

Supplementary Information

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

V. Gupta^{1,†}, T. Würger^{2,†}, J.M.C. Mol³, S.V. Lamaka^{2*}, C. Feiler^{2,4,5}, D. A. Winkler^{6,7,8}, M. Laleh⁹, 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.

Common and Distinguishing features of surface reactions in Catalysis and Corrosion

Comparison of surface aspects that are common and different for catalysis and corrosion is shown in **Table SI 1**. These are based on catalysis metrics developed by Medford et al. ¹ to describe catalytic surfaces. Both areas of research involve the interaction of surfaces with environment. In catalysis these interactions can lead to surface restructuring while in corrosion they lead to irreversible surface changes due to corrosion reactions.

Comparison Between Corrosion and Catalysis

Table SI 1: Comparison between corrosion and catalysis.

Catalysis	Corrosion
Interaction of a molecule with a surface – a local surface phenomenon, i.e., bulk properties not directly relatable.	Interaction of an electrolyte with a surface. Bulk properties may be important for propagation, but surface structure is responsible for initiation and can be locally influenced by bulk composition and microstructure.
“Featurisation challenge” – difficult to identify salient features that control fundamental catalytic behaviour since it involves active sites that have a structure that may change under reaction conditions at the atomic and nanoscales.	“Featurisation challenge” is different – surface reactions are rapid, but there is a switch at early stages from random chemical/electrochemical reaction to electrochemically driven highly active sites. These are challenging to featurise as it is not clear how these start – maybe stochastic.

Catalysts aim to maximise surface area using nanoparticles ² . They try to maximise the number of reactive sites, often by increasing the surface area.	Typically, corrosion consumes the surface, so the aim is to minimise surface activity and maximise passivation.
Multiscale phenomena in <i>time</i> and <i>scale</i> .	Similar
Activity highly dependent on local surface composition and structure. Local structure may depend on presence of the adsorbed reactant that creates a feedback loop that effectively acts as a process on the surface.	Initial corrosion activity is local (nanoscale) but cedes to electrochemical reactions on the micro- or even macro-scale with time.
Quantum mechanical approaches can be applied but require knowledge of the atomic structure – this must be obtained from experiments and makes emergent structures difficult to model.	Quantum mechanical approaches have been applied to highly local adsorption onto metal clusters representing termination and localisation of a larger surface. The metal clusters represent low index planes and do not facilitate relevant temporal changes due to corrosion progress.
The “process” aspect of the surface is maintained by a quasi-equilibrium structure which facilitates the reaction of reactants to products. The catalyst is not consumed.	The process is complex to model since it can vary from simple adsorption to oxidation, dissolution, and formation of products that may quench surface reactions, not eliminate them.
A small number of active sites may have a big impact on activity, e.g., the B5 site on Ru nanoparticles has 10,000 times the reactivity of the Ru single crystal surface ³ . The number of these sites is maximised on 5 nm hexagons plates.	Probably similar localised attack but at micro-macro-scale, since a small number of active sites can lead to a fast failure.

MILES Tokens (Two Letter Codes) and Their Levels

MILES tokens are the building blocks of a MILES string and are expressed as two capital letters followed by an “argument” which can be null (i.e., nothing follows), a value expressed as $\langle \square \rangle$ (where the angle brackets are omitted), a string expressed as “ $\square$ ” (with quotes), or something more complex depending on the descriptor. Tokens corresponding to lists are marked with empty squared brackets as $\langle \square \rangle []$. In the instance where there is no argument then the token acts more like a binary code or Boolean; if its present, then it is included as a substring as presented in **Figure 4** in the main paper, otherwise it is not included. These types of tokens can be used when distinguishing extruded from cast material for example. In the former case the MILES string would have an EX and, in the latter, it would be CA thus introducing a differentiator on the method of manufacture. A series of numbers, e.g., FTIR measurements given as 2938, 1650 and 890, are coded by repeating tokens, such as FT2938FT1650FT890.

Table SI 2 lists all currently implemented MILES tokens and indicates to which level the tokens are attached. Furthermore, there is information regarding the data type, unit (if relevant) and token definition. Occasionally, tokens are paired such as, heat treatment time and temperature in $TI \langle \square \rangle TE \langle \square \rangle$, where the first token has a unit of hours and the second has a unit of degree centigrade. For reference, the currently implemented levels are as follows: **L0** = material, **L1** = metal, **L2** = alloy, **L3** = conventional manufacturing, **L4** = electrolyte, **L5** = additive manufacturing. However, implementation of further levels for other use cases is likely and will be developed in the community. To track future developments of MILES, a version number is included at the beginning of a MILES string. The format is based on the InChI standard, where the version number (e.g., 1) is followed by the version type (e.g., the standard version is indicated by “S”). Newer versions should always be backwards-compatible to ensure that older versions are still usable. Analogous to the NInChI project, we use an alpha version (“A”) for the first development period to facilitate interaction with the community. The alpha version starts at “0.1A”.

Table SI 2: List of all currently implemented MILES tokens, along with their assigned levels, data types, units (if given) and definitions.

