Corrosion inhibition of aerospace alloys through organic molecules

An end-to-end materials discovery approach from surface analytical and electrochemical experiments to predictive machine learning relationships

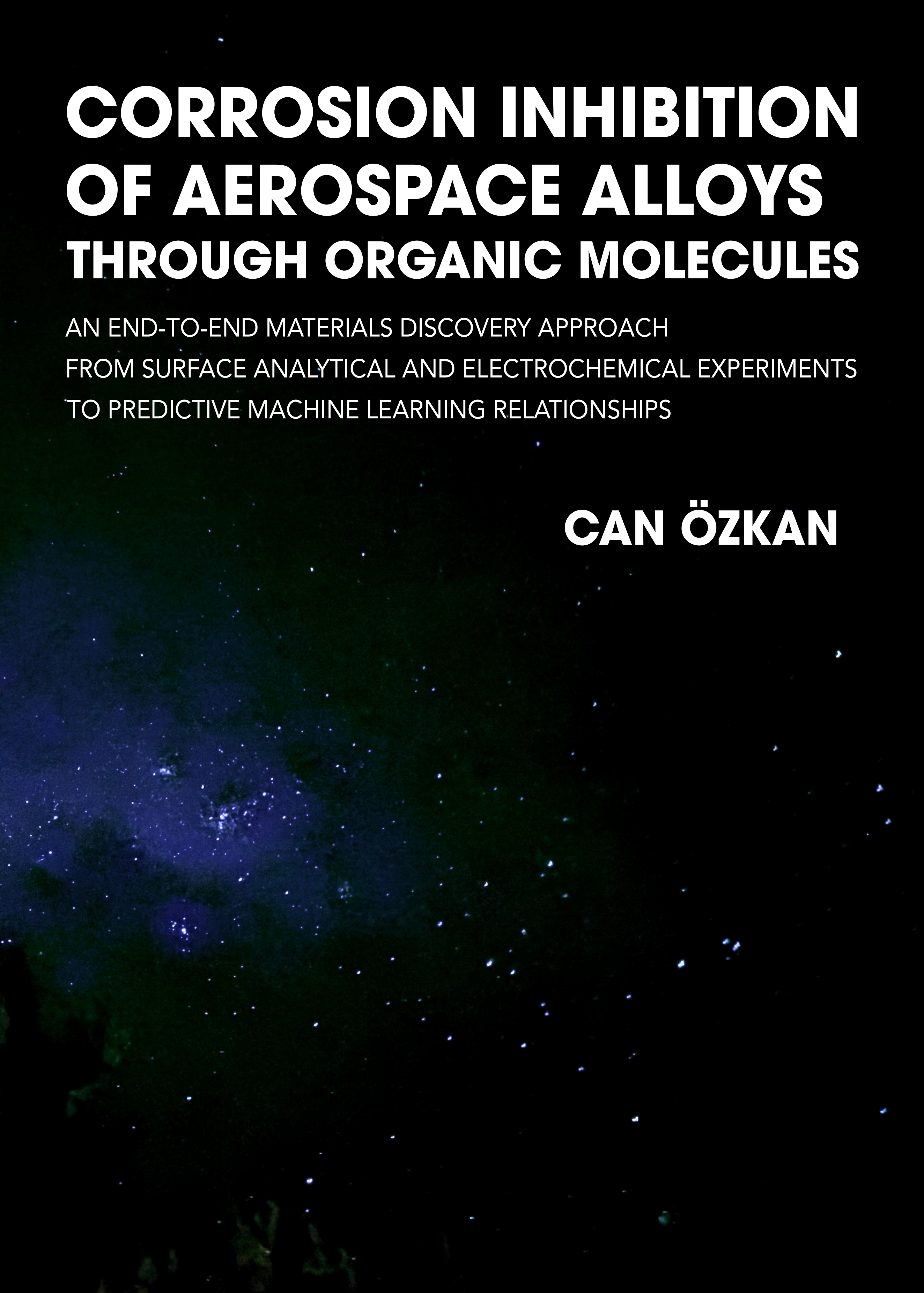
Özkan, C.

DOI
10.4233/uuid:81ec461b-3d47-462a-8f25-2437bf7eac84

Publication date
2026

Document Version
Final published version

Citation (APA)
Özkan, C. (2026). *Corrosion inhibition of aerospace alloys through organic molecules: An end-to-end materials discovery approach from surface analytical and electrochemical experiments to predictive machine learning relationships*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:81ec461b-3d47-462a-8f25-2437bf7eac84



CORROSION INHIBITION OF AEROSPACE ALLOYS THROUGH ORGANIC MOLECULES

AN END-TO-END MATERIALS DISCOVERY APPROACH
FROM SURFACE ANALYTICAL AND ELECTROCHEMICAL EXPERIMENTS
TO PREDICTIVE MACHINE LEARNING RELATIONSHIPS

CAN ÖZKAN

CORROSION INHIBITION OF AEROSPACE ALLOYS THROUGH ORGANIC MOLECULES

AN END-TO-END MATERIALS DISCOVERY APPROACH FROM SURFACE ANALYTICAL AND ELECTROCHEMICAL EXPERIMENTS TO PREDICTIVE MACHINE LEARNING RELATIONSHIPS

CORROSION INHIBITION OF AEROSPACE ALLOYS THROUGH ORGANIC MOLECULES

AN END-TO-END MATERIALS DISCOVERY APPROACH FROM SURFACE ANALYTICAL AND ELECTROCHEMICAL EXPERIMENTS TO PREDICTIVE MACHINE LEARNING RELATIONSHIPS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus Prof. dr. it. T.H.J.J. van der Hagen,
chair of the board for doctorates,
to be defended publicly on
Friday 9th of January 2026 at 10:00

by

Can ÖZKAN

Master of Science in Materials Science and Engineering,
Delft University of Technology, Delft, the Netherlands,
born in İstanbul, Türkiye.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. J.M.C. Mol	Technische Universiteit Delft, promotor
Dr. P. Taheri	Technische Universiteit Delft, copromotor
Independent members:	
Dr. Y. Gonzalez-Garcia	Technische Universiteit Delft
Dr. S. Kumar	Technische Universiteit Delft
Prof. dr. R. Benedictus	Technische Universiteit Delft
Prof. dr. M. Zheludkevich	Helmholtz-Zentrum Hereon, Germany
Prof. dr. P. Marcus	Chimie ParisTech, France
Reserve member:	
Prof. dr. J. Dik	Technische Universiteit Delft, reserve member

The research carried out in this dissertation is a part of the VIPCOAT project *(Virtual Open Innovation Platform for Active Protective Coatings Guided by Modelling and Optimisation)* funded by the European Union Horizon 2020 research and innovation programme with grant agreement number 952903.

Keywords: corrosion inhibition, small organic molecules, aluminium alloy, aerospace materials, chromate replacement, molecular descriptors, electrochemistry, spectroscopy, explainable machine learning

Printed by: Ridderprint ‖ http://www.ridderprint.nl

Front & Back: R. de Rijk, *Aoraki Mackenzie dark sky reserve, NZ.*

ISBN 978-94-6518-099-1

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

T. S. Elliot

Contents

Summary

Corrosion inhibitors are vital for protecting metallic substrates, either as standalone treatments present in surrounding electrolytes, or as leaching components in active protective coatings. While organic molecules offer tremendous versatility due to their nearly infinite structural tunability, their electrochemical performance still falls short of traditional chromate-based systems, especially under dynamic environments present in service conditions. This dissertation aims to analyse the potential of organic molecules as corrosion inhibitors for aerospace alloys by applying a systematic and multidisciplinary approach to evaluate, understand, and ultimately improve the electrochemical performance, stability, and long-term efficacy.

Chapter 1 establishes the societal relevance of corrosion inhibition and provides the scientific background required to contextualise the findings presented in this work.

In **Chapter 2**, a robust experimental framework is developed to generate a high-quality electrochemical library for surface-molecule interactions. AA2024-T3 substrates were exposed to 0.1 M NaCl with more than 100 molecules in 1 mM concentrations, where time-resolved data were created for 24 hours of electrolyte exposure using linear polarisation resistance, electrochemical impedance spectroscopy and potentiodynamic polarisation experiments. This consistent data collection paved the way for analysis of trends across a broad molecular landscape. Extended testing (over six hours), the usage of inhibition power rather than inhibition efficiency as the performance metric, and the inclusion of experimental characteristics such as pH in machine learning models, were found to be critical in more reliable differentiation of the top-performing inhibitors.

