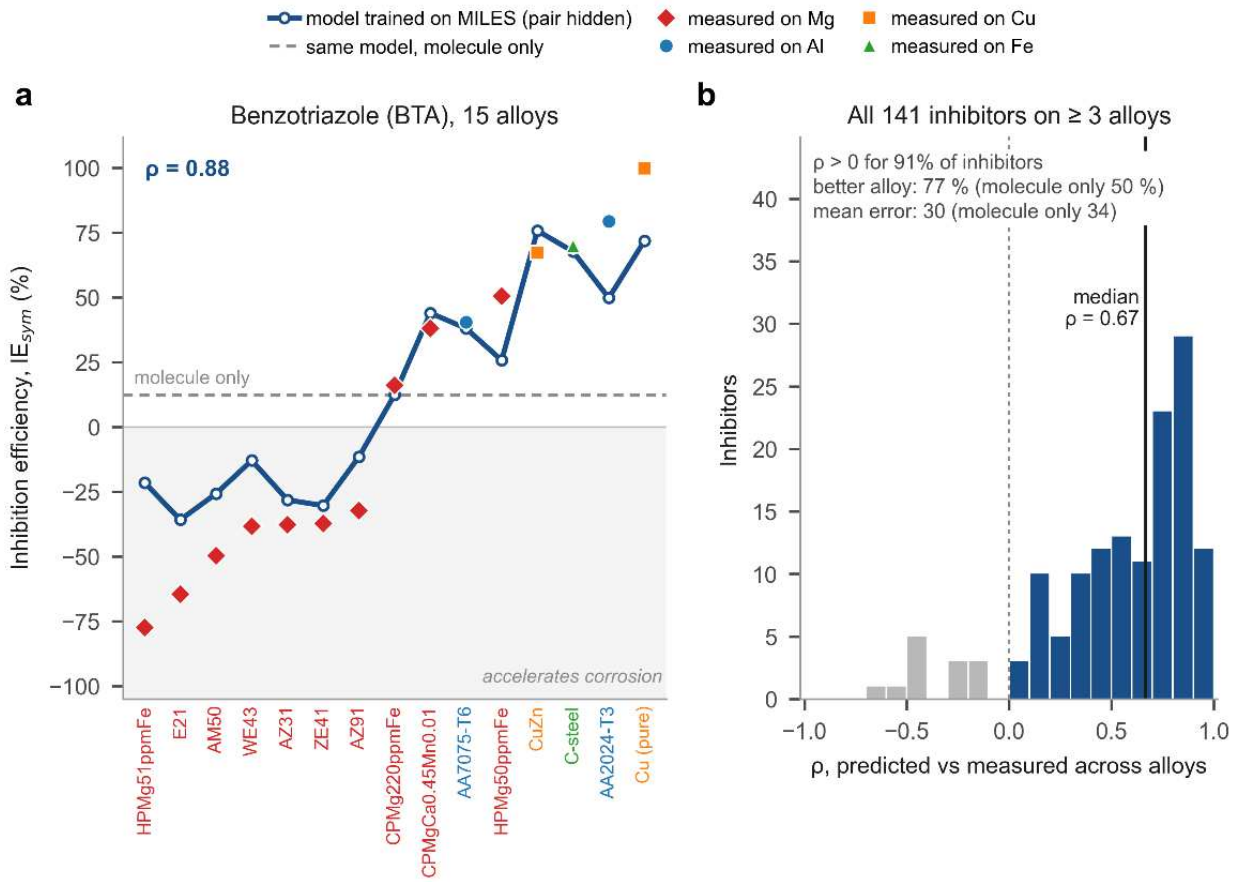


***Figure 8**. A model trained on MILES predicts changes in inhibitor performance across alloys for which it has not seen the data. **(a)** IE of Benzotriazole on each alloy: measured values (symbols, coloured by substrate; alloy names in the same colours) and values predicted by a random forest trained on MILES strings with the respective inhibitor-alloy pair hidden (line). The dashed line is the single value the same model predicts from the molecule alone. The shaded region marks acceleration of corrosion (IE < 0). ρ is the Spearman correlation between predicted and measured IE across alloys. **(b)** Spearman ρ for each of the 141 inhibitors tested on at least three alloys (ρ > 0 in blue); the vertical line marks the median. Predictions are averaged over three repeats of a ten-fold split over inhibitor-alloy pairs. The MILES encoder was trained without IE values. IE is shown in its symmetrised form: IE for IE ≥ 0 and IE/(1 − IE/100) for IE < 0, so that -50 % corresponds to a doubled corrosion rate.*

Discussion

Many complex material and surface properties need to be assessed in “real world” environments. Simple ML models are surprisingly good at modelling and predicting the properties of materials of diverse types in restricted domains, but they are naturally limited in their scope and applicability. The simplifications used result in loss of relevant information about the material and its environment. We showed that MILES strings capture many aspects of materials composition, microstructure, properties and performance across multiple length and time scales. MILES strings encode both the input and output properties as metrics. The coding system is modular, with optional modules, allowing MILES to be extended to new applications in materials science and other disciplines. The MILES tokens can be followed by a <value>, “string” or can be standalone where they define a Boolean. MILES is designed to span more complex multicomponent materials and much larger ranges of time and scale than NInChI. There is no clear rule for where MILES should begin and end, i.e., what information should be encoded within the strings. MILES leaves the order of fields within a level free; since every field starts with a two-letter code, a tokenizer can restore a fixed order before training, which we verified on all the ExCorr-derived MILES strings. MILES is flexible enough that coding can be simple, for instance when capturing information from a paper's experimental section. It could therefore define minimum requirements for reporting certain data types in publications, serving as both quality control and a computer-readable repository for published data (see **Figures 4** and **5**). Lodging this type of information is becoming an

increasingly important requirement for publication. Clearly, the concept of MILES can be adopted widely across many fields. The current MILES tokens could also be extended to include important metadata such as names of instrument manufacturers and details of instrument operating conditions. In this spirit, interfaces and plugins could also be developed to connect MILES with Electronic Lab Notebooks and existing materials databases, allowing available experimental data to be automatically converted into MILES strings. A significant advantage of a system like MILES is that papers could provide an easy-to-read machine format for searching and reporting. Discussions with the community about introducing new MILES tokens and/or levels will be accomplished via the MILES project GitHub.

One limitation of the current version of MILES is the normalization of quantifiable entities. MILES currently assumes that a given quantity is represented using a consistent unit (see the **SI** for the unit associated with each descriptor). However, the same quantity may be reported in the literature or databases using different units. This requires either a normalization step within the information extraction workflow or post-processing using tools such as grobid-quantities [60, 61] and ontologies such as QUDT [62].