Metric	Level	Type(s)	Unit	Definition
AB"type"<grade>	L3	string number	– –	Information about abrasion process including tool type and abrasion grade. Here "type" can be one of the following: "Al", "SiC", "Al ₂ O ₃ ", "acidic SiC" or "diamond". A grade of 400 grit is written as 0400, 1200 grit is 1200. The finer grade follows the coarser one, building up the grinding sequence.
AN	L4	boolean	–	Indicator that the experiment was performed under anaerobic conditions.
BE<ratio>	L3	number	–	The extrusion ratio for extrusion applications (cross section ratio of billet/extrusion).
BL<length>	L2	number	m/cm ²	Mean grain boundary length.
CA	L3	boolean	–	Indicator that the material has been cast.
CC	L4	boolean	–	Indicator for a crevice. There could be metrics for crevice configuration, crevice depth. No tokens to describe details have yet been developed.
CD	L3	boolean	–	Indicates that a material has been manufactured using cold drawing. No tokens to input further processing details have been developed.
CL<thickness>	L3	number	µm	Clad layer thickness in microns. Relevant for galvanised products but can also be used with other clad systems. It can be followed by an elemental composition (refer to EC).
CN<#>	L2	number	–	Number of phases in a material.
CN"name"	L4	string	–	Chemical compound name.
CR<rate>	L4	number	mm/year	Corrosion rate, expressed in mm/year as per ASTM G31-72(2004).
CP"phase"<element, content>[]	L2	string string number	– – %	Crystal phase of any phase, e.g., BCC or FCC, present in a major quantity in the material, along with the according composition, as obtained from EBSD.
C<direction> CO<percent>(<miller>)	L2	string number number[]	– % –	Crystal orientation as obtained from phase maps using EBSD. <direction> is used to indicate IPFX, IPFY or IPFZ, and can be X, Y or Z. The orientation is defined as a list of percentages of orientation and their according Miller indices.
CY	L4	boolean	–	Indicator that the environment is cyclic.
DA<distance>	L3	number	µm	Dendrite arm spacing.
DA<angle>	L4	number	°	Droplet contact angle in degrees.
DD<density>	L4	number	#/dm ²	Droplet density in number per square decimetre.
DE<thickness>	L3	number	µm	Deformed layer thickness (<value>) in microns. It can be followed by an elemental composition (refer to EC) for the deformed layer.
DI<element, content>	L2	string content	– at.%	Nominal composition of present dispersoid particles. Similar to EC.
DR	L3	boolean	–	Indicator that a material has been drawn, but the history is not known (i.e., whether it is hot or cold drawn).

DS<size>	L4	number	mm	Droplet size.
EC<element, content>[]	L2 L3	string number	– at.%	Elemental composition expressed as a series of elements followed by the content in at.% e.g. Cu20.0Yy2.0Fe1.3. The element/number junction is the delimiter. It can be used to describe also layered materials, e.g., in conjunction with clad layers.
ED<time>	L4	number	minutes	Total duration of the experiment.
EG<value>	L3	number	%	Breaking elongation.
ELCN" name"SM"smiles"[<molarity>]	L4	string string number	– – mol/L	Electrolyte composition (<i>EL</i>) as a combination of compound names (<i>CN</i>), SMILES strings (<i>SM</i>) and concentrations given in molarity. Note: <i>IT</i> <medium> can be used for in-vitro electrolytes.
ES<speed>	L3	number	m/s	Extrusion speed.
ET<temperature>	L4	number	°C	Corrosion temperature.
EM"mode"	L3	string	–	Extrusion mode. Options are <i>direct</i> , <i>backward</i> , and <i>equichannel</i> .
ET<temperature>	L3	number	°C	Extrusion temperature.
ER<time>	L4	number	hours	Electrolyte renewal frequency, often used in biometal applications, instead of flow-through electrolyte cells.
EW<element, content>[]	L2 L3	string number	– wt.%	Identical to <i>EC</i> but in weight percent.
EX	L3	boolean	–	Indicator that the material has been extruded.
FD<density>	L4	number	mm ⁻¹	Site density of filiform corrosion filaments expressed as number per millimetre.
FI	L4	boolean	–	Indicator for filiform corrosion. There are other metrics which specify more detailed aspects of filiform corrosion.
FL<length>	L4	number	mm	Average filament length.
FO	L3	boolean	–	Indicator that the material has been forged.
FT<wavenumber>	L2	number	cm ⁻¹	Fourier transform infrared measurements; multiple peak wavenumbers can be stringed together.
GE<rate> CD<depth>	L3	number number	µm/hr µm	Generalised corrosion (<i>GE</i>) over a broad area which has a corrosion rate (<i>CR</i>) and a corrosion depth (<i>CD</i>). Reserved for situations where there is little localised corrosion but extensive generalised corrosion.
GA<value>	L2	number	–	Grain averaged map (GAM) information providing an indication of residual strain accumulation across a grain.
GD<value>	L2	number	#/m ²	Geometrically necessary dislocations defined as number per square metre.
GP<element, content>[]	L2	string number	– at.%	Grain boundary phase decorating the grain boundaries, including the phase stoichiometry / composition.
GS<size>	L2	number	µm	Average grain size.

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.
GZ<value>	L2	number	nm	Grain boundary depletion zone.
HA<element, concentration>	L3	string number	– ppm	Composition of the heat treatment atmosphere.
HM<hardness>	L3	number	N/mm ²	Hardness value from micro-indenter. Multiple hardness values can be stringed together.
HN<hardness>	L3	number	N/mm ²	Hardness value from nano-indenter. Multiple hardness values can be stringed together.
HP<element, content>[][]	L2	string number	– at.%	Hardening precipitates as an array of compositions.
HS<length>	L3	number	mm	Separation length between hardness measurements; should be provided directly after the respective hardness measurements.
HT"\"type" Tl<time> TE<temperature>	L3	string number number	– hours °C	Heat treatment type (<i>HT</i>), time (<i>Tl</i>) and temperature (<i>TE</i>). Available heat treatment types are <i>annealing</i> , <i>quenching</i> , <i>tempering</i> and <i>aging</i> . Time-temperature pairs can be cycled over different time and temperature regimes to generate a temperature-time profile, e.g. T11TE40T12TE50.
HV<hardness>	L3	number	N/mm ²	Vickers hardness measurement.
IB	L3	boolean	–	Indicator that ion beam milling was used at the end of polishing.
IB<potential>	L4	number	V vs. RE	Breakdown potential from potentiodynamic polarisation. Should be combined with the corresponding reference electrode (see <i>RE</i>).
IC<value>	L2	number	–	Intermetallic clustering normally obtained from the peak of the radial distribution function using centroid positions for all particles
IC<current>	L4	number	µA/cm ²	Corrosion current at OCP.
IE<efficiency>	L4	number	%	Inhibition efficiency expressed as a percentage. Calculated from the formula $IE\% = 100 * (I_{corr} - I_{inh}) / I_{corr}$.
IM<element, content>[][]	L2	string number	– at.%	Intermetallic composition as an array of compositions.
IM<time>	L4	number	hours	Immersion time.
IP<value>	L4	number	–	Inhibition power, defined as $IP = 10 * \log_{10}(I_{corr}/I_{inh})$.
IS<efficiency>			%	Symmetrized inhibition efficiency, defined as $IE_{sym} = IE$ if $IE \geq 0$ otherwise $IE_{sym} = -100 * (I_{inh} - I_{corr}) / I_{inh}$
IT<medium>	L4	string	–	In vitro test medium. <medium> is identical to the electrolyte composition <i>EL</i> , defined as a list of compound names, SMILES strings and molarities.