Chapter 3 discusses a perspective on the way forward for corrosion inhibitor discovery and optimisation studies, by identifying the influence of key factors that can lead to false positives when seeking to replace hexavalent chromium-based corrosion inhibitors. The electrochemical corrosion inhibition performance of several promising organic molecules in comparison to sodium dichromate was investigated, with a focus on often-overlooked but essential parameters including inhibitor concentration, exposure time, post-presence efficacy, and polarisation stability. Our electrochemical analyses reveal that, although organic molecules can match chromates under certain conditions, their protective performance may degrade significantly under realistic and dynamic environments, potentially leading to misleading conclusions when evaluated in narrow contexts.

Chapter 4 focuses on the *physicochemical stability* previously determined to be critical in corrosion protection during service conditions. A promising molecule was investigated further to determine the root cause of the observed quasi-sustained corrosion inhibition behaviour: remaining partial inhibition even when the molecule is no longer sustained in the environment. The goal was to uncover the underlying mechanisms that could eventually support the design of stable inhibition systems. Through potentiodynamic polarisation, atomic force microscopy and scanning Kelvin probe force microscopy, X-ray photoelectron spectroscopy, attenuated total reflectance Fourier transform infrared spectroscopy, shell-isolated nanoparticle-enhanced Raman spectroscopy, and time-of-flight secondary ion mass spectrometry complemented by density functional theory calculations, the observed quasi-sustained corrosion inhibition was attributed to a sulfate-like adsorption developed between the Al-(hydr)oxide and the thione S moiety of the molecule. Although intermetallics also had sustained molecule adsorption, a change in bonding configuration diminished their corrosion inhibition capabilities. The combined influence of matrix and intermetallic adsorption phenomena resulted in sustained corrosion inhibition, albeit at a reduced efficacy compared to when molecule is present in the environment.

Chapter 5 transitions into the use of machine learning to extract hidden patterns and relationships from the experimental data, offering new insights that go beyond conventional statistical analysis. This approach was designed to address the challenges of working with small datasets, where traditional methods may lack interpretability and fail to uncover the underlying mechanistic drivers of the studied scientific phenomenon. By systematically evaluating over 12,000 model configurations across 29 molecular featurisation strategies and 9 target representations, and using algorithmic feature elimination as a spectroscopy-like tool to understand decision-making process of the most predictive representations, we integrate data-driven patterns with chemical intuition, revealing robust structure-property relationships for corrosion inhibition offered by organic molecules. This methodology emphasises the interpretability of the model's decision-making process, allowing for the identification of meaningful trends that guide the design of novel, non-toxic corrosion inhibitors. Its example application to a toxicity database led to the discovery of the novel non-toxic promising organic corrosion inhibitor 2-thiobarbituric acid, which was validated via electrochemical testing. Ultimately, the framework not only advances machine learning in low-data regimes but also provides actionable insights that drive experimental validation, offering an accelerated pathway for replacing hazardous materials like hexavalent chromium with environmentally sustainable alternatives.

As a conclusion, **Chapter 6** summarises the key findings of the dissertation and outlines future research directions that could build upon this work to develop next-generation corrosion inhibition strategies.

Samenvatting

Corrosie-inhibitoren zijn essentieel voor de bescherming van metalen substraten, zowel als op zichzelf staande behandelingen in de omringende elektrolyt, als uitlogende componenten in actieve beschermende coatings. Hoewel organische moleculen enorme veelzijdigheid bieden door hun bijna oneindige structurele mogelijkheden, lopen hun elektrochemische prestaties achter op traditionele op chromaat gebaseerde systemen, vooral in dynamische omgevingen die van toepassing zijn onder realistische praktijkomstandigheden. Dit proefschrift beoogt het potentieel van organische moleculen als corrosie-inhibitoren voor luchtvaartlegeringen te analyseren door de toepassing van een systematische en multidisciplinaire benadering om de elektrochemische prestaties, stabiliteit en langetermijneffectiviteit te evalueren, te begrijpen en uiteindelijk te verbeteren.

Hoofdstuk 1 stelt de maatschappelijke relevantie van corrosie-inhibitie vast en biedt de wetenschappelijke achtergrond die nodig is om de bevindingen in dit proefschrift te duiden.

In **Hoofdstuk 2** wordt een robuust experimenteel kader ontwikkeld om een hoogwaardige elektrochemische database voor oppervlakte-molecuulinteracties te genereren. AA2024-T3-substraten werden blootgesteld aan 0,1 M NaCl met meer dan 100 moleculen in concentraties van 1 mM. Met behulp van lineaire polarisatieweerstand, elektrochemische impedantiespectroscopie en potentiodynamische polarisatie-experimenten werden hierbij gedurende 24 uur elektrolytblootstelling tijdsafhankelijke gegevens verkregen. Deze consistente gegevensverzameling vormde de basis voor de analyse van trends in een breed moleculair landschap. Uitgebreid testen (meer dan zes uur), het gebruik van inhibitievermogen in plaats van inhibitie-efficiëntie als prestatiemaatstaf, en de opname van experimentele kenmerken zoals pH in machinelearningmodellen werden als cruciaal beschouwd voor een betrouwbaardere differentiatie van de best presterende inhibitoren.

Hoofdstuk 3 bespreekt een perspectief op weg naar de ontdekking en optimalisatie van corrosie-inhibitoren. Dit wordt gedaan door de invloed van belangrijke factoren te identificeren die kunnen leiden tot valse positieven bij de uitdaging om op zeswaardig chroom gebaseerde corrosie-inhibitoren te vervangen. De elektrochemische corrosie-inhibitieprestaties van verschillende veelbelovende organische moleculen in vergelijking met natriumdichromaat werden onderzocht. Hierbij lag de nadruk op essentiële parameters, die vaak over het hoofd worden gezien, zoals inhibitorvariëteit, blootstellingstijd, effectiviteit na de aanwezigheid van inhibitoren en polarisatiestabiliteit. Hoewel organische moleculen onder bepaalde omstandigheden chromaten kunnen evenaren, onthullen onze elektrochemische analyses dat hun beschermende prestaties aanzienlijk kunnen afnemen in realistische en dynamische omgevingen, wat kan leiden tot misleidende conclusies bij een evaluatie in beperkte blootstellingscontext.

Hoofdstuk 4 richt zich op de *fysisch-chemische stabiliteit*, waarvan eerder al werd

vastgesteld dat die cruciaal is bij corrosiebescherming onder realistische blootstellingsomstandigheden. Een veelbelovend molecuul werd verder onderzocht om de oorzaak van het waargenomen quasi-stabiele corrosie-inhibitiegedrag te bepalen: blijvende gedeeltelijke inhibitie, zelfs wanneer het molecuul niet langer in de omgeving aanwezig is. Het doel was om de onderliggende mechanismen te ontdekken die uiteindelijk het ontwerp van stabiele inhibitiesystemen kunnen ondersteunen. Door gebruik te maken van potentiodynamische polarisatie, atoomkrachtmicroscopie en scanning Kelvin-probekrachtmicroscopie, röntgenfoto-elektronenspectroscopie, totaalreflectie-Fourier-transformatie-infraroodspectroscopie, schil-geïsoleerde nanopartikel-versterkte Raman-spectroscopie, en time-of-flight secundaire ionen-massaspectrometrie aangevuld met dichtheidsfunctionaaltheorieberekeningen, werd het waargenomen quasi-stabiele corrosie-inhibitiegedrag toegeschreven aan een sulfaatachtige adsorptie die zich ontwikkelde tussen de Al-(hydr)oxide en het thione S-gedeelte van het molecuul. Hoewel intermetallische verbindingen ook duurzame molecuuladsorptie vertoonden, verminderde een verandering in de bindingsconfiguratie hun corrosiewerende capaciteiten. De gecombineerde invloed van matrix- en intermetallische adsorptiefenomenen resulteerde in een duurzame corrosie-inhibitiereactie, zij het met verminderde effectiviteit in vergelijking met een situatie waarin het molecuul aanwezig is in de omgeving.