For ML modelling of corrosion, we have shown how MILES strings can provide information or features, i.e., a tag for a property, variable, and performance metric. When these MILES tokens are combined into complex MILES text strings, they contain information on the underlying alloy and its microstructure, the electrolytes to which it is exposed, its performance, and details of the experimental conditions at which the properties were measured. Our results demonstrate that this information can be learned by transformer-based encoders, with the resulting latent representations capturing substrate- and alloy-dependent structure and changes in material behaviour across experiment environments. Furthermore, the ablation study provided an effective demonstration of the potential of the MILES concept. The lack of a *significant* improvement in agreement with IE changes may be due to limited sample data (only 74 shared inhibitors), the encoder's capacity, or the need for further enrichment of the MILES strings. These factors remain to be explored in future work. However, it already shows how the ability of language models to learn meaningful representations of materials can be improved by augmenting MILES strings with additional information, motivating the research community to further develop the MILES framework. Beyond learning representations, MILES strings can be used directly for prediction. Given MILES strings instead of the molecule alone, a standard random forest predicted how an inhibitor performance changes across alloys for which it had not seen the data, including the change from corrosion acceleration on Mg alloys to protection of Al, Fe, and Cu alloys. This information is lost when inhibitors are described by their molecular structure alone. These results are a first proof of concept, obtained with sparse strings and a small encoder; we expect the predictive performance to improve as MILES strings are enriched with further information, such as alloy composition and microstructure, and as more data points become available. Collectively, these findings support the central premise of MILES as an extensible materials representation and demonstrate its potential as a materials language. By providing a common language for encoding materials and experimental context, MILES creates a framework for further language-model development, including masked-token prediction and conditional generation of MILES strings, forming the basis of future generative materials design and modelling.

Current advances in ML are rapidly unlocking the full potential of textual representations. LLMs can process complex, information-rich inputs and can identify deep underlying relationships between materials, surface information and performance metrics. LLM-based interfaces such as `MILES-GPT` can be used to extract relevant surface information from multiple sources, such as research papers, and process it into MILES strings, opening the pathway not only for a new form of data storage, but for the generation of truly FAIR data: data that is *Findable*, through defined tokens and metadata; *Accessible*, via machine-readable formats and open interfaces; *Interoperable*, through consistent use of ontologies and shared data logics; and *Reusable*, by ensuring rich contextual information enabling reproducibility and integration across studies. We have demonstrated this using a proof-of-concept employing the MILES encoder. In this way, large amounts of data required for training ML models can be automatically generated from available data sources, thus reusing the results of previous research. Future research will target the automatic extraction of more complex

information, such as EBSD data. By including surface information, models trained on these information-rich descriptors could make more broadly general predictions, e.g., on the corrosion behaviour for different substrates, as opposed to multiple surface-specific or alloy-specific models as is currently the case. Although we have exemplified MILES for inhibitor assessment in corrosion control, the framework offers a new paradigm for encoding, storing, and modelling materials data more broadly.

Methods

MILES Tooling

Naturally, manual conversion of experimental data into a MILES string would be arduous. The variety of surface metrics would render the process rather complicated and compiling the tokens by hand would be prone to errors. For a standardized conversion, digital tools have been created that help with the encoding, decoding, and usage of MILES strings. The digital tools are accessible on GitHub via https://github.com/MILES4Materials. Over time, the tools will be updated with new features, and new tools will become available. Users from the community will also have the chance to implement new tokens for their specific use cases. Further documentation will be shared in the web application and the GitHub repository.

For direct encoding, the surface data can be entered using a web application (https://miles.hereon.de) that converts this information in real-time into a usable MILES string. Dedicated forms are organised according to the layered structure of the MILES string itself. The input data is structured as JSON and sent to a backend where it is processed into the final string format. Analogously, a given MILES string can be decoded to JSON output or an LLM-generated textual summary. The backend uses our developed Python package `milescoder` that provides some basic operations for MILES, including encoding from (JSON to string) and decoding (from string to JSON), and can be accessed via PyPi or the MILES project on GitHub. The user can install the package via `pip` for direct use in their workflows to process MILES strings for their own ML applications. For further information, please refer to the GitHub repository.

Data Mining for Metric Extraction from the literature

To build ML models that can accurately predict across different substrates, surface conditions, and environments, the information-rich nature of MILES strings provides greatly enhanced amounts of data for training. Recent advancements in data mining, specifically rapidly increasing capabilities of LLMs, suggest they may also be useful for generating MILES strings from information in corrosion papers, and more generally in materials science. Several works describe real or potential data mining approaches for the corrosion literature [12, 63] and curated machine-readable inhibitor databases, such as CORDATA [64] and ExCorr [12], have benefited from these efforts. LLMs are increasingly used to extract information automatically from publications generally. Patiny and Godin recovered FAIR (Findable, Accessible, Interoperable, Reusable) data from journal articles, validating extracted molecular properties with cheminformatics tools [65]. Jami et al. built CCU-Llama by pretraining and transfer-learning Llama-2 on carbon capture and utilisation articles retrieved through the Elsevier API, then queried it for technology and impact information [66]. Ansari and Moosavi developed Eunomia, in which an LLM plans and executes extraction tasks using external tools which matched or exceeded fine-tuned methods without task-specific training [67]. To address the multifactorial nature and complex interconnections of scientific content, current research also focuses on multimodal information extraction, retrieving data not only from text but also from figures and tables [68]. Busch et al. benchmarked general-purpose and domain-specific LLMs for predicting magnesium dissolution modulator efficiency, finding GPT-4o outperformed the neural-network baseline model, while a domain-specific chemical LLM showed essentially no correlation with experiment [69].

A potential workflow is to first ingest a research paper, which is usually available in JATS-XML or PDF format, into a standardized machine-readable format, such as TEI [70] or the S2ORC-based JSON format [71]. The process requires an appropriate parser, e.g., GROBID [60] for PDF-to-TEI conversion. In parallel, a Pydantic

model and the corresponding JSON schema can be defined, conforming to the MILES descriptors. The next step is to develop an LLM agent to extract entities defined in the Pydantic-based JSON schema from the machine-readable content of the literature. In this regard, the agent can leverage the structured output mechanism to ensure that the generated output follows the JSON schema. Besides this, the agent requires a suitable prompt that defines a set of rules and possibly incorporates few-shot examples to guide the extraction. For example, the extraction routine must not infer values that are not available in a paper or should carefully disambiguate named entities into an array of MILES strings when multiple data points (sets) are available. Our developed MILES-GPT service conceptually follows the above-described routine. In simple terms, it takes a research paper as input and automatically processes it into one (or multiple) JSON file(s), which are then converted into MILES strings.