IV<animal>	L4	string	–	Animal model; currently implemented options are: <i>rat, mouse, rabbit, guinea pig, dog, cat, non-human primate, and human.</i>
IW<implantation site>	L4	string	–	Implantation site; currently implemented options are: <i>joint replacement, spine, small bones joints, cranio maxillofacial, dental, arteries, heart, pacemaker, defibrillator, tendon, ligament, subcutaneous, muscle, ureteral, bladder, gastrointestinal, nerves, brain</i>
KA<value>	L2	number	%	Kernel averaged map (KAM) information as obtained from EBSD that reflects the level of dislocations locally, providing an indication of residual strain; defined as the percentage of the largest misorientation angle of 6° of all misorientations 6° and under.
LA<value>	L2	number	%	Percentage of low-angle grain boundaries (<15°) as obtained from EBSD.
LD<percent>	L5	number	%	Percent of part density measured using Archimedes principle compared to full density in laser powder bed fusion (LPBF) additive manufacturing (AM).
LG<element>	L5	string	–	Major cover gas in laser powder bed fusion (LPBF) additive manufacturing (AM).
LH<spacing>	L5	number	µm	Hatch spacing for the scanning pattern of the laser in laser powder bed fusion (LPBF) additive manufacturing (AM).
LL<size>	L5	number	µm	Laser spot size in laser powder bed fusion (LPBF) additive manufacturing (AM).
LP<power>	L5	number	Watts	Laser power in laser powder bed fusion (LPBF) additive manufacturing (AM).
LS<speed>	L5	number	mm/s	Laser speed in laser powder bed fusion (LPBF) additive manufacturing (AM).
LT<thickness>	L5	number	µm	Powder bed layer thickness in laser powder bed fusion (LPBF) additive manufacturing (AM).
LX<value>	L5	string	–	Scanning strategy for the laser in laser powder bed fusion (LPBF) additive manufacturing (AM). Current options are <i>zigzag, meander, contour, raster, and spiral.</i>
LW	L5	boolean	–	Indicator that WAAM (wire arc additive manufacturing) was used.
ME	L3	boolean	–	Indicator that the sample has experienced mechanical stress.
ME	L2	boolean	–	Presence of mischmetal
NC<element, content>[]	L2	string number	– at.%	Nominal composition is coded the same way as EC.
OC<element, content>[]	L3	string number	– at.%	Oxide composition (identical to EC).
OX<thickness>	L3	number	µm	Oxide thickness.

PA	L3	boolean	–	Indicator that an organic coating is present.
PT<thickness>	L3	number	μm	Organic coating thickness.
PB<ratio>	L5	number	–	Pilling-Bedworth ratio defined as the ratio of the density of the oxide volume compared to the metal.
PD<value>	L5	number	μm	Mean powder size distribution in laser powder bed fusion (LPBF) additive manufacturing (AM).
PH<value>	L4	number	–	pH value of the electrolyte used for a experiment.
PO" type" <grade>	L3	string number	– –	Information about polishing process including type of grinding compound and polishing grade. Here "type" can be one of the following: "Al", "SiC", "Al ₂ O ₃ ", "acidic SiC" or "diamond". A grade of 400 grit is written as 0400, 1200 grit is 1200. The finer grade follows the coarser one, building up the grinding sequence.
PI PD<depth> PW<width> PV<volume> PE<value> PR<density>	L4	boolean number number number number number	– μm μm μm ³ μm #/cm ²	Pitting corrosion (<i>PI</i>), including the pit depth (<i>PD</i>), pit width (<i>PW</i>), pit volume (<i>PV</i>), any extreme value information (<i>PE</i>) that might be available from extreme values statistics, and the pitting density (<i>PR</i>) defined as the number of pits per square centimetre.
PT<temperature>	L5	number	°C	Preheat temperature of the powder bed prior to heating in laser powder bed fusion (LPBF) additive manufacturing (AM).
RA<wave number>	L2	number	cm ⁻¹	Raman spectroscopy measurement (similar to <i>FT</i>); multiple peak wavenumbers can be stringed together.
RE"electrode"	L4	string	–	Reference electrode for electrochemical measurements.
RH<percent>	L4	number	%	Relative humidity.
RO NP<#> NG<thickness>	L3	boolean number number	– – mm	Indicator for an as-rolled surface with a folded oxide with embedded metal and a subsurface deformed zone. Other information is the number of rolling passes (<i>NP</i>) and the gauge reduction thickness (<i>NG</i>) designating the thickness of the sheet.
SA<value>	L4	number	°	Angle for washing a corroding surface. The higher the angle, the greater the degree of washing; relevant for ASTM B117 neutral salt spray.
SC	L4	boolean	–	Indicates that stress corrosion cracking is the form of degradation.
SE<value>	L4	number	cm ²	Surface area of working electrode.
SF<value>	L2		–	Schmid Factor as obtained from EBSD analysis, varying from 0.0 to 0.5. Maximum slip occurs when the slip plane is at 45° to the tensile stress.
SH	L3		–	Indicator that shingling is present. This occurs in a number of manufacturing processes where friction between a forming surface (e.g. roller) and the manufactured surface causes folding of surface oxides in the surface layer.
SM<SMILES>[<molarity>]	L4	string number	– g/mol	SMILES string, e.g., for the components of the electrolyte followed by the molarity.
SP	L3	boolean	–	Indicator that the surface has been shot-peened.
SS<stress>	L3	number	N/m ²	Shear Stress.

ST<percent>	L2	number	%	Percentage of $\Sigma 3$ grain boundaries as obtained from EBSD analysis.
TC<code>	L2	number	–	Temper code for wrought Al alloys (with “T” omitted)
VA<ratio>	L4	number	ml/cm ²	The ratio of electrolyte volume to sample surface area.
VO<value>	L4	number	cm ³	Volume of the electrolyte for immersion experiments.
WR	L3	boolean	–	Indicator that a material has been wrought.
YM<modulus>	L3	number	GPa	Young’s Modulus.
YS<strength>	L3	number	GPa	Yield Strength.
ZB<size>	L4	number	mm	Maximum blister size for blistering of the layer after 1000 hours of exposure in salt spray tests. Blistering extending to 5% or more of the surface area or blisters >10 mm in diameter are considered unacceptable.
ZC<creep>	L4	number	mm	Edge after 1000 hours of exposure in salt spray tests.
ZI<time>	L3	number	s	Immersion time in plating bath.
ZK<bar>	L3	number	bar	Air knife pressure for galvanised layer removal.
ZM<mass>	L4	number	g/m ² /day	Mass loss after 1000 hours of exposure in salt spray tests.
ZR<time>	L4	number	hours	Time at which the first red rust formation covers around 5% of the surface.
ZT<temperature>	L3	number	°C	Plating bath temperature of sacrificial metallic anode.