Hoofdstuk 5 maakt de overstap naar het gebruik van machine learning om verborgen patronen en relaties uit de experimentele gegevens te halen, en biedt nieuwe inzichten die verder gaan dan conventionele statistische analyses. Deze benadering is ontwikkeld om de uitdagingen van werken met kleine datasets aan te pakken, waar traditionele methoden mogelijk interpreteerbaarheid missen en er daarom niet in slagen de onderliggende mechanistische drivers van het wetenschappelijke fenomeen dat wordt bestudeerd te onthullen. Door voor 29 strategieën voor moleculaire kenmerken en negen doelrepresentaties systematisch meer dan 12.000 modelconfiguraties te evalueren en door gebruik te maken van algoritmische kenmerkverwijdering, als een spectroscopie-achtig middel om het besluitvormingsproces van de meest voorspellende representaties te begrijpen, integreren we datagedreven patronen met chemische intuïtie, wat robuuste structuur-eigenschaprelaties voor corrosie-inhibitie door organische moleculen laat zien. Deze methodologie benadrukt de interpreteerbaarheid van het besluitvormingsproces van het model, waardoor het mogelijk wordt om betekenisvolle trends te identificeren die het ontwerp van nieuwe, niet-giftige corrosie-inhibitoren sturen. De toepassing van deze methodologie op een toxiciteitsdatabase leidde tot de ontdekking van de veelbelovende nieuwe, niet-toxische organische corrosie-inhibitor 2-thiobarbituurzuur, die werd gevalideerd door middel van elektrochemisch testen. Uiteindelijk bevordert dit kader niet alleen machine learning in lage-dataregimes, maar biedt het ook bruikbare inzichten die experimentele validatie stimuleren. Tegelijkertijd presenteert het een versneld traject voor het vervangen van gevaarlijke materialen zoals zeswaardig chroom door milieuvriendelijke alternatieven.

Concluderend vat **Hoofdstuk 6** de belangrijkste bevindingen van het proefschrift samen en schetst toekomstige onderzoeksrichtingen die hierop voort kunnen bouwen om strategieën te ontwikkelen voor een volgende generatie corrosie-inhibitoren.

Özet

Korozyon inhibitörleri (yavaşlatıcıları), ister çevreleyen elektrolitte (örneğin, akü ya da deniz suyu) tek başına bulunarak, isterse aktif koruyucu kaplamaların yapısındaki çözünüp açığa çıkan bileşenler olarak, metalik yüzeylerin korunmasında hayati öneme sahiptir. Organik moleküller, neredeyse sınırsız yapısal ayarlanabilirlikleri sayesinde büyük esneklik sunsalar da, geleneksel kromat bazlı sistemler ile karşılaştırıldıklarında değişken gerçek hayattaki koşullar altında elektrokimyasal performansları hâlâ yetersiz kalmaktadır. Bu tez çalışması, organik moleküllerin havacılık alaşımları için korozyon inhibitörü potansiyelini sistematik ve çok disiplinli bir yaklaşımla değerlendirip anlamayı ve nihayetinde elektrokimyasal performans, korumanın sabitliği ve uzun dönem etkinliği çerçevesinde iyileştirmeyi amaçlamaktadır.

Bölüm 1 korozyon inhibisyonunun iktisadi ve toplumsal maliyetini ortaya koyar ve bu çalışmada sunulan bulguları ortak bir bağlam içine yerleştirmek için gerekli olan bilimsel temel arka planı sağlar.

Bölüm 2'de yüzey–molekül etkileşimlerine ilişkin kapsamlı bir elektrokimyasal veri kütüphanesi oluşturmak üzere sağlam bir deneysel çerçeve geliştirilmiştir. AA2024-T3 alaşım yüzeyleri, 0.1 M NaCl çözeltisinde 1 mM konsantrasyonda 100'den fazla organik moleküle maruz bırakılmış; 24 saatlik elektrolit teması süresince lineer polarizasyon direnci, elektrokimyasal empedans spektroskopisi ve potensiyodinamik polarizasyon deneyleri sayesinde zamana dayanan veriler toplanmıştır. Bu tutarlı veri toplama yöntemi, geniş bir moleküler yelpazedeki eğilimlerin analizine olanak tanımıştır. Altı saati aşan uzun süreli testler, performans metriği olarak inhibisyon verimliliği yerine inhibisyon gücünün kullanılması, ve pH gibi deneysel özelliklerin makine öğrenmesi modellerine dahil edilmesi, en yüksek performanslı inhibitörlerin daha güvenilir biçimde tespit edilmesinde önemli bulunmuştur.

Bölüm 3 heksavalent kromat esaslı inhibitörlerin yerine geçebilecek adayları araştırırken yanlış pozitif sonuçlara yol açabilecek temel etkenleri belirleyerek, korozyon inhibitörü keşif ve optimizasyon çalışmalarının nasıl ilerlemesi gerektiğine dair bir perspektif sunar. İnhibitör i) konsantrasyonu, ii) maruz kalma süresi, iii) koruyucu etkinliğinin çözelti ortamından uzaklaştırıldıktan sonraki kalıcılığı, ve iv) polarizasyon kararlılığı, gibi sıklıkla göz ardı edilen ancak aslında hayati öneme sahip olan parametreler dikkatle incelenmiştir. Bu etkenler dikkate alınarak, bir dizi umut verici organik molekülün korozyon inhibisyon performansı sodyum dikromatla karşılaştırılmistir. Elektrokimyasal analizlerimiz, belirli koşullar altında organik moleküllerin kromatlarla eşdeğer koruma sağlayabildiğini, ancak doğal değişken ortamlarda performanslarının önemli ölçüde düşebileceğini göstermiş, ayrıca dar kapsamlı değerlendirmelerdeki yanıltıcı sonuçlara varma riskine dikkat çekilmiştir.

Bölüm 4 bir önceki bölümde inhibitör keşfinde önemli olduğu belirlenen fizikokimyasal sabitlik üzerine odaklanır. Ümit vadeden bir molekül, ortamdan uzaklaştırıldıktan sonra bile kısmen devam eden inhibitör etkisinin kaynağını ortaya çıkarmak amacıyla derinlemesine incelenmiştir. Potensiyodinamik polarizasyon, atomik kuvvet mikroskobu, taramalı Kelvin sondası mikroskobu, X-ışını fotoelektron spektroskopisi, azaltılmış toplam yansıma Fourier dönüşümlü kızılötesi spektroskopisi, kabuk yalıtılmış nanoparçacık destekli Raman spektroskopisi, ve zaman uçuşlu ikincil iyon kütle spektrometrisi gibi deneysel teknikler yoğunluk fonksiyonel teorisi hesaplamalarıyla desteklenmiştir. Gözlenen kısmi-kalıcı korozyon inhibisyonunun aluminyum (hidr)oksit yüzeyi ile molekülün tiyol S grubu arasında gelişen sülfat benzeri bir kalıcı yüzeye tutunma sayesinde gerçekleştiği sonucuna varılmıştır. İntermetalik bölgelerde de devam eden molekül yüzeye tutunması gözlenirken, bağlanma şeklindeki farklılıklar korozyon korumasının yok olmasına neden olmuştur. Matris ve intermetalik yüzeye tutunma etkileşimlerinin birleşik etkisi, molekülün ortamdan çekildikten sonra bile belirli bir miktarda koruma sagladığını, ancak etkinliğin molekülün ortamda bulunduğu zamanki kadar yüksek oranda olmadığını göstermiştir.

Bölüm 5 deneysel verilerde gözlemlenen ilişkileri ortaya çıkarmak için yapay zeka kullanımına odaklanmaktadır. Burada geliştirilen metodoloji, küçük veri kümeleriyle çalışmanın zorluklarına odaklanıp, geleneksel yöntemlerin anlaşılabilirlik eksikliği ve mekanistik sebepleri ortaya çıkaramaması sorunlarına çözüm arar. 29 moleküler özellik çıkarım stratejisi ile 9 hedef temsilini içeren 12.000'den fazla model konfigürasyonu sistematik olarak değerlendirilmiştir. Algoritmik özellik eleme yönteminin spektroskopi-benzeri bir araç olarak kullanımı sayesinde, organik moleküllerin yapısı ile korozyon inhibitörlüğü özellikleri arasındaki ilişkiler ortaya çıkarılmıştır. Bu yöntem, modelin karar verme sürecinin anlaşılabilirliğini vurgulayarak, inhibitörlük özelliğinde önemli olan molekül yapı taşlarının belirlenmesini sağlar. Örnek bir zehirlilik veri tabanını kullanan çalışmada, elektrokimyasal testlerle doğrulanan 2-tiyobarbitürik asit adlı yeni, zehirsiz bir inhibitör keşfedilmiştir. Sonuç olarak, bu metod hem sınırlı veri şartlarında makine öğrenmesini ilerletmekte, hem de halihazırda kullanılan tehlikeli maddeler yerine sürdürülebilir alternatiflere geçiş için deneysel doğrulamayı sağlayan hızlandırılmış bir yol sunmaktadır.

Bölüm 6 tezin temel bulgularını özetler, ve bu çalışmayı ilerletip gelecek nesil korozyon inhibisyon stratejilerinin geliştirilmesi için araştırma yönlerini ortaya koyar.

Preface

New directions in science are launched by new tools much more often than by new concepts. The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained.

Freeman Dyson

Materials science is currently at an exciting crossroads, shaped by two powerful forces.