Processing research papers in this manner could help to extract MILES-related metrics, e.g., alloy compositions such as those determined by the researcher. Three studies by Laleh et al.[72], Zhang et al.[73] and Peng et al. [74] serve as proofs-of-concept for surface information extraction from research publications and its automatic conversion to MILES strings. For information extraction, we have used sub-routines from the tool `scilitminer` developed by some of the co-authors [75]. It allows different LLMs (both open-source and proprietary) to be used, as well as custom JSON schemas and system prompts. In the following example, we use GPT-4.1 as the underlying LLM. **Figure 9** shows a flowchart for the different ways to encode or decode surface information in MILES. The latest version of the extraction logic is available as `MILES-GPT` (https://miles.hereon.de/miles-gpt) .

One example is provided to showcase the extraction quality. In the study of Zhang *et al.*[73], a binary Mg-Zn magnesium alloy and pure Mg were studied as degradable biomedical implant materials. Processing the respective publication as PDF, results in a list of five MILES strings provided below.

```
0.1A/L0ME/L1Mg/L2ID"Mg6Zn"EWFe0.0038Si0.0016Ni0.0005Cu0.0005Al0.0085Mn0.0
004Zn5.621Mg94.3642NCMg97.5Zn2.5NWMg94.0Zn6.0/L3HT"annealing"TI2.0TE350.0
HAAr999900.0EG18.8YM42.3YS169.5AB"SiC"1200BE8.0EM"direct"ET250.0CAEXWR/L4
CR0.2VO20.0ELCN"NaCl"[0.116]CN"CaCl2"[0.0018]CN"KCl"[0.0054]CN"MgSO4"[0.0
01]CN"NaHCO3"[0.0262]CN"Na2HPO4"[0.00089]CN"NaH2PO4"[0.00021]PH7.44PH9.32
VA20.0ET37.0IM72.0SE1.0

0.1A/L0ME/L1Mg/L2ID"Mg6Zn"EWFe0.0038Si0.0016Ni0.0005Cu0.0005Al0.0085Mn0.0
004Zn5.621Mg94.3642NCMg97.5Zn2.5NWMg94.0Zn6.0/L3HT"annealing"TI2.0TE350.0
HAAr999900.0EG18.8YM42.3YS169.5AB"SiC"1200BE8.0EM"direct"ET250.0CAEXWR/L4
CR0.07VO20.0ELCN"NaCl"[0.116]CN"CaCl2"[0.0018]CN"KCl"[0.0054]CN"MgSO4"[0.
001]CN"NaHCO3"[0.0262]CN"Na2HPO4"[0.00089]CN"NaH2PO4"[0.00021]PH7.44PH9.3
2VA20.0ET37.0IM720.0SE1.0

0.1A/L0ME/L1Mg/L2ID"Mg6Zn"EWFe0.0038Si0.0016Ni0.0005Cu0.0005Al0.0085Mn0.0
004Zn5.621Mg94.3642NCMg97.5Zn2.5NWMg94.0Zn6.0/L3HT"annealing"TI2.0TE350.0
HAAr999900.0EG18.8YM42.3YS169.5AB"SiC"1200BE8.0EM"direct"ET250.0CAEXWR/L4
CR2.32IM1344.0IV"rabbit"IW"bone"

0.1A/L0ME/L1Mg/L2ID"Mg"EWMg99.99NCMg100.0NWMg100.0/L3AB"SiC"1200/L4CR0.43
VO20.0ELCN"NaCl"[0.116]CN"CaCl2"[0.0018]CN"KCl"[0.0054]CN"MgSO4"[0.001]CN
"NaHCO3"[0.0262]CN"Na2HPO4"[0.00089]CN"NaH2PO4"[0.00021]PH7.44PH9.22VA20.
0ET37.0IM72.0SE1.0

0.1A/L0ME/L1Mg/L2ID"Mg"EWMg99.99NCMg100.0NWMg100.0/L3AB"SiC"1200/L4CR0.1V
O20.0ELCN"NaCl"[0.116]CN"CaCl2"[0.0018]CN"KCl"[0.0054]CN"MgSO4"[0.001]CN"
NaHCO3"[0.0262]CN"Na2HPO4"[0.00089]CN"NaH2PO4"[0.00021]PH7.44PH9.22VA20.0
ET37.0IM720.0SE1.0
```

The extracted information was split into multiple datasets, differentiating between *in vivo* (MILES string three above) and *in vitro* testing, for Mg6Zn alloy and pure Mg (strings four and five). The alloy elemental composition was correctly identified, and the Mg "balance" value was correctly resolved to the corresponding weight percent in the composition. Nominal composition is correctly presented using at.% and wt.%. The electrolyte temperature, initial pH and volume to surface area ratio were correctly identified. The pH value after the *in vitro* tests was captured correctly when it was directly stated in the text (MILES string two, above) but misinterpreted when only an intermediate value was described in the text (MILES string one, above). The electrolyte composition was correctly converted to mol/L from the g/L given in the original paper. Mechanical properties were correctly extracted from tables together with other specifics concerning the manufacturing process. Corrosion metrics were also extracted, including correct information regarding the corrosion rate (CR) result from several testing conditions (*in vitro* vs *in vivo*, electrochemical measurement vs immersion test and immersion time 3 vs 30 days).

Similar levels of success arise for other studies. For example, Laleh et al. [72], studying the grain boundary character distribution in an austenitic stainless steel, found that the extracted data was correctly split into as-built and annealed conditions. However, processing of very complex grain information, e.g., grain interfaces, graphical results etc., could not be fully captured and demands further development. A third study by Peng et al. [74] compared the inhibition performance of two novel rare-earth complexes with the well-researched corrosion inhibitor lanthanum 4-hydroxycinnamate, La(4-OHcin)$_3$, on AS1020 mild steel. `MILES-GPT` was able to correctly identify results for blank and inhibited conditions together with the respective corrosion inhibition performance. Other extracted information is shown as MILES strings in the **SI**. It is noteworthy that in all studies, reasonable information could be extracted and converted. The extraction logic correctly identified alloy and electrolyte compositions, different datapoints, and most experimental conditions and properties from text and tables.