61

62

63 MILES String Examples

64 The first MILES coding example of a wrought aluminium alloy AA2024-T3 is presented in the main text of this
65 paper. The second example is corrosion of 316 stainless steel. It has been abraded using silicon carbide to
66 2000 grit, then polished with diamond to 1 micron but left with a surface deformed layer 1.2 μm thick, half
67 of which is oxide with a composition of $\text{Fe}_{40}\text{O}_{60}$ (Fe_2O_3). Its bulk microstructure from EBSD is a mixture of
68 grains parallel to the rolling direction. It has MnS and Al-Ca oxide as inclusions at 1% and 0.2% respectively.
69 The grain size is 20 μm and the length of the grain boundary is 36 metres/cm², with grain boundary
70 precipitates CrC and a grain boundary depletion zone of 2nm. It is exposed to a 0.1M NaCl electrolyte
71 containing 0.01M Na₂SO₄ for 24 hours. Finally, pitting was observed, with a pit mouth size of 20 μm and depth
72 of 5 μm . No inhibitors were used in this case. Note here that the token *IM* is used both in **L2** (intermetallics)
73 and **L4** (immersion time) but have different meanings in each level. This example translates to

74 0.1A/**L0**ME/**L1**Fe/**L2**ID"316"IMMn50.0Ss50.0Al0.0001Ca0.0001BL36.0GPCrCcGZ2.0GS
75 20.0/**L3**AB"SiC"2000DE1.2OCFe40.00o60.0OX0.6PO"diamond"/**L4**PIPD5.0PW20.0ELCN
76 "NaCl"[0.1]CN"Na2SO4"[0.01]IM24.0

77 The MILES strings for the third example were constructed from an earlier publication by Liu *et al.*⁴.
78 Degradation of Mg-alloy WE43 tubes was studied in a simulated body fluid. Its elemental composition was
79 defined by the authors in the previous publication⁵. The content of noble impurities (Fe, Cu, Ni and Si), critical
80 for corrosion performance of Mg alloys, was not reported in these papers and hence not reflected in the
81 MILES string. Precipitates, Mg₂₄Y₅ and Mg₄₁Nd₅, are uniformly distributed throughout the WE43 alloy. After

82 extrusion and several rounds of cold rolling, the tubes with wall thickness 180-micron exhibit highly deformed
83 grains with various proportions of twins and secondary particles. When tubes are annealed at 450 °C for 1 h,
84 finer microstructures appear, with equiaxed, recrystallized grain sizes of 15.0 µm and EBSD deduced Schmid
85 Factor of 0.29. Reported yield strength and breaking elongation were 113 MPa and 10% correspondingly.
86 Corrosion rate was 0.85mm/y, measured after 48 hours of exposure at 37°C, in HBSS electrolyte of disclosed
87 composition, renewed every 24 hours and volume to surface area ratio 40 ml/cm². The MILES string for the
88 example above would resolve to:

```
89 0.1A/L0ME/L1Mg/L2ID"WE43"ECMg98.3Yy1.1Gd0.15Zr0.14Nd0.4IMMg24.0Yy5.0IMMg4  
90 1.0Nd5.0SF0.29GS15.0/L3HT"annealing"TI1.0TE450.0EG10.0YS0.113BE20.0ET350.  
91 0DREXFORONG0.18/L4CR0.85ELCN"NaCl"[0.137]CN"KCl"[0.0053]CN"CaCl2"[0.00126  
92 ]CN"NaHCO3"[0.00417]CN"MgCl2"[0.00049]CN"MgSO4"[0.00041]CN"KH2PO4"[0.0004  
93 4]CN"Na2HPO4"[0.000168]CN"C6H12O6"SM"C([C@H]1[C@H]([C@H]([C@H](C(O1)O)O  
94 )O)O)O"[0.0055]ER24.0VA40.0ET37.0IM48.0
```

95 The MILES string can accommodate the final microstructure without including how that microstructure was
96 obtained, although it is possible to include some aspects of the manufacture history. The reason for this is
97 that the final microstructure is the one exposed to the electrolyte. Similar MILES strings were also extracted
98 using [MILES-GPT](#) for AZ31 and JDBM alloys, studied along with WE43⁴, see section below MILES-GPT: Proof
99 of Concept.

100

101 MILES-GPT: Proof of concept

102 The study by Laleh et al.⁶ deals with grain boundary character distribution in an austenitic stainless steel
103 produced by additive manufacturing, under as-built and annealed conditions. After processing the PDF file,
104 MILES-GPT returned MILES strings are

```
105  
106 0.1A/L0ME/L1Fe/L2ID"316L"CP"FCC"FeCrNiMoCPST30.6/L5LGArlEArLH80.0LP150.0L  
107 S400.0LT30.0LX"meander"PT200.0
```

```
108  
109 0.1A/L0ME/L1Fe/L2ID"316L"CP"FCC"FeCrNiMoCPST30.4/L3HT"annealing"TI0.166TE  
110 1000.0/L5LGArlEArLH80.0LP150.0LS400.0LT30.0LX"meander"PT200.0
```