On one hand, the urgency of the environmental crisis looms large. Climate change, driven by greenhouse gas emissions, is causing rising temperatures that shrink habitats, exacerbate food and water scarcity, and trigger extreme weather events. Global CO_2 emissions reached a record high in 2024, a year that also marked the first time global temperatures exceeding 1.5°C above pre-industrial levels. The consequences of climate crisis are neither distant nor abstract; even as I am writing this in Delft, the Netherlands, the sun has not emerged for over two consecutive weeks. This intensifying climate emergency poses escalating threats to biodiversity, human health, and critical infrastructure, creating an urgent demand for sustainable and cost-effective material solutions.

On the other hand, the rapid advancement of new technologies is opening exciting new possibilities. Self-driving laboratories, which have revolutionized drug discovery by identifying active small molecules, are now making their way into fields such as catalysis, carbon capture, and corrosion inhibition. While new datasets are emerging, they are often too small for traditional machine learning approaches. As a result, innovative methods for applying machine learning to smaller science datasets are being developed, paving the way for breakthroughs in the discovery of new materials.

This dissertation aims to strike a balance between these two forces — the pull of environmental necessity and the push of technological advancement — in the hope of contributing to a brighter, more sustainable future.

Can ÖZKAN
Delft, January 2025

LIST OF FIGURES

List of Tables

1

On Corrosion Inhibition

Iron which has been acted upon by fire is spoiled, unless it is forged with the hammer.
It is not in a fit state for being hammered when it is red-hot,
nor, indeed, until it has begun to assume a white heat.

By sprinkling vinegar or alum upon it, it acquires the appearance of copper.
It is protected from rust by an application of ceruse, gypsum, and tar;
a property of iron known by the Greeks as "antipathia".

Some say that this may be ensured by the performance of certain religious ceremonies,
and that there is in existence at the city of Zeugma, upon the Euphrates,
an iron chain,

by means of which Alexander the Great constructed a bridge across the river;
the links of which that have been replaced are infested with rust,
while the original links are totally exempt from it.

Pliny the Elder, Natural History Book 34 (AD 77)

Parts of this chapter have been published as *Impact of inhibition mechanisms, automation, and computational models on the discovery of organic corrosion inhibitors*, Progress in Materials Science (2024) [1].

1.1. Context

1.1.1. Societal importance of finding novel corrosion inhibitors

Corrosion is the reversal of metals to thermodynamically more stable states such as their oxides or hydroxides [2]. Despite being a relatively slow process, it is one of the leading causes of loss of integrity for metal products [3]. In fact, the economic impact of corrosion surpasses that of all other natural disasters combined [4]. It is estimated to eat away 3.4 % of the annual global gross domestic product (2.5 trillion U.S. dollars in 2002, inflation corrected to be approximately €750 per person each year in 2025) by causing catastrophic structural failures of engineered systems resulting in loss of life, injury, and substantial economic damage across a wide range of industries [5]. Notably, improvements in corrosion control could enhance the lifespan of metal components significantly - according to a recent NACE report, implementing best practices for corrosion prevention could save between 15% and 35% of these costs, potentially totaling up to 875 billion dollars [6].

One of the most efficient corrosion control approaches is corrosion inhibition. A corrosion inhibitor is a substance that minimises metal degradation, reduces hydrogen embrittlement, and/or prevents pitting when added in small concentrations to a given environment. Small is relative for any industry, but in the context of aerospace, this would mean concentrations on the order of mM. Corrosion inhibitors can either be used as is, or added to protective coating formulations to act as active pigments in the event of a coating damage. Figure 1.1 depicts this concept for an aircraft protection scheme.

Chromate compounds have been historically employed as universal corrosion inhibitors for a wide array of metals, but their extreme toxicity with their reprotoxic, mutagenic, and carcinogenic properties [7, 8] (one estimate reported that workers exposed to chromate residues during aircraft repainting had 250,000-fold higher risk of cancer

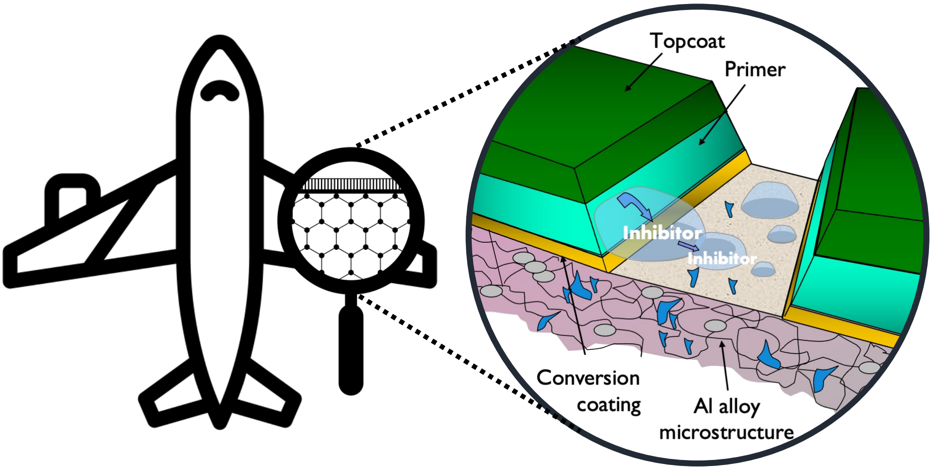


Figure 1.1: Graphical representation of a plane coating cross-section. In the case of a coating breach, inhibitor material would leach out of from the primer coating and interact with the metallic substrate, inhibiting corrosion.

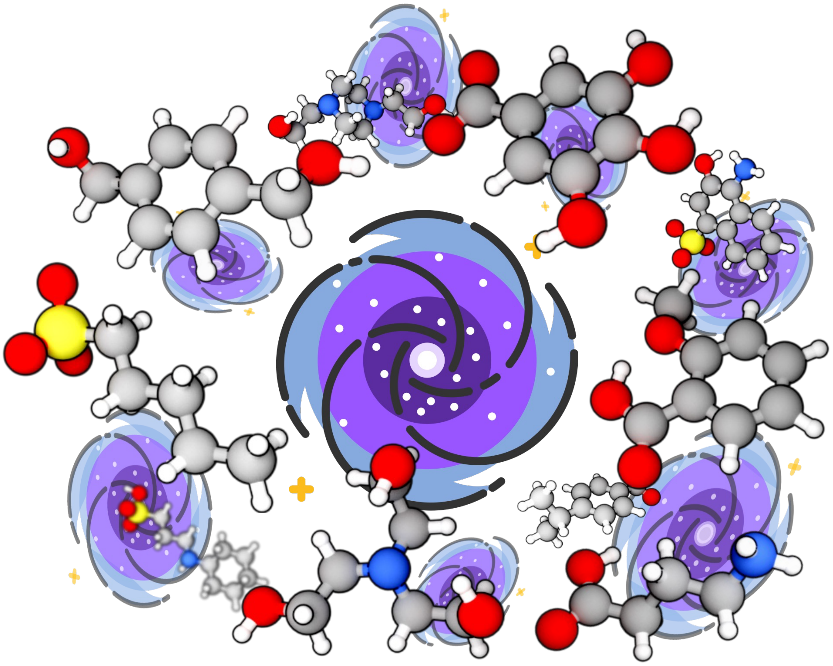


Figure 1.2: A comparison of the number of stars in the universe to the number of synthesisable compounds. There are approximately 10^{21} stars in the observable universe, and 10^{63} synthesisable compounds. Image by courtesy of Tim Wuerger.

than the rest of the public [9, 10]), resulted in it being phased out in socially responsible countries [8, 10, 11].

Despite progress in reducing chromate use, challenges remain with regard to the availability and cost of effective replacements. While the substitution of chromates in corrosion protection applications is possible, these alternatives must be equally cost-effective and efficient, particularly in highly specialised fields such as aerospace.

To this end, organic molecules offer a broad range of potential corrosion inhibition solutions due to their diversity, stability, and selectivity under operational conditions. Small organic molecules typically function by adsorbing onto the surfaces of metal substrates, displacing water, and forming protective thin films that inhibit the ingress of corrosion inducing aggressive ions like chlorides. They can be completely safe to use, or much less harmful than chromates, but they can also be quite toxic depending on their structure. The balance between toxicity and corrosion inhibition performance should serve as a key consideration in the pursuit of safer and more effective organic corrosion inhibitors.

Unfortunately, no organic inhibitor currently matches the wide applicability and effectiveness of chromates. Finding an organic molecule replacement requires a structured search through potential small organic molecules. However, this is no trivial task. Figure 1.2 highlights the potential chemical search space and its comparison to the cosmic space. Space is vast. Chemical space, even more so, with estimated 10^{63} synthesisable organic molecules [12–14]. Considering the size of such a large chemical search space of organic molecules, the scale of an exhaustive search is daunting, but the opportunity of a bespoke solution for every surface science problem in need of an answer equally exhilarating.