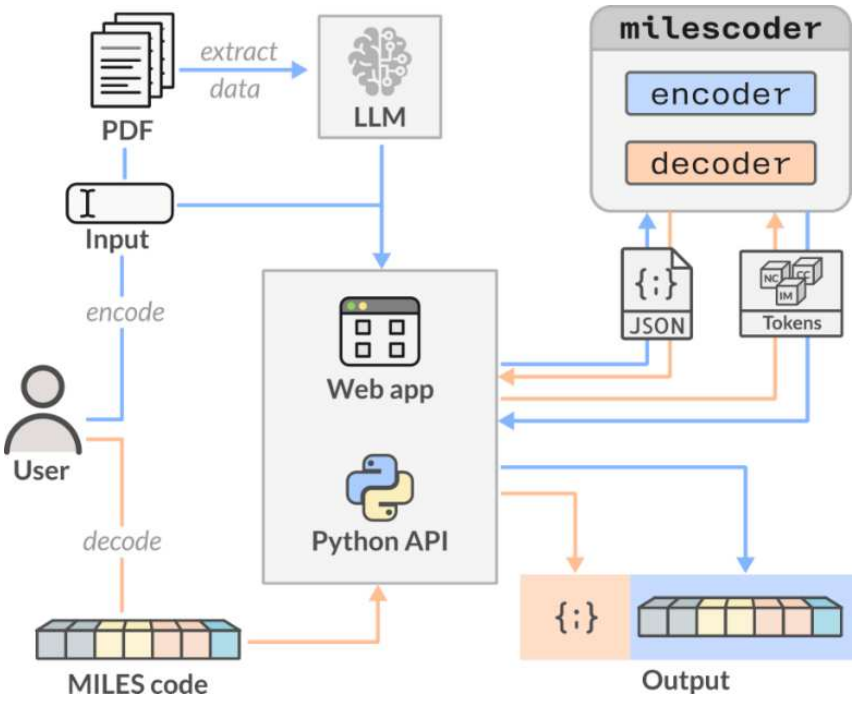


***Figure 9**: Flowchart for encoding and decoding surface information using the MILES digital tools. For encoding and decoding, input data can be directly processed via the web app or Python API. Input data can also be generated by extracting relevant information from PDFs using the LLM-service MILES-GPT.*

This feature will be helpful for building meaningful training corpora for ML models. Although MILES-GPT is still under development, and current limitations exist, such as consecutive extraction runs generating slightly different MILES strings and maximum output token limit, its capabilities already allow relevant data to be collected from research papers and subsequently used for training ML models. The maximum output limit of 32,768 tokens in GPT-4.1 resulted in truncated data. However, this limitation is largely addressed in the newer models that allow substantially higher output tokens, e.g., GPT-5.2 offers 128,000 tokens. One possible

practical solution to mitigate this effect could also be limiting the to-be-extracted metrics to a domain-relevant, user-defined selection, which we plan to implement in the future. Other limitations include extracting multi-modal data; some inconsistency in extracted results; and generation of error statistics. The initial implementation is limited to textual and tabular information. Furthermore, the LLM and manual input of MILES tokens make no explicit assumption about the nature of the statistics of the performance metric. It is expected that this will, most commonly, be the average of performance data, but, equally, it could be an individual performance datum. The statistics of the measurement using higher moments of a sample distribution such as the standard deviation are not currently employed but are under consideration by the authors, as is consideration of the errors in the measurements. Despite these limitations, MILES-GPT already enables the automatic extraction of MILES-related metrics, which will significantly ease the use and adaptation of the MILES format. In the future, it shall be possible to also extract relevant information from figures, where certain graph types, e.g. EBSD data, are automatically identified as such and processed accordingly. In this study, we have assessed the literature mining-constructed MILES strings qualitatively; the underlying information extraction framework was evaluated quantitatively in previous work [76], at both the unique-experiment-set and field levels, in the creep domain. This framework is a multiagent system that enables ontology-guided structuring and reflexive valuation for consistent outputs and effective disambiguation of multiple datasets [76].

Masked Language Modelling for MILES

To exemplify MILES, we used the ExCorr database, a curated collection of corrosion inhibition experimental data, such as inhibitor chemistry, substrate materials, experimental conditions, and corrosion inhibition performance [12]. For example, a record in the database describes the use of carboxymethyl cellulose (IUPAC: acetic acid;2;3;4;5;6-pentahydroxyhexanal) as a corrosion inhibitor for pure Al in a 0.5 M H_2SO_4 corrosive medium. The corrosion IE was reported as 40% at an operating concentration of 0.1 mM using the potentiodynamic polarisation method. The molecular structure of the inhibitor was represented by the SMILES string of the repeating unit CC(=O)O.O=CC(O)C(O)C(O)C(O)CO. Additional metadata included the publication source.

Each ExCorr entry was converted into a MILES string that encodes both molecular and experimental information. To ensure data quality, MILES strings were filtered using a set of validation rules. Specifically, only entries containing valid substrate and alloy identifiers were retained. Entries with unknown substrates or unknown alloys were removed. Additionally, IE values were validated, retaining only entries with IE values less than or equal to 100%. After applying these filtering criteria, a total of 5,330 valid MILES strings remained for model training.

We tokenised these MILES strings with a structure-aware scheme that keeps the main parts of each record intact. The tokenizer first separates hierarchy markers, quoted text, and bracketed blocks, then splits the remaining segments into meaningful units, such as alloy codes, inhibitor names, concentrations, and experimental conditions. This preserves the experimental structure without forcing a rigid handwritten grammar. Each string was converted into a token sequence and mapped to a fixed vocabulary with reserved padding, masking, and unknown tokens. Sequences were limited to 64 tokens, sufficient to cover the full-length distribution while keeping the input compact. The representation therefore preserves both local content and broader experimental context. The encoder model was implemented as a lightweight transformer architecture with learned token and positional embeddings of dimension 96. The model consisted of two transformer encoder layers with four self-attention heads per layer and a feed-forward network dimension of 192. A dropout rate of 0.1 was applied throughout the network to improve generalization. We trained the model using an MLM objective, in which 15% of tokens were randomly selected for masking. MLM is a self-supervised training objective where a model hides a random sample of input tokens and learns to predict the original tokens using bidirectional context. Optimization was performed using AdamW with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-2}. Training was conducted with a batch size of 32 for up to 150 epochs, and 10% of the dataset was held out for validation. The learning rate schedule employed 200 warmup steps followed by cosine decay. Early stopping, with a patience of 15 epochs,

was used to prevent overfitting and select the best-performing model. During training, padding tokens were masked from the attention mechanism so that they did not influence the learned latent representations. After training, the best-performing encoder was used to generate a 96-dimensional latent representation for each of the 5,330 MILES. The Transformer produces token-level representations, which were mean-pooled across non-padding and non-mask positions to obtain a single representation for each MILES.