111
112 The study by Peng et al.⁷ compares two novel rare-earth based complexes with the well-researched corrosion
113 inhibitor lanthanum 4-hydroxycinnamate La(4-OHcin)₃ for AS1020 mild steel. The inhibition information is
114 given in the following MILES strings, accordingly (*Note: due to the limited context window size, the extracted
115 output had to be truncated*):

```
116  
117 0.1A/L0ME/L1Fe/L2ID"AS1020"NCCc0.26Mn0.45Cr0.14Si0.19Al0.18Fe98.78/L3PO"S  
118 iC"1200/L4VO150.0ELCN"NaCl"[0.01]PH6.0PH6.5IM0.5IC2.74SE0.785
```

```
119  
120 0.1A/L0ME/L1Fe/L2ID"AS1020"NCCc0.26Mn0.45Cr0.14Si0.19Al0.18Fe98.78/L3PO"S  
121 iC"1200/L4VO150.0ELCN"NaCl"[0.01]PH6.0PH6.5IM24.0IC2.79SE0.785
```

```
122  
123 0.1A/L0ME/L1Fe/L2ID"AS1020"NCCc0.26Mn0.45Cr0.14Si0.19Al0.18Fe98.78/L3PO"S  
124 iC"1200/L4VO150.0ELCN"NaCl"[0.01]CN"Lambp3"[0.00025]PH6.0PH6.5IE82.0IM0.5  
125 IC0.5SE0.785
```

```
126  
127 0.1A/L0ME/L1Fe/L2ID"AS1020"NCCc0.26Mn0.45Cr0.14Si0.19Al0.18Fe98.78/L3PO"S  
128 iC"1200/L4VO150.0ELCN"NaCl"[0.01]CN"Lambp3"[0.00025]PH6.0PH6.5IE82.1IM24.  
129 0IC0.5SE0.785
```

```
130  
131 0.1A/L0ME/L1Fe/L2ID"AS1020"NCCc0.26Mn0.45Cr0.14Si0.19Al0.18Fe98.78/L3PO"S
```

132 iC"1200/**L4**VO150.0ELCN"NaCl"[0.01]CN"La4-OHcin3"[0.00025]PH6.0PH6.5IE93.0I
133 M0.5IC0.2SE0.785
134
135 0.1A/**L0**ME/**L1**Fe/**L2**ID"AS1020"NCCc0.26Mn0.45Cr0.14Si0.19Al0.18Fe98.78/**L3**PO"S
136 iC"1200/**L4**VO150.0ELCN"NaCl"[0.01]CN"Ymbp3"[0.00025]PH6.0PH6.5IE97.0IM0.5I
137 C0.09SE0.785

138 [MILES-GPT](#) extracted the following strings for AZ31 and JDBM alloys, that were studied along with WE43 by
139 Liu et al.⁴ exemplified in the section above.

140 0.1A/**L0**ME/**L1**Mg/**L2**ID"AZ31"NCMg96.5Al3.0ZnMn0.5GS12.9/**L3**HT"annealing"TI1.0T
141 E450.0EG16.0YS172.0BE20.0ET350.0CDDREXWR/**L4**CR16.4 ELCN"NaCl"[0.137]
142 CN"KCl"[0.0053]CN"CaCl2"[0.00126]CN"NaHCO3"[0.00417]CN"MgCl2"[0.00049]CN"
143 MgSO4"[0.00041]CN"KH2PO4"[0.00044]CN"Na2HPO4"[0.000168]CN"C6H12O6"SM"C([C
144 @H]1[C@H]([C@H]([C@H](C(O1)O)O)O)O)O"[0.0055]ER24.0VA40.0ET37.0IM48.0
145 ELER24.0VA40.0ET37.0IM48.0

146 0.1A/**L0**ME/**L1**Mg/**L2**ID"JDBM"NCMg94.62Nd2.5ZnZr1.88GS10.9/**L3**HT"annealing"TI1.
147 0TE450.0EG26.0YS123.0BE20.0ET350.0CDDREXWR/**L4**CR1.44 ELCN"NaCl"[0.137]
148 CN"KCl"[0.0053]CN"CaCl2"[0.00126]CN"NaHCO3"[0.00417]CN"MgCl2"[0.00049]CN"
149 MgSO4"[0.00041]CN"KH2PO4"[0.00044]CN"Na2HPO4"[0.000168]CN"C6H12O6"SM"C([C
150 @H]1[C@H]([C@H]([C@H](C(O1)O)O)O)O)O"[0.0055]ER24.0VA40.0ET37.0IM48.0
151 ELER24.0VA40.0ET37.0IM48.0

152 0.1A/**L0**ME/**L1**Mg/**L2**ID"WE43"NCMg95.5Yy4.0Nd0.5GS15.0/**L3**HT"annealing"TI1.0TE4
153 50.0EG10.0YS113.0BE20.0ET350.0CDDREXWR/**L4**CR0.85 ELCN"NaCl"[0.137]
154 CN"KCl"[0.0053]CN"CaCl2"[0.00126]CN"NaHCO3"[0.00417]CN"MgCl2"[0.00049]CN"
155 MgSO4"[0.00041]CN"KH2PO4"[0.00044]CN"Na2HPO4"[0.000168]CN"C6H12O6"SM"C([C
156 @H]1[C@H]([C@H]([C@H](C(O1)O)O)O)O)O"[0.0055]ER24.0VA40.0ET37.0IM48.0
157 ELER24.0VA40.0ET37.0IM48.0

158

159 Additional substrate/alloy-dependent inhibitor behaviour

160 **Figure SI 1** shows the additional PCA plots of substrate/alloy-dependent inhibitor behaviour ranked 3-11
161 according to substrate coverage. ExCorr database⁸ was used to source all the data. As in the primary examples
162 (**Figure 7b**), the same inhibitor occupies quite different positions in latent space when tested on different
163 substrates or alloys. In several cases, these shifts coincide with changes in IE, with corrosive or weakly
164 inhibiting conditions tending to occur separately from strongly inhibiting conditions. These examples provide
165 further support that the latent representation reflects differences in inhibitor behaviour across metallurgical
166 contexts, rather than simply encoding the molecular identity of the inhibitor.