1.1.2. Challenges of materials discovery with machine learning

Machine learning has emerged as one of the most suitable methods for screening vast chemical and material spaces to identify promising compounds [15–25]. However, fundamental research in molecular and materials science often grapples with the challenge of smaller data sets, making it difficult to train predictive models effectively and identify the most appropriate machine learning algorithms [26–28]. These smaller data sets, frequently derived from human-conducted experiments or subjectively collected information, present significant limitations for complex analyses that aim to uncover causal relationships. Despite advances in automated data acquisition techniques, such as emerging self-driving labs [29, 30], a substantial portion of the natural science data used in machine learning applications still qualifies as "small data". This issue is particularly pronounced in fields such as corrosion inhibition studies, where the lack of extensive, high-quality data sets persists [31–35].

In addition to limited data availability, inconsistencies between different studies further complicate the process of selecting and training machine learning models. Small data approaches are fraught with challenges such as data imbalance, as well as the risks of model overfitting or underfitting, due to the limited scale of the data or the high dimensionality of features required by many machine learning models. These challenges hinder the ability to draw reliable predictions and insights, slowing down progress in the adoption of advanced computational methods.

The scarcity of large, chemically diverse, high-quality data sets (typically exceeding 1000 compounds) is a significant barrier to the broader application of data-driven machine learning. This limitation is particularly restrictive for state-of-the-art architectures such as transformer models and graph neural networks, which rely on extensive and diverse datasets in the order of millions, if not billions of datapoints. Promising approaches, such as transfer learning and active learning, offer potential workaround solutions. However, their success hinges on the availability of either large, relevant datasets or the integration of intelligent experimental designs to guide data collection. Overcoming these obstacles is critical for unlocking the full potential of machine learning in molecular and materials research.

1.2. Scientific Background

1.2.1. Corrosion of aluminium alloys

Despite being the most abundant metallic element on Earth, aluminium (Al) was mostly hidden away from the engineering stage until the invention of Hall-Heroult process in 1886. Thanks to its lightness, high performance, and corrosion resistance, it has grown to be one of the most important engineering materials.

During their first manned flight in 1903, the Wright brothers chose Al as their pre-

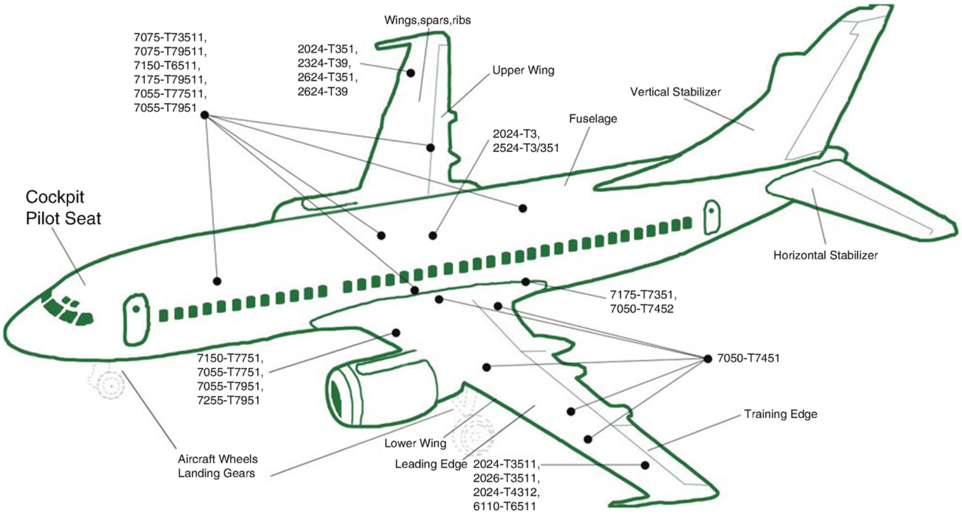


Figure 1.3: Most widely used aluminium alloy grades in an aircraft [38].

ferred material to make the cylinder block and other engine parts. Since then, Al alloys have been widely used in aerospace structures ranging from rocket fuel tanks to fuselages thanks to their high specific strength, low density, good machinability and formability, good maintainability, and low cost. Despite the growing use of composites in widebody aircraft, aluminium-copper 2xxx series alloys and aluminium-zinc 7xxx series alloys are still dominant in small aircraft [36]. Figure 1.3 displays the distribution of aluminium alloys in such an aircraft.

For structural engineering applications, aluminium is alloyed with elements such as magnesium, copper and silicon to improve the mechanical properties. Further strengthening of the aluminium alloy is achieved through thermal and mechanical processes. In this way, the low strength of pure aluminium which is around 10 MPa can be increased up to 800 MPa through alloying and age hardening processes [37]. The processing modifies the microstructural properties of the alloy such as the grain size, grain boundary properties, secondary phase constituents, inclusions and texture. Besides being important for the mechanical response, these microstructural features determine the electrochemical properties of the material.

MICROSTRUCTURE OF ALUMINIUM ALLOYS

The most important microstructural features are usually intermetallics that are formed of Al and transition metals such as Cu, Cr, Mn, Fe, Zr, and other abundant alloying elements such as Si, Li and Mg. The combination of different alloying mixtures with further thermal and/or mechanical processing results in a plethora of microstructural features with different scales: comparatively large irregularly shaped insoluble constituent phases (0.5-10 μm) that are formed out of Fe, Mn and Si, dispersoids (0.1 - 0.5 μm) which limit grain growth through pinning of the grain boundaries among others, and age-hardening precipitates (< 0.2 μm) which provide strengthening primarily by impeding dislocation motion [39, 40]. The size and chemical compositions of these components vary significantly, and equilibrium and non-equilibrium phases coexist together. As a result, the microstructures display a complex behaviour that follows a quasi rather than full thermodynamical equilibrium.

Unfortunately high-strength aluminium alloys are often susceptible to corrosion as a direct result of their microstructure. The two high-strength Al alloys studied the most for their corrosion related properties are the legacy alloys AA2024-T3 and AA7075-T6. This work focuses on AA2024-T3, an alloy widely used in fuselage, wings, shear bars, webs, and other structural parts that require high strength [36].

AA2024-T3

For decades AA2xxx alloys have been one of the top choices of the transportation industry due to their mechanical properties. Despite their corrosion susceptibility, they are widely used in aircraft fuselage manufacturing thanks to their high strength to weight ratio, fracture toughness and good fatigue crack growth resistance.

AA2024-T3 is a wrought aluminium alloy. Main composition of AA2024 includes Cu (3.5-5 wt.%) and Mg (1-2 wt.%); in less concentrations Mn, Si, Ti, Fe, Zr and other trace elements. The microstructure undergoes an age-hardening tempering process named T3 to further improve the mechanical properties. During the T3 tempering process, alloy is solution heat-treated, cold worked, and naturally aged.

Due to its high alloying content and constituent particles with complex stoichiometries, the microstructure of AA2024-T3 is one the most complicated among the Al alloys [37]. Typical intermetallics recognised in the 2xxx series alloys are: precipitates Al_2Cu (ϑ-phase) and Al_2CuMg (S-phase), dispersoid $Al_{20}Mn_3Cu_2$, and constituents $Al_{20}Mn_3Cu$, $Al_6(CuFeMnSi)$ and Al_7Cu_2Fe. 2.7% of the total surface area is covered by the S-phase, which may occupy up to 60% of the total intermetallic surface [41, 42].

However, many intermetallics are not single phase but in fact of multiphase nature, as they show heterogenous stoichiometries in one region [43]. On top of that, periphery dispersoid-free regions were examined around intermetallic particles. The complicated character of the AA2024-T3 microstructure directly influences the electrochemical properties, resulting in a bewitching corrosion behaviour.

Corrosion of aluminium

Pure aluminium is a very electrochemically active metal with a standard reduction potential of -1.66 V. Figure 1.4 presents the Pourbaix diagram of Al which demonstrates this behaviour, as well as the stability regions of common phases. Despite its electrochemical activity, between pH values 4-9, the thermodynamically stable form of Al is a passive oxide layer. Because of this thermodynamical tendency, when exposed to the environment the Al surface reacts with its surroundings and forms an amorphous (hydr)oxide layer with a thickness on the order of magnitude of nanometers. This passive layer protects the underlying Al substrate against the ingress of corrosive species and decreases the corrosion rate [44, 45].