For downstream analysis, the trained encoder was used to generate a latent representation

$$z \in \mathbb{R}^{96}$$

for each MILES string. Prior to dimensionality reduction, latent representations were standardised using feature-wise z-score normalization,

$$\tilde{z} = \frac{z - \mu}{\sigma}$$

where μ and σ denote the mean and standard deviation computed for each latent dimension across all MILES strings. The standardised latent representations were subsequently analysed using PCA and UMAP.

Formally, given an input token sequence $x = (x_1, \ldots, x_n)$, a subset of tokens M was randomly selected for masking. The model was trained to predict the original identities of the masked tokens by minimising the cross-entropy loss

$$\mathcal{L}_{\mathcal{MLM}} = -\frac{1}{|M|}\sum_{i \in M} \log P\left(x_i \mid x_{\setminus M}\right)$$

where $x_{\setminus M}$ denotes the partially masked sequence and $P\left(x_i \mid x_{\setminus M}\right)$ is the predicted probability assigned to the original token at position i.

Property Prediction using MILES

To exclude IE information from the input, the IE token was removed from all 5,330 strings and the encoder was retrained with the protocol described above. Token representations were mean-pooled within each MILES level (L1 – L4) and concatenated (384 dimensions). A random forest regressor (scikit-learn; 300 trees, $\sqrt{p}$ features per split; where p is the number of input features, minimum two samples per leaf, n_jobs = -1, random_state = 0) was trained on two inputs: the molecule alone (217 RDKit 2D descriptors and a 1,024-bit Morgan count fingerprint, radius 2) or the same molecular features combined with the MILES representation. We selected the 141 inhibitors tested on at least three alloys (1,710 MILES strings, 1,123 inhibitor-alloy pairs). Each inhibitor-alloy pair was held out with all its strings, including strings of the same molecule on that alloy listed under another name, and the model was trained on the remaining strings with a measured IE (ten-fold split over pairs, three repetitions). As a reference, each held-out pair was also predicted by the mean IE of its alloy in the training data. Performance was evaluated using the mean absolute error of the symmetrised IE, the Spearman correlation ρ across the alloys of each inhibitor, and the fraction of alloy pairs ($|\Delta IE| \geq 10$) for which the better-protected alloy was correctly identified.

Code Availability

The code repositories for the web application (https://miles.hereon.de), the digital tools are available under the GitHub organization MILES4Materials (https://github.com/MILES4Materials) and the source code distributed as the Python package (https://pypi.org/project/milescoder). The data that is used in this study is openly available at https://excorr.web.app/database.

Acknowledgements

The authors would like to acknowledge the IT infrastructure and resources provided by the Helmholtz-Zentrum Hereon. Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number

535656357, OPTIMA (Identification of optimal corrosion inhibitors for bare and PEO coated magnesium alloy by combining machine learning and robotic testing). This work was supported by the Digital Twins of Complex Systems (DToCS) Platform at Helmholtz-Zentrum Hereon, which is jointly funded by the German Federal Government and the state governments of Schleswig-Holstein, Hamburg, Lower Saxony, and Brandenburg.

The authors used Claude Opus 5 (Anthropic) to assist with coding, to shorten certain sections of the text, and to improve the readability of the manuscript. The authors revised all edits and take full responsibility for the content of this publication.

Author Contributions

Conceptualization - lead: AEH, JMCM, VG, TW; equal: SVL, CF, DAW, ML, MLZ.
Methodology - lead: AEH, JMCM, VG, TW; equal: SVL, CF, DAW, ML, MLZ.
Software - lead: VG, TW; support: DAW, CF.
Validation - lead: AEH, VG, TW, SVL, MLZ; equal: JMCM, CF, DAW, ML.
Data curation - lead: VG, TW, AEH.
Visualization - lead: TW, VG, AEH; support: SVL.
Writing - original draft: lead: AEH, VG, TW; support: SVL, DAW.
Writing - review & editing - all coauthors.
Funding acquisition - SVL, MLZ.

Competing Interests

The authors declare no competing interests.

References

[1] C.L. Mellor, K.E. Tollefsen, C. LaLone, M.T.D. Cronin, J.W. Firman, In Silico Identification of Chemicals Capable of Binding to the Ecdysone Receptor, Environmental Toxicology and Chemistry, 39 (2020) 1438-1450. https://doi.org/10.1002/etc.4733

[2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature, 596 (2021) 583-589. https://doi.org/10.1038/s41586-021-03819-2

[3] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, Machine learning assisted design of high entropy alloys with desired property, Acta Materialia, 170 (2019) 109-117.https://doi.org/10.1016/j.actamat.2019.03.010

[4] C. Zeng, A. Neils, J. Lesko, N. Post, Machine learning accelerated discovery of corrosion-resistant high-entropy alloys, Computational Materials Science, 237 (2024) 112925.https://doi.org/10.1016/j.commatsci.2024.112925

[5] H.C. Ozdemir, A. Nazarahari, B. Yilmaz, D. Canadinc, E. Bedir, R. Yilmaz, U. Unal, H.J. Maier, Machine learning – informed development of high entropy alloys with enhanced corrosion resistance, Electrochimica Acta, 476 (2024) 143722.https://doi.org/10.1016/j.electacta.2023.143722

[6] J.M. Rickman, H.M. Chan, M.P. Harmer, J.A. Smeltzer, C.J. Marvel, A. Roy, G. Balasubramanian, Materials informatics for the screening of multi-principal elements and high-entropy alloys, Nature Communications, 10 (2019). https://doi.org/10.1038/s41467-019-10533-1

[7] K. Hippalgaonkar, Q. Li, X. Wang, J.W. Fisher, J. Kirkpatrick, T. Buonassisi, Knowledge-integrated machine learning for materials: lessons from gameplaying and robotics, Nature Reviews Materials, 8 (2023) 241-260. https://doi.org/10.1038/s41578-022-00513-1

[8] M. Seifrid, S. Lo, D.G. Choi, G. Tom, M.L. Le, K. Li, R. Sankar, H.T. Vuong, H. Wakidi, A. Yi, Z. Zhu, N. Schopp, A. Peng, B.R. Luginbuhl, T.Q. Nguyen, A. Aspuru-Guzik, Beyond molecular structure: critically

assessing machine learning for designing organic photovoltaic materials and devices, Journal of Materials Chemistry A, 12 (2024) 14540-14558. https://doi.org/10.1039/d4ta01942c