167

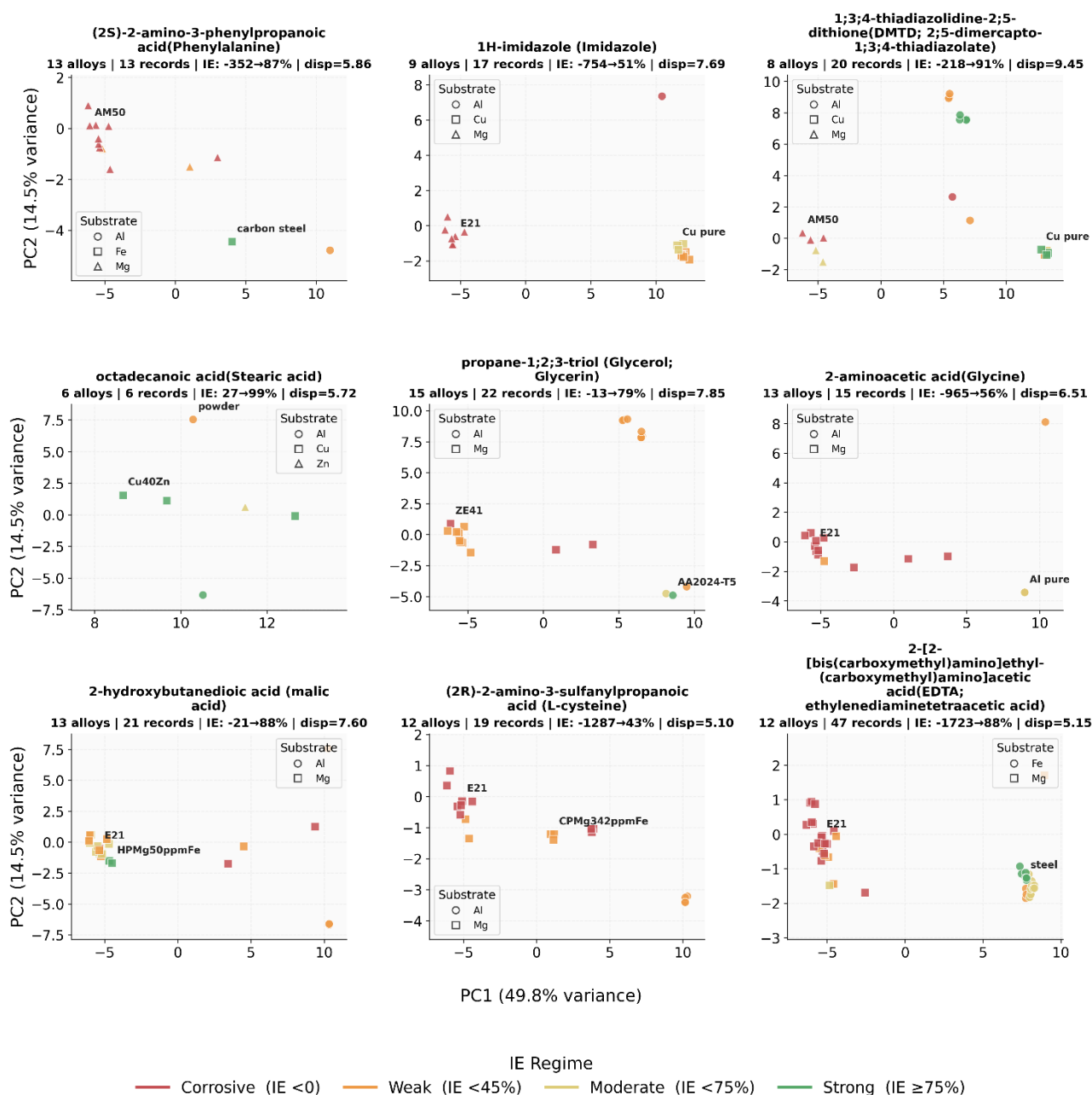


Figure SI 1: Inhibitor-specific latent representations across alloy environments. PCA projections of latent representations for nine inhibitors selected for broad substrate/alloy coverage. Each panel shows the representations of a single inhibitor across different alloys; points are coloured by IE regime. The separation of representations across alloy environments and the corresponding variation in IE illustrate substrate- and alloy-dependent changes in inhibitor representation within the latent space.

Quantitative assessment of latent-space organization

The preceding analyses suggested that substrate identity, alloy composition, and corrosion behaviour each contribute to the organization of the latent space, albeit to different degrees. To quantify these effects, we computed silhouette scores using substrate labels and IE regimes as independent grouping variables.

The latent representations exhibited stronger organization by substrate identity. Across the entire dataset, the global silhouette score was 0.384 for substrate labels. This quantitative difference is consistent with the PCA and UMAP projections, where substrate-dependent structure was readily apparent. Together, these results indicate that substrate-related information represents a dominant source of organization within the latent space. The strength of this organization varied considerably between substrates. Zn-, Fe-, and Mg-based

183 systems exhibited high silhouette scores (0.957, 0.832, and 0.799, respectively), reflecting the compact and
184 well-separated clusters observed in the PCA projection. In contrast, Cu-based systems showed only weak
185 separation (0.111), while Al-based systems exhibited a slightly negative silhouette score (-0.116), indicating
186 substantial overlap with neighbouring regions of the latent space. This behaviour mirrors the diffuse structure
187 observed for Al-containing systems in the PCA analysis and suggests that Al-associated corrosion systems
188 occupy a broader and more heterogeneous region of the latent space than the other substrate classes. This
189 unusually weak cohesion of Al-based systems, together with the strong separation observed for Fe-, Mg-, and
190 Zn-based systems, points to substrate-specific differences in the complexity of the underlying corrosion
191 landscape. This raises the possibility that certain substrate classes, particularly aluminium alloys, encompass
192 a wider range of chemical and metallurgical environments than is captured by substrate labels alone,
193 motivating a more detailed analysis of alloy-specific organization and context-dependent inhibitor behaviour.

194 In contrast to substrate identity, IE regimes exhibited a markedly different pattern. Although the global regime
195 silhouette score was low (0.032), this average masked substantial differences between individual behavioural
196 classes. Corrosive systems formed the most cohesive regime group, with a silhouette score of 0.43, indicating
197 that observations associated with negative IEs occupied a relatively well-defined region of the latent space.
198 In contrast, moderate inhibitors exhibited a strongly negative silhouette score (-0.28), while weak inhibitors
199 were distributed broadly throughout the latent space. Strong inhibitors showed only modest cohesion (0.12).
200 These pronounced differences between regime-specific silhouette scores indicate that inhibition behaviour is
201 not represented uniformly across the latent space. Rather than forming discrete global clusters, behavioural
202 regimes display varying degrees of local organization. The strong cohesion observed for corrosive systems
203 contrasts with the highly heterogeneous distribution of moderate inhibitors, suggesting that intermediate
204 inhibition performance spans a broad range of chemical, metallurgical, and experimental contexts.
205 Consequently, IE does not define a single dominant axis of organization within the latent space. Instead,
206 behavioural information appears superimposed upon a representation whose primary structure is governed
207 by substrate and alloy identity.