Al corrosion consists of two electrochemical half-cell reactions that keep the electroneutrality during corrosion. An anodic part where the aluminium loses electrons through oxidation and becomes Al^{3+} cation:

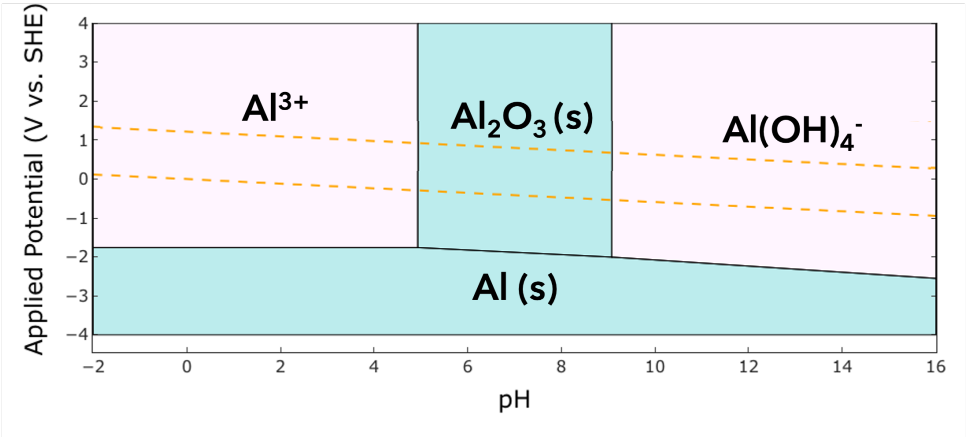


Figure 1.4: Pourbaix diagram of Al calculated through `pymatgen` package [46]. Pink regions display ionic phases, cyan regions solid phases. Dashed lines show water stability lines.

$$Al \rightarrow Al^{3+} + 3e^{-} \tag{1.1}$$

and a cathodic part where the excess electrons are consumed via oxygen reduction reaction (ORR):

$$O_{2(aq)} + 2H_2O + 4e^{-} \rightarrow 4OH^{-} \tag{1.2}$$

or in acidic environment via hydrogen evolution reaction (HER):

$$2H^{+} + 2e^{-} \rightarrow H_2 \tag{1.3}$$

As the corrosion progresses anodic zones will become acidic and cathodic zones will become caustic as the half-reactions proceed. Conversion of atomic Al to ionic Al occurs very rapidly, at time-scales faster than 10^{-5} seconds. In about a microsecond, ionic Al undergo hydration to form hexa-coordinated complexes [47]. It is hypothesised that a variety of mono-nuclear:

$$Al^{3+} + H_2O \rightleftharpoons Al(OH)^{2+} + H^{+} \tag{1.4}$$

$$Al^{3+} + 2H_2O \rightleftharpoons Al(OH)^{2+} + 2H^{+} \tag{1.5}$$

$$Al^{3+} + 3H_2O \rightleftharpoons Al(OH)^{2+} + 3H^{+} \tag{1.6}$$

$$Al^{3+} + 4H_2O \rightleftharpoons Al(OH)^{2+} + 4H^{+} \tag{1.7}$$

and poly-nuclear:

$$2Al^{3+} + 2H_2O \rightleftharpoons Al_2(OH)_4 + 2H^{+} \tag{1.8}$$

$$3Al^{3+} + 4H_2O \rightleftharpoons Al_3(OH)_4^{5+} + 4H^+ \tag{1.9}$$

$$13Al^{3+} + 28H_2O \rightleftharpoons Al_{33}O_4(OH)_{24}^{7+} + 32H^+ \tag{1.10}$$

hydrolysis products are formed. In the case of passivation, released Al^{3+} ions react to the surrounding environment with disassociation of water and form a protective (hydr)oxide layer through:

$$2Al + 3H_2O \rightleftharpoons Al_2O_3 + 6H^+ + 6e^- \tag{1.11}$$

When aluminium does corrode, the required environmental conditions are often one of the two cases. First case requires extreme pH environments, where thermodynamically stable form of Al becomes Al^{3+} in acidic ($pH < 4.5$), and $Al(OH)_4^-$ in alkaline ($pH > 9$) environments. This results in uniform corrosion. Second case requires the presence of halide ions such as Cl^-, which can initiate localised corrosion through the breakdown of the passive film. Cl^- ions form highly soluble corrosion products:

$$Al^{3+} + Cl^- \rightleftharpoons AlCl^{2+}/AlCl_2^+/AlCl_3 \tag{1.12}$$

$$AlCl_3 + 3H_2O \rightleftharpoons Al(OH)_3 + 3HCl \tag{1.13}$$

which promote the breakdown of passivation.

In most cases Al corrosion progresses in the form of localised corrosion pits and crevices rather than the overall corrosion of the surface.

Localised corrosion

Unlike pure aluminium, the oxide layer of Al alloys is defective due to the diverse surface microstructure. The diverse composition is a direct result of the alloying. An aluminium matrix rich in intermetallics that range from nanometers to micrometers, forms a heterogenous oxide morphology which leads to an electrochemically active surface prone to localised corrosion [48, 49].

Intermetallics interact by forming micro-galvanic cells among each other and with the aluminium matrix due to their distinct corrosion potentials. It was commonly believed that intermetallics either take the role of the cathode and dissolve the surrounding Al matrix, or take the role of the anode and dissolve themselves based on their electrochemical potential. However, a recent study combined OCP evaluation with high-resolution optical microscopy and demonstrated that both anodic and cathodic intermetallics of AA2024-T3 show a similar degradation morphology [50]. It was suggested that all intermetallics become active, go through rapid early-stage dealloying and proceed with an order of magnitude slower local trenching process around the intermetallic. The driving force is the elemental composition dependent electrochemical properties. As the composition of intermetallics change with time, the electrochemical properties also change. This results in a dynamic anode/cathode nature of intermetallics with respect to their surrounding surface.

Electrochemical characteristics of a span of intermetallics of Al alloys have been documented previously [39, 40, 51]. With corrosion potentials that range in a span of 0.5 V [41], Al intermetallics act as significant defect sites which result in micro-galvanic corrosion of the surface.

Such diversity of microstructure phases results in different modes of localised corrosion. The main forms can be encapsulated under the classes of intergranular, pitting, crevice, filiform corrosion; exfoliation and stress corrosion cracking. Out of these, pitting and intergranular corrosion is especially important for Al with high alloying content.

In intergranular corrosion, active intermetallics with less noble corrosion potential precipitate into the grain boundaries, making it anodic compared to the rest of the grain. This attack on grain boundaries cause cracks to precipitate along the grain boundaries. An example of this phenomena is seen after the heat treatment of AA2024-T3. Copper depleted zones of Al_2Cu and Al_2CuMg intermetallics form at the grain boundaries, resulting in severe intergranular corrosion through anodisation of the grain boundaries [41, 52].

In pitting corrosion, deep tunnels form on the surface due to electrochemical activity. In chloride environments pitting is the main corrosion mechanism. Chloride ions adsorb to the surface of the oxide film and initiate pitting through the breakdown of the passive film. Pits develop through the surface via the dissolution of exposed Al matrix, forming locally acidic conditions that exacerbate corrosion.

Figure 1.5 displays a schematic of the pitting process. The pit acts as the anode and surrounding intermetallics undertake the cathodic reactions. Acidic conditions inside the pit prevent repassivation. Surrounding larger cathodic area enables rapid anodic dissolution and propagation of the pit. Released Al^{3+} ions electrostatically attract the Cl^{-} ions from the solution and hydrolyse the water inside the pit. This produces hydrochloric acid inside the pit, which is kept inside the pit as a result of the formed $Al(OH)_3$ cap:

$$Al^{3+} + 3Cl^{-} + 3H_2O \rightarrow Al(OH)_3 + 3HCl \tag{1.14}$$

This cap is impermeable enough to keep the inside of the pit extremely acidic, yet porous enough to allow further Cl^{-} transport inside the pit. The consequence is an autocatalytic process that forms deep cracks propagating through the structure until a rapid and unforeseen mechanical failure [41].