[9] J. Messina, R. Luo, K. Xu, G. Lu, H. Deng, M.A. Tschopp, F. Gao, Machine learning to predict aluminum segregation to magnesium grain boundaries, Scripta Materialia, 204 (2021). https://doi.org/10.1016/j.scriptamat.2021.114150

[10] L. Xu, P. Yu, S. Wen, Y. Tang, Y. Wang, Y. Tian, T. Mao, C. Li, Visualization and Analysis of Oil and Gas Pipeline Corrosion Research: A Bibliometric Data-Mining Approach, Journal of Pipeline Systems Engineering and Practice, 15 (2024). https://doi.org/10.1061/JPSEA2.PSENG-1605

[11] S. Ma, Y. Du, S. Wang, Y. Su, Application of machine learning in material corrosion research, Corrosion Reviews, 41 (2023) 417-426. https://doi.org/10.1515/corrrev-2022-0089

[12] 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, Progress in Materials Science, 149 (2025). https://doi.org/10.1016/j.pmatsci.2024.101392

[13] 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?, Corrosion Science, 179 (2021) 108856.https://doi.org/10.1016/j.corsci.2020.108856

[14] 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, Corrosion Science, 193 (2021) 109900. https://doi.org/10.1016/j.corsci.2021.109900

[15] C. Feiler, D. Mei, B. Vaghefinazari, T. Würger, R.H. Meißner, B.J.C. Luthringer-Feyerabend, D.A. Winkler, M.L. Zheludkevich, S.V. Lamaka, In silico screening of modulators of magnesium dissolution, Corrosion Science, 163 (2020). https://doi.org/10.1016/j.corsci.2019.108245

[16] 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 Chemistry, 16 (2014) 3349-3357. https://doi.org/10.1039/C3GC42540A

[17] L.B. Coelho, D. Zhang, Y. Van Ingelgem, D. Steckelmacher, A. Nowé, H. Terryn, Reviewing machine learning of corrosion prediction in a data-oriented perspective, npj Materials Degradation, 6 (2022) 8. https://doi.org/10.1038/s41529-022-00218-4

[18] L. Yan, Y. Diao, Z. Lang, K. Gao, Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach, Science and Technology of Advanced Materials, 21 (2020) 359-370. https://doi.org/10.1080/14686996.2020.1746196

[19] Y. Diao, L. Yan, K. Gao, Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features, Materials & Design, 198 (2021) 109326.https://doi.org/10.1016/j.matdes.2020.109326

[20] M. Liu, W. Li, Prediction and analysis of corrosion rate of 3C steel using interpretable machine learning methods, Materials Today Communications, 35 (2023). https://doi.org/10.1016/j.mtcomm.2023.106408

[21] A. Rojas Z, S. Bakhtiari, K. Wang, V.M. Calo, M. Iannuzzi, Revisiting the Copson Curve Using Data Science, Journal of the Electrochemical Society, 170 (2023). https://doi.org/10.1149/1945-7111/acd7ab

[22] A.R. Murthy, K. Lakshmi, S. Vishnuvardhan, M. Saravanan, Prediction of SIF range for plain API 5L Grade X65 steel under corrosion using AI & ML models, Materials Today Communications, 36 (2023) 106543.https://doi.org/10.1016/j.mtcomm.2023.106543

[23] Y. Wang, B. Goh, M. Moorehead, J. Hattrick-Simpers, A. Couet, High-Throughput Electrochemistry to Study Materials Degradation in Extreme Environments, Analytical Chemistry, 94 (2022) 16528-16537. https://doi.org/10.1021/acs.analchem.2c03325

[24] Y. Zhi, D. Fu, D. Zhang, T. Yang, X. Li, Prediction and Knowledge Mining of Outdoor Atmospheric Corrosion Rates of Low Alloy Steels Based on the Random Forests Approach, in: Metals, 2019. https://doi.org/10.3390/met9030383

[25] O.C. Olawole, D.K. De, O.F. Olawole, R. Lamba, E.S. Joel, S.O. Oyedepo, A.A. Ajayi, O.A. Adegbite, F.I. Ezema, S. Naghdi, T.D. Olawole, O.O. Obembe, K.O. Oguniran, Progress in the experimental and