208

209 Substrate-induced latent semantic displacement

210 To quantify how metallurgical context influences inhibitor representation, we computed the cosine distance
211 between all cross-substrate observation pairs for each inhibitor appearing on two or more substrates. The
212 reported mean displacement is pair-weighted; to account for the non-independence of pairs sharing the same
213 inhibitor, confidence intervals were estimated by inhibitor-level bootstrap, in which the 57 inhibitors rather
214 than individual pairs were resampled. Across these 57 inhibitors generating 4,107 unique cross-substrate
215 pairs, the mean pairwise displacement was 0.1554 (95% CI: [0.1362, 0.1684]), corresponding to roughly 15%
216 of the unit cosine-distance range and indicating a geometrically meaningful shift rather than a marginal
217 perturbation. A permutation test in which substrate labels were randomly reassigned yielded a null mean of
218 0.0066 ($z = 1286$, $p < 0.001$), confirming that the observed displacement cannot be attributed to random
219 substrate assignment. The distribution of pairwise distances is broadly spread (IQR = 0.048, 5th-95th
220 percentile range: 0.054-0.262) with a visible bimodal structure, suggesting two populations of inhibitor-
221 substrate pairs: one with modest displacement and one with substantially higher cross-substrate divergence.
222 This finding is consistent with the substrate-conditioned PCA trajectories in **Figure SI 2**, and together these
223 results show that substrate identity systematically shifts inhibitor position in the learned latent space:
224 molecular structure alone does not determine where an inhibitor is represented.

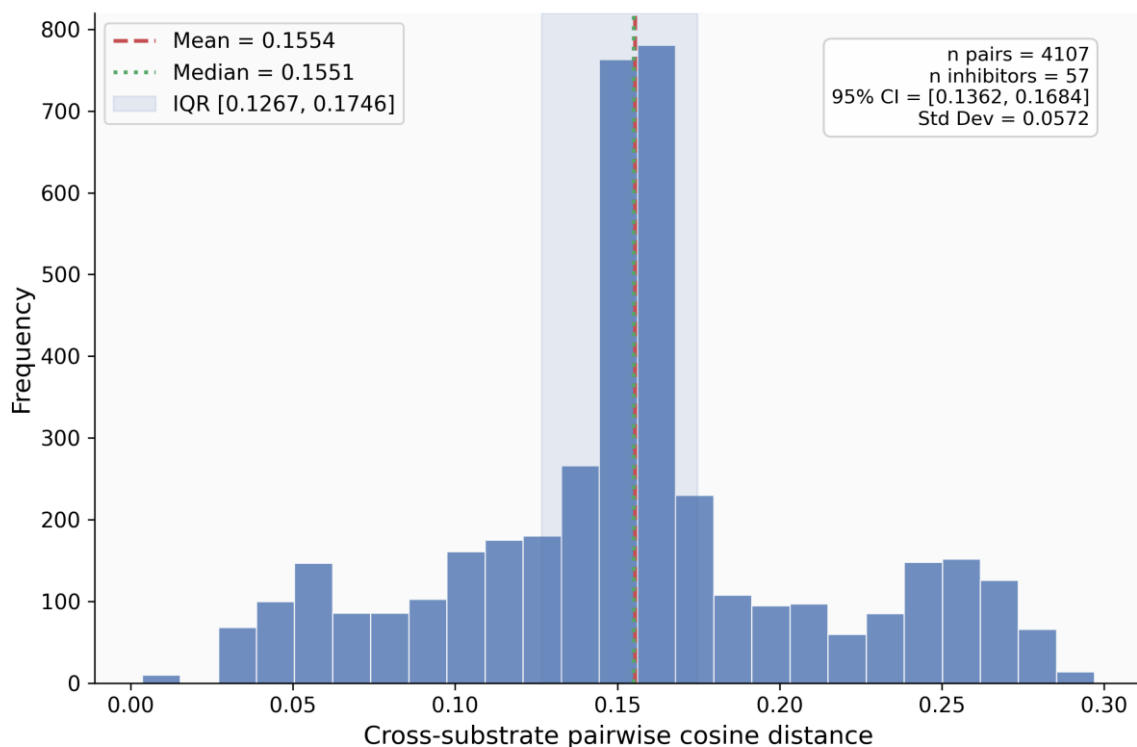


Figure SI 2: Substrate-induced latent semantic displacement. Distribution of cross-substrate pairwise cosine distances between latent representations of the same inhibitor. The analysis includes 4,107 cross-substrate pairs from 57 inhibitors. The mean pairwise displacement was 0.1554 (95% CI: 0.1362-0.1684), with a median of 0.1551. The shaded region denotes the interquartile range (0.1267-0.1746), and the dashed and dotted lines indicate the mean and median, respectively. The distribution shows substantial variation in the displacement of inhibitor representations across substrate environments, indicating that substrate context systematically influences the position of an inhibitor in the latent space.