It must be noted that although the potential difference between the phases is the driving force of the corrosion process, the rate of corrosion is determined by the kinetics. A larger potential difference of the micro-galvanic cell does not directly correspond to a higher corrosion rate/current. This has been remarked on by the paper of Birbilis and Bucheit [40], where electrochemical behaviour of intermetallic particles commonly seen in Al alloy microstructures were analysed with a microcapillary electrochemical cell setup. Largest anodic/cathodic currents were not observed in samples with least/most noble corrosion potential values. It was argued that the magnitude of the current of intermetallics at the corrosion potential of the Al alloy is more important than the difference in corrosion potentials of the intermetallics and Al alloy. Additionally, enhanced corrosion is recognised in samples with higher copper content, which demonstrates that simple more/less noble concepts are not enough to describe Al alloy corrosion. It was also observed that size and distribution of the intermetallic particles through the Al ma-

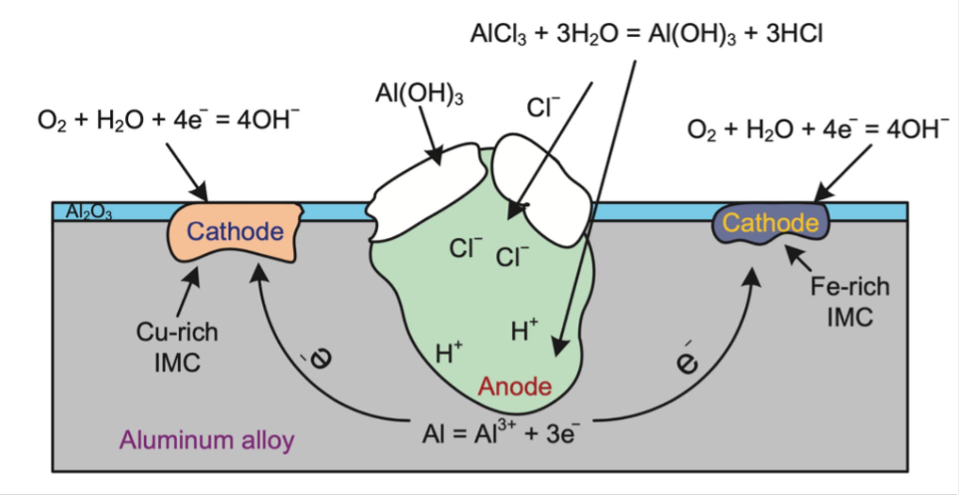


Figure 1.5: Pitting of an Al alloy in chloride-containing aqueous environment [41].

trix matter – for example, dispersoids appear to have a negligible effect on localised corrosion due to their homogenous distribution and small size in comparison to larger constituent particles.

LOCALISED CORROSION OF AA2024-T3

Two main corrosion mechanisms that affect AA2024-T3 in brine environments is intergranular and pitting corrosion. Deaerated anodic polarisation measurements of AA2024-T3 in 1M NaCl environments showed two breakdown potentials where anodic current increases rapidly during polarisation: supported by electron microscope analysis less noble increase in current was attributed to pitting by the transient dissolution of Al_2CuMg, and more noble one the intergranular corrosion attack [53].

In both corrosion mechanisms, intermetallics play an important role by creating micro/nanogalvanic electrochemical interactions among the different intermetallics and Al matrix. Localised corrosion may start from the vicinity of an intermetallic and propagate through the locally most corrosion-prone part of the Al matrix, the grain boundaries, only to continue dissolution of the matrix in the direction of intermetallic clusters buried inside the matrix, potentially emerging from a different part of the alloy surface again [54].

It is now understood that this interaction between different precipitate/constituent intermetallics and grain boundaries is of high importance in explaining the full complexity of the corrosion phenomena of AA2024-T3. A combination of studies that analyse the corrosion of isolated [42, 55, 56] and clustered [51, 57–59] intermetallics demonstrated that understanding of both systems is crucial in understanding the complete corrosion mechanism.

Isolated corrosion attack on intermetallics happens in two different ways. In the case of an anodic particle corrosion progresses with the self-dissolution of the intermetallic. In the case of an initially cathodic particle, circumferential pits and trenching appear around the intermetallic due to dissolution of the surrounding Al matrix. These pits may initiate as metastable pits and stop, or develop into stable pits. It is proposed that a stable pit can form under the condition that the product of pit depth and current density is larger than 10^{-2} A/cm [39]. However, studies show that trenching around isolated particles do not fulfill this condition, and not lead to severe corrosion [37].

A dealloying driven local corrosion mechanism is observed with in/ex-situ analytical TEM for Al_2CuMg and Al_2Cu precipitates [55]. Local corrosion starts with a surface initiation stage, where passive layer covering the alloy destabilises and locally dissolves around the intermetallic. Mg/Al dissolve and hydrolyse from the intermetallics, meanwhile Cu diffuses to the surface rim of the pits. Al_2Cu exhibits a relatively slower corrosion initiation as a consequence of formed $Al(OH)_3$. Nanogalvanic interactions form inside the intermetallic, and Cu rich cathodic sites start producing OH^-. This makes the local surface top of the intermetallics basic. Due to faster dealloying, local chemistry becomes more basic for Al_2CuMg. Increase in pH locally dissolve the surrounding Al oxide passive layer. This triggers trench initiation. Pits start to propagate in depth of the matrix: cathodic ORR take place on the surface of the pits, while the Al matrix around the intermetallic dissolves until the particle is undercut. When the remaining Cu rich intermetallic remnant is undercut, Cu reaches its corrosion potential, dissolves into the solution and gets redeposited on the Al matrix or other intermetallic surfaces. Cu plating can create more cathodic areas and promote further localised corrosion of the surrounding Al matrix. Due to faster kinetics of the Al_2CuMg, Cu ions liberated earlier from its dissolution can be redeposited onto the Al_2Cu precipitates and other intermetallics.

A follow-up TEM study [42] demonstrated similar degradation mechanisms for isolated constituent particles $Al_{76}Cu_6Fe_7Mn_5Si_6$, $Al_7Cu_2Fe(Mn)$. Nano-pits initiate with dealloying attack of active elements Al, Mg, Mn while Cu and Fe rich cathodic zones undertake the reduction of oxygen. Local dissolution rate increases with increasing exposure. Dealloyed zones of the intermetallic become more cathodic, and start the trenching through dissolving surrounding Al matrix. Depth propagation occurs after trench initiation at the area surrounding the intermetallic particles. It is seen that Si inhibits the reaction through oxidising into stable SiO_2, while Mn actively dissolves away. Due to the higher electrochemical stability of constituent particles than the Al_2CuMg precipitate, copper ions released during the Al_2CuMg corrosion may deposit onto the constituent particles in vicinity. This would increase the cathodic activity of intermetallic particles and increase local dissolution.

Figure 1.6 summarises the co-operative corrosion mechanism through intermetallic clusters. As mentioned before, even the most active intermetallic Al_2CuMg cannot reach the stable pit formation current density of 10^{-2} A/cm while isolated. However, in the presence of a cluster of intermetallics this can be supplied through a corrosion attack on a larger scale. Boag et al. [51] have identified that more than a normal amount of intermetallic particles with opposite electrochemical activity were present around stable pits. Study of Hughes et al. [57] supported these findings, and claimed that clustering results in a co-operative corrosion process. During their experiments they have spotted rapidly formed corrosion rings around clusters of intermetallics. They have identified center of the rings to be electrochemically active domes with increased Cl^- presence and H_2 evolution in the area. Furthermore, surface and subsurface attack to the grain boundaries

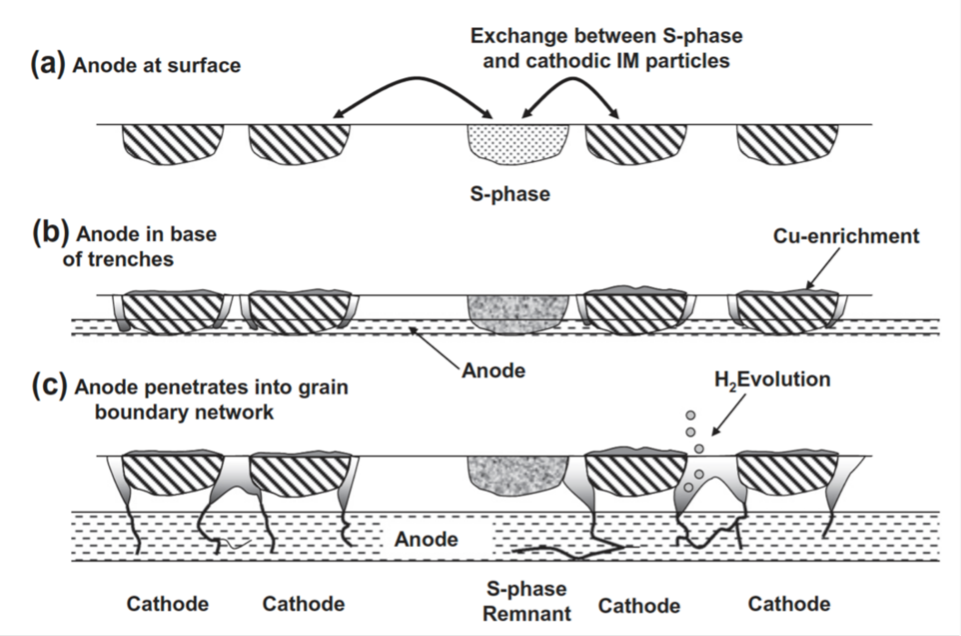

(a) Anode at surface
Exchange between S-phase and cathodic IM particles
S-phase
(b) Anode in base of trenches
Cu-enrichment
Anode
(c) Anode penetrates into grain boundary network
H_2Evolution
Anode
Cathode
Cathode
S-phase Remnant
Cathode
Cathode

Figure 1.6: Co-operative corrosion mechanism: (a) initiation, (b) trenching, (c) propagation [57].

was observed. Intergranular attack propagates through the active grain boundaries, and at a later stage in the case of a surfaced intermetallic corrosion of the isolated intermetallic takes place as described previously [55, 57]. Previous studies have observed a lateral propagation tendency for this pitting and intergranular attack [37, 39].