948 computational methods of work function evaluation of materials: A review, Heliyon, 8 (2022).
949 https://doi.org/10.1016/j.heliyon.2022.e11030
950 [26] D. Rajendran, T. Sasilatha, S. Rajendran, S.K. Selvaraj, C. Lacnjevac, S.S. Prabha, R.J. Rathish, Application
951 of machine learning in corrosion inhibition study, Materials Protection, 63 (2022) 280-290.
952 https://doi.org/10.5937/zasmat2203280R
953 [27] K. Haruna, S.I. Abba, J. Usman, A.G. Usman, A. Musa, T.A. Saleh, I.H. Aljundi, Machine learning insight
954 into inhibition efficiency modelling based on modified graphene oxide of diaminohexane (DAH-GO) and
955 diaminooctane (DAO-GO), Carbon Trends, 16 (2024) 100373.https://doi.org/10.1016/j.cartre.2024.100373
956 [28] J. Du, J. Zheng, Y. Liang, N. Xu, Q. Liao, B. Wang, H. Zhang, Deeppipe: Theory-guided prediction method
957 based automatic machine learning for maximum pitting corrosion depth of oil and gas pipeline, Chemical
958 Engineering Science, 278 (2023). https://doi.org/10.1016/j.ces.2023.118927
959 [29] F.R. Elmas, M.P. Rios, E.R. Almeida Lima, R.G. Gusmão Caiado, R.S. Santos, Prediction of external
960 corrosion rate in Oil and Gas platforms using ensemble learning: a Maintenance 4.0 approach, Brazilian
961 Journal of Operations and Production Management, 20 (2023). https://doi.org/10.14488/BJOPM.1952.2023
962 [30] V.C. Anadebe, V.I. Chukwuike, C.C. Okoye, L. Guo, R.C. Barik, Artificial intelligence and machine learning
963 in corrosion research, in: Handbook of Research on Corrosion Sciences and Engineering, 2023, pp. 1-23.
964 [31] G. Grethe, G. Blanke, H. Kraut, J.M. Goodman, International chemical identifier for reactions (RInChI),
965 Journal of Cheminformatics, 10 (2018) 22. https://doi.org/10.1186/s13321-018-0277-8
966 [32] X. Yan, T. Yue, D.A. Winkler, Y. Yin, H. Zhu, G. Jiang, B. Yan, Converting Nanotoxicity Data to Information
967 Using Artificial Intelligence and Simulation, Chemical Reviews, 123 (2023) 8575-8637.
968 https://doi.org/10.1021/acs.chemrev.3c00070
969 [33] C. Ren, L. Ma, D. Zhang, X. Li, A. Mol, High-throughput experimental techniques for corrosion research:
970 A review, Materials Genome Engineering Advances, 1 (2023) e20.https://doi.org/10.1002/mgea.20
971 [34] S.V. Lamaka, B. Vaghefinazari, D. Mei, R.P. Petrauskas, D. Höche, M.L. Zheludkevich, Comprehensive
972 screening of Mg corrosion inhibitors, Corrosion Science, 128 (2017) 224-
973 240.https://doi.org/10.1016/j.corsci.2017.07.011
974 [35] T. Würger, L. Wang, D. Snihirova, M. Deng, S. Lamaka, D.A. Winkler, D. Höche, M. Zheludkevich, R.H.
975 Meißner, C. Feiler, Data-driven Selection of Electrolyte Additives for Aqueous Magnesium Batteries, Journal
976 of Materials Chemistry A, (2022).https://doi.org/10.1039/D2TA04538A
977 [36] T. Würger, D. Mei, B. Vaghefinazari, D.A. Winkler, S.V. Lamaka, M.L. Zheludkevich, R.H. Meißner, C.
978 Feiler, Exploring structure-property relationships in magnesium dissolution modulators, npj Materials
979 Degradation, 5 (2021) 2.https://doi.org/10.1038/s41529-020-00148-z
980 [37] T. Wurger, C. Feiler, F. Musil, G.B.V. Feldbauer, D. Hoche, S.V. Lamaka, M.L. Zheludkevich, R.H. Meissner,
981 Data Science Based Mg Corrosion Engineering, Frontiers in Materials, 6 (2019) 53.ARTN 53.
982 https://doi.org/10.3389/fmats.2019.00053
983 [38] E.J. Schiessler, T. Würger, S.V. Lamaka, R.H. Meißner, C.J. Cyron, M.L. Zheludkevich, C. Feiler, R.C. Aydin,
984 Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning
985 models, npj Computational Materials, 7 (2021). https://doi.org/10.1038/s41524-021-00658-7
986 [39] C. Feiler, D. Mei, B.J.C. Luthringer-Feyerabend, S.V. Lamaka, M.L. Zheludkevich, Rational Design of
987 Effective Mg Degradation Modulators, Corrosion, 77 (2021) 204-208. https://doi.org/10.5006/3597
988 [40] E.J. Schiessler, T. Würger, B. Vaghefinazari, S.V. Lamaka, R.H. Meißner, C.J. Cyron, M.L. Zheludkevich, C.
989 Feiler, R.C. Aydin, Searching the chemical space for effective magnesium dissolution modulators: a deep
990 learning approach using sparse features, npj Materials Degradation, 7 (2023)
991 74.https://doi.org/10.1038/s41529-023-00391-0
992 [41] The Statue of Liberty on Liberty Island. Photograph by Carol M. Highsmith, 2017. Library of Congress,
993 Prints & Photographs Division, Carol M. Highsmith Archive, Public domain.
994 [42] Astronaut photograph ISS026-E-6255 (30 Nov 2010, Nikon D3S, 180 mm), NASA Johnson Space Center /
995 ISS Crew Earth Observations. Cropped and contrast-enhanced by NASA Earth Observatory. Public domain.

[43] A.J. Medford, M.R. Kunz, S.M. Ewing, T. Borders, R. Fushimi, Extracting Knowledge from Data through Catalysis Informatics, ACS Catalysis, 8 (2018) 7403-7429. https://doi.org/10.1021/acscatal.8b01708

[44] J.A. Esterhuizen, B.R. Goldsmith, S. Linic, Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys, Chem, 6 (2020) 3100-3117. https://doi.org/10.1016/j.chempr.2020.09.001

[45] N. Birbilis, Y.M. Zhu, S.K. Kairy, M.A. Glenn, J.F. Nie, A.J. Morton, Y. Gonzalez-Garcia, H. Terryn, J.M.C. Mol, A.E. Hughes, A closer look at constituent induced localised corrosion in Al-Cu-Mg alloys, Corrosion Science, 113 (2016) 160-171. https://doi.org/10.1016/j.corsci.2016.10.018

[46] M. Gao, C. Feng, R.P. Wei, An analytical electron microscopy study of constituent particles in commercial 7075-T6 and 2024-T3 alloys, Metallurgical and Materials Transactions A, 29 (1998) 1145-1151,

[47] A.E. Hughes, C. MacRae, N. Wilson, A. Torpy, T.H. Muster, A.M. Glenn, Sheet AA2024-T3: A new investigation of microstructure and composition, Surface and Interface Analysis, 42 (2010) 334-338. https://doi.org/10.1002/sia.3163

[48] Y. Chen, Y. Zhang, H. Zhou, C.T. Leung, H. Gao, Use of Machine Learning and Large Language Models in Chemical Information Extraction, Annual Review of Chemical and Biomolecular Engineering, 17 (2026) 417-441. https://doi.org/10.1146/annurev-chembioeng-100724-080433

[49] M.C. Ramos, C.J. Collison, A.D. White, A review of large language models and autonomous agents in chemistry, Chemical Science, 16 (2025) 2514-2572. https://doi.org/10.1039/d4sc03921a

[50] M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M.V. Gil, S. Miret, C.T. Koch, J.A. Márquez, K.M. Jablonka, From text to insight: large language models for chemical data extraction, Chemical Society Reviews, 54 (2025) 1125-1150. https://doi.org/10.1039/d4cs00913d

[51] C.W. Kosonocky, A.L. Feller, C.O. Wilke, A.D. Ellington, Using alternative SMILES representations to identify novel functional analogues in chemical similarity vector searches, Patterns, 4 (2023) 10.1016/j.patter.2023.100865

[52] C. Qian, H. Tang, Z. Yang, H. Liang, Y. Liu, Can large language models empower molecular property prediction?, arXiv preprint arXiv:2307.07443, (2023).