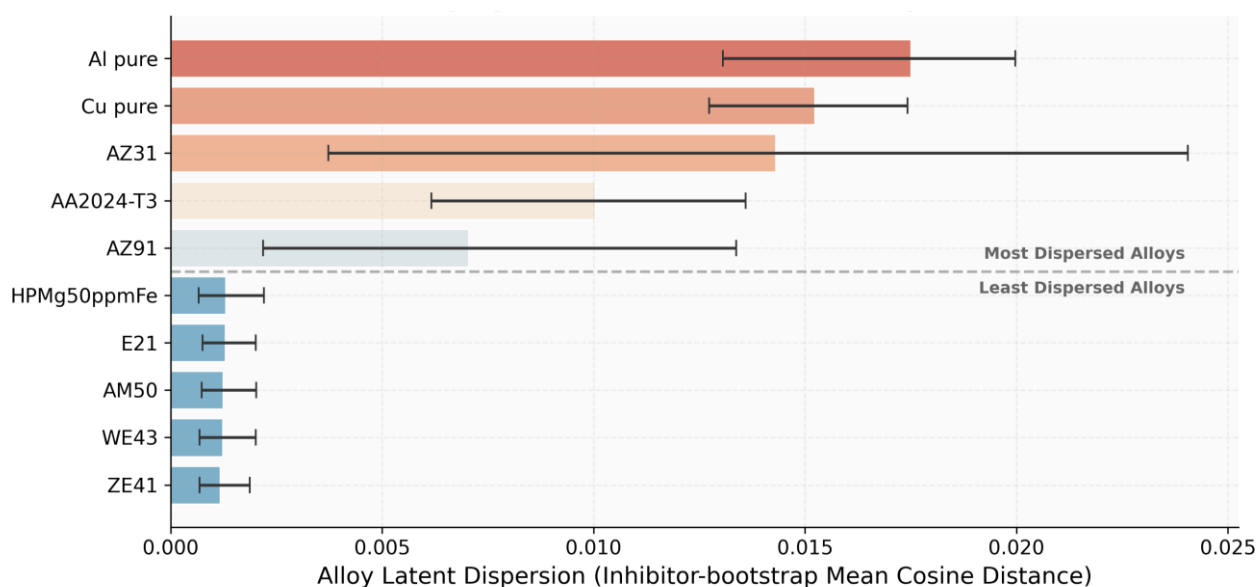
Alloy-dependent latent dispersion

We also examined the internal geometric structure of alloy-specific latent representations by quantifying how broadly the MILES encoder distributes inhibitors within each alloy region of the latent space. We define this quantity, measured as the mean cosine distance between individual inhibitor latent representations and their corresponding alloy centroid, as latent dispersion. To ensure comparable estimates across alloys with different sampling densities, analysis was restricted to alloys represented by at least 50 observations and 30 distinct inhibitors. Dispersion was estimated using an inhibitor-level bootstrap (1,000 iterations), in which inhibitors, rather than individual observations, were resampled. This approach avoids bias arising from repeated measurements of the same compounds, which would otherwise artificially reduce centroid dispersion. Figure SI.3 shows the bootstrap mean latent dispersion and 95% confidence intervals for the five most and five least dispersed alloys.

Pure Al (0.0175) and pure Cu (0.0153) exhibit the highest latent dispersion, **Figure SI 3**, followed by AZ31 (0.0144), AA2024-T3 (0.0101), and AA7075-T6 (0.0064). The broad latent distributions observed for the pure metals are consistent with their role as canonical benchmark substrates in corrosion research, for which chemically diverse inhibitor classes have been investigated. Accordingly, the MILES encoder distributes these chemically diverse inhibitors across a broader region of the latent space. Although AZ31 ranks third, its substantially wider confidence interval (95% CI: 0.005-0.024) overlaps that of AA2024-T3, indicating that their relative ordering cannot be resolved confidently with the current dataset. This uncertainty reflects sparse per-inhibitor sampling. Specifically, AZ31 has 106 distinct inhibitors across only 143 observations, which is a consequence of broad but shallow coverage across independent literature sources.

253 In contrast, the five least dispersed alloys, specifically ZE41, AM50, WE43, E21, and HPMg50ppmFe, are all
 254 specialty magnesium alloys, with dispersion values tightly clustered between 0.0012 and 0.0013 and narrow
 255 confidence intervals that do not overlap with those of the highest-dispersion group. The approximately 15-
 256 fold difference between the extremes demonstrates substantial variation in the internal organization of the
 257 latent space across alloys. The consistently low dispersion of the magnesium alloys is likewise consistent with
 258 the comparatively restricted inhibitor chemistries reported for these systems, where studies have focused on
 259 a narrower range of compounds developed for Mg corrosion, providing the encoder with less chemical
 260 diversity to distribute within the latent space.

261 Together, these results demonstrate that the information encoded in MILES is reflected at multiple levels in
 262 the latent representations, capturing not only relationships between materials but also the internal structure
 263 of their associated chemical and performance landscapes. Such organizations distinguish MILES from
 264 conventional descriptor-based representations.



265
 266 **Figure SI 3: Alloy-specific latent dispersion.** Mean cosine distance between inhibitor latent representations and their corresponding
 267 alloy centroid for the five most and five least dispersed alloys. Error bars indicate 95% confidence intervals estimated by inhibitor-level
 268 bootstrap. Pure Al and pure Cu show the highest latent dispersion, whereas the five least dispersed alloys are magnesium alloys with
 269 substantially tighter distributions.

270

271

272 References

273

- 274 1. Medford AJ, Kunz MR, Ewing SM, Borders T, Fushimi R. Extracting Knowledge from Data through
 275 Catalysis Informatics. *ACS Catalysis* **8**, 7403-7429 (2018).
- 276 2. Nørskov JK, Bligaard T, Hvolbæk B, Abild-Pedersen F, Chorkendorff I, Christensen CH. The nature of
 277 the active site in heterogeneous metal catalysis. *Chemical Society Reviews* **37**, 2163-2171 (2008).
- 278 3. Song Z, Cai T, Hanson JC, Rodriguez JA, Hrbek J. Structure and reactivity of Ru nanoparticles
 279 supported on modified graphite surfaces: A study of the model catalysts for ammonia synthesis.
 280 *Journal of the American Chemical Society* **126**, 8576-8584 (2004).
- 281
- 282

283

284 4. Liu F, *et al.* The processing of Mg alloy micro-tubes for biodegradable vascular stents. *Materials*
285 *Science and Engineering: C* **48**, 400-407 (2015).

286 5. Mao L, *et al.* Nanophasic biodegradation enhances the durability and biocompatibility of
287 magnesium alloys for the next-generation vascular stents. *Nanoscale* **5**, 9517-9522 (2013).
288

289 6. Laleh M, Hughes AE, Tan MY, Rohrer GS, Primig S, Haghdadi N. Grain boundary character
290 distribution in an additively manufactured austenitic stainless steel. *Scripta Materialia* **192**, 115-119
291 (2021).
292

293 7. Peng Y, *et al.* A study of rare-earth 3-(4-methylbenzoyl)-propanoate compounds as corrosion
294 inhibitors for AS1020 mild steel in NaCl solutions. *Corrosion Science* **145**, 199-211 (2018).
295

296 8. Winkler DA, *et al.* Impact of inhibition mechanisms, automation, and computational models on the
297 discovery of organic corrosion inhibitors. *Progress in Materials Science* **149**, (2025).
298
299
300