[53] J.M.Z. Chaves, E. Wang, T. Tu, E.D. Vaishnav, B. Lee, S.S. Mahdavi, C. Semturs, D. Fleet, V. Natarajan, S. Azizi, Tx-LLM: A Large Language Model for Therapeutics, arXiv preprint arXiv:2406.06316, (2024),

[54] S. Sadeghi, A. Bui, A. Forooghi, J. Lu, A. Ngom, Can large language models understand molecules?, BMC bioinformatics, 25 (2024) 225. https://doi.org/10.1186/s12859-024-05847-x

[55] N.T. Runcie, C.M. Deane, F. Imrie, Assessing the Chemical Intelligence of Large Language Models, Journal of Chemical Information and Modeling, 66 (2025) 216-227. https://doi.org/10.1021/acs.jcim.5c02145

[56] H. Qiu, Z.-Y. Sun, Exploring SMILES and BigSMILES: Toward Efficient and Scalable Polymer Representations for Machine Learning, Macromolecules, 58 (2025) 7084-7093. https://doi.org/10.1021/acs.macromol.5c00604

[57] J.M. Cavanagh, K. Sun, A. Gritsevskiy, D. Bagni, Y. Wang, T.D. Bannister, T. Head-Gordon, SmileyLlama: modifying large language models for directed chemical space exploration, Nature Computational Science, 6 (2026) 856-869. https://doi.org/10.1038/s43588-026-00986-y

[58] R. Jacobs, M.P. Polak, L.E. Schultz, H. Mahdavi, V. Honavar, D. Morgan, Regression with Large Language Models for Materials and Molecular Property Prediction, arXiv preprint arXiv:2409.06080, (2024)

[59] Q. Zhang, K. Ding, T. Lyv, X. Wang, Q. Yin, Y. Zhang, J. Yu, Y. Wang, X. Li, Z. Xiang, Scientific large language models: A survey on biological & chemical domains, arXiv preprint arXiv:2401.14656, (2024)

[60] P. Lopez, GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications, in: M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas (Eds.) Research and Advanced Technology for Digital Libraries, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 473-474

[61] GROBID-quantities (2015-2025) <https://github.com/kermitt2/grobid-quantities>, accessed 29.10.2026

[62] FAIRsharing.org:, QUDT; Quantities, Units, Dimensions and Types , DOI: 10.25504/FAIRsharing.d3pqw7, Last Edited: Thursday, March 26th 2026, 10:27, Last Editor: Lea.Girard,Last Accessed: Thursday, October 29th 2026, 11:45,Last Reviewed:Tuesday, March 25th 2025, 10:11. 10.25504/FAIRsharing.d3pqw7

[63] S. Zhao, N. Birbilis, Searching for chromate replacements using natural language processing and machine learning algorithms, npj Materials Degradation, 7 (2023) 2. https://doi.org/10.1038/s41529-022-00319-0

[64] T.L.P. Galvão, I. Ferreira, A. Kuznetsova, G. Novell-Leruth, C. Song, C. Feiler, S.V. Lamaka, C. Rocha, F. Maia, M.L. Zheludkevich, J.R.B. Gomes, J. Tedim, CORDATA: an open data management web application to select corrosion inhibitors, npj Materials Degradation, 6 (2022) 48. https://doi.org/10.1038/s41529-022-00259-9

[65] L. Patiny, G. Godin, Automatic extraction of FAIR data from publications using LLM, ChemRxiv, 2023 (2023). doi: 10.26434/chemrxiv-2023-05v1b-v2

[66] H.C. Jami, P.R. Singh, A. Kumar, B.R. Bakshi, M. Ramteke, H. Kodamana, CCU-Llama: A Knowledge Extraction LLM for Carbon Capture and Utilization by Mining Scientific Literature Data, Industrial & Engineering Chemistry Research, 63 (2024) 17585-17598. https://doi.org/10.1021/acs.iecr.4c01656

[67] M. Ansari, S.M. Moosavi, Agent-based learning of materials datasets from the scientific literature, Digital Discovery, 3 (2024) 2607-2617. https://doi.org/10.1039/d4dd00252k

[68] G. Khalighinejad, S. Scott, O. Liu, K.L. Anderson, R. Stureborg, A. Tyagi, B. Dhingra, MatViX: Multimodal Information Extraction from Visually Rich Articles, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 3636-3655. https://doi.org/10.18653/v1/2025.naacl-long.185

[69] 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, Corrosion Science, 255 (2025) 113080. https://doi.org/10.1016/j.corsci.2025.113080

[70] T. Consortium, TEI P5: Guidelines for Electronic Text Encoding and Interchange. https://www.tei-c.org/guidelines/p5/

[71] K.W. Lo, Lucy Lu; Neumann, Mark; Kinney, Rodney; Weld, Daniel S2ORC: The Semantic Scholar Open Research Corpus, in: D.C. Jurafsky, Joyce; Schlut, Natalie; Tetreault, Joel (Ed.) 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4969-4983. https://doi.org/10.18653/v1/2020.acl-main.447

[72] M. Laleh, A.E. Hughes, M.Y. Tan, G.S. Rohrer, S. Primig, N. Haghdadi, Grain boundary character distribution in an additively manufactured austenitic stainless steel, Scripta Materialia, 192 (2021) 115-119. https://doi.org/10.1016/j.scriptamat.2020.10.018

[73] S. Zhang, X. Zhang, C. Zhao, J. Li, Y. Song, C. Xie, H. Tao, Y. Zhang, Y. He, Y. Jiang, Y. Bian, Research on an Mg–Zn alloy as a degradable biomaterial, Acta Biomaterialia, 6 (2010) 626-640. https://doi.org/10.1016/j.actbio.2009.06.028

[74] Y. Peng, A.E. Hughes, G.B. Deacon, P.C. Junk, B.R.W. Hinton, M. Forsyth, J.I. Mardel, A.E. Somers, A study of rare-earth 3-(4-methylbenzoyl)-propanoate compounds as corrosion inhibitors for AS1020 mild steel in NaCl solutions, Corrosion Science, 145 (2018) 199-211.https://doi.org/10.1016/j.corsci.2018.09.022

[75] V. Gupta, J.D.H. Paul, I. Schmitt, F. Pyczak, SciLitMiner: An Intelligent System for Scientific Literature Mining and Knowledge Discovery, Advanced Intelligent Systems, 8 (2026) e202501235. https://doi.org/10.1002/aisy.202501235

[76] V. Gupta, A holistic system for scientific literature mining and knowledge discovery: application to advanced intermetallic titanium aluminide alloys. PhD Thesis, 2026. https://doi.org/10.26127/BTUOpen-7490

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MILESSI01Nov2026.docx