Figure 1.7 presents an overview of Al intermetallic corrosion time scales. Dispersoid particles $Al_{20}Mn_3Cu_2$ exhibit a similar dealloying driven local corrosion behaviour to the constituent particles, although on a slower time and a smaller length scale [42]. In fact, a recent analysis of AA2024-T3 microstructure with open circuit potential and optical microscopy measurements suggests that all intermetallic phases show similar micro/-nanogalvanic activation, dealloying and trenching behaviour [50]. Main difference between the intermetallic particles was found in the dealloying step – trenching occurred at similar rates independently from intermetallic composition. On the other hand, it is reported that compared to Al-Cu-Mn-Fe phases, Al-Cu-Mg phases account for most of the increase in ORR on AA2024-T3 relative to pure Al [60].

Corrosion protection of aluminium alloys

Corrosion protection is especially important in demanding environments such as the service conditions of the aerospace industry. Different materials must be isolated to prevent galvanic coupling among themselves, and corrosion protection system must be designed with respect to external constraints. Protection must be sustained in various chemical media such as chlorides and de-icing liquids, various humidity conditions such as wet and dry cycles, in large temperature ranges from -50 to 100 °C, during the lifetime of the aircraft, for at least three decades [61].

One of the most widely used and economically viable corrosion protection solutions is the application of functional coatings on the alloy substrate. In this way, surface exposure to the aggressive outside environment is limited with the application of a thin layer of coating. Coatings consist of a set of functional ingredients (anticorrosive pigments, fillers, among others) spread over a host polymer binder. They are applied to the substrate while liquid and after application transform into a solid layer with a thickness comparable to the human hair.

Figure 1.8 shows an example multilayered corrosion protection coating system used in the aerospace industry. A passive barrier protection provided by the polymer paint and anodic oxide layer prevents the ingress of corrosive species to the metal substrate. In the case of a breach this protection is obsolete, and protective properties are lost. To counter that, an active protection is also employed through leaching of anticorrosive inhibitor species from the primer layer and preferential dissolution of the clad layer. Inhibitor species decrease the rate of corrosion through suppression of anodic and cathodic reactions, while clad layer acts as a sacrificial anode.

For the past century, both active and passive corrosion protection mechanisms have heavily relied on hexavalent chromium. Despite significant research efforts since the 1980s, hexavalent chromium remains the benchmark for corrosion prevention, particularly in the aerospace industry [62]. However, its severe health and environmental hazards are now well-documented. Hexavalent chromium is not only highly carcinogenic but also causes irreversible damage to the skin, nose, throat, eyes, and DNA. Its genotoxic effects extend to aquatic and plant life, making it an ecological threat. As a result, there is an urgent need to develop novel, green alternatives for corrosion protection.

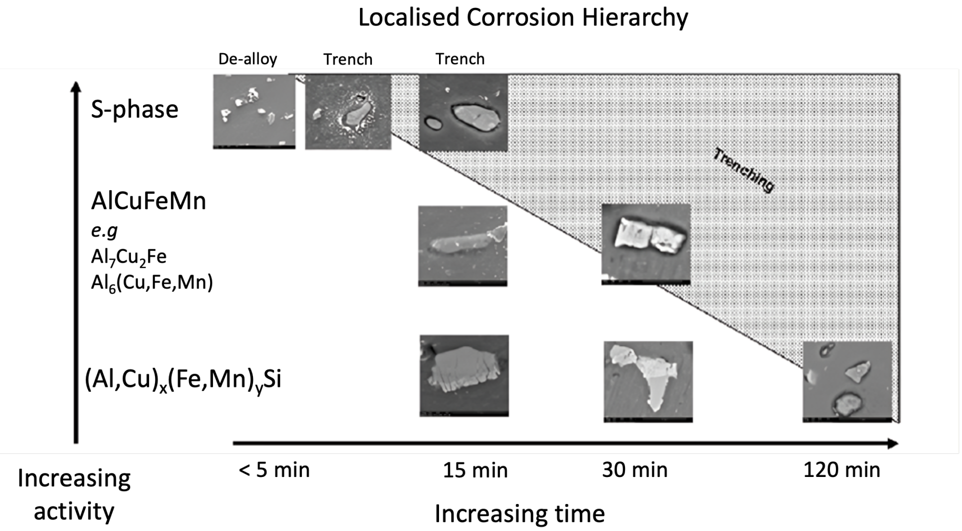


Figure 1.7: Overview of corrosion time scale of various Al intermetallics [56].

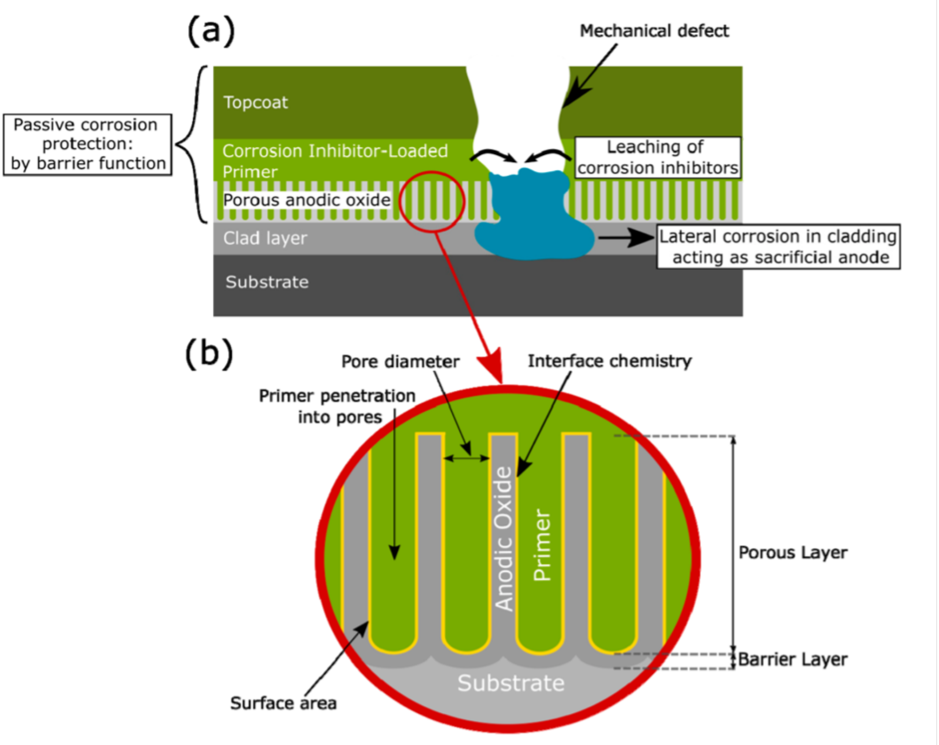

(a)
Mechanical defect
Passive corrosion protection: by barrier function
Topcoat
Corrosion Inhibitor-Loaded Primer
Leaching of corrosion inhibitors
Porous anodic oxide
Clad layer
Lateral corrosion in cladding acting as sacrificial anode
Substrate
(b)
Pore diameter
Interface chemistry
Primer penetration into pores
Anodic Oxide
Primer
Porous Layer
Barrier Layer
Substrate
Surface area

Figure 1.8: A typical corrosion protection system used in the aerospace industry [61]. (a) Cross-section of multiple protective layers, (b) detail of the anodised oxide layer features.

The selection of suitable inhibitors presents a multifaceted challenge influenced by various factors including the pH of the corrosive environment, the presence of aggressive species, the underlying corrosion mechanisms, and the application method of the inhibitor. Depending on whether inhibitors are encapsulated or directly embedded, the encapsulation system, polymer binder material, and other parameters critically shape the final efficacy of the inhibitor.

Achieving superior corrosion protection performance while ensuring environmental compatibility is key in developing next-generation corrosion inhibitors. Understanding the intricate mechanisms by which inhibitors function is the first step in designing effective, sustainable replacements to hexavalent chromium.

1.2.2. CORROSION INHIBITION

ORGANIC MOLECULES AS CORROSION INHIBITORS

ORGANIC corrosion inhibitors primarily function by forming protective films on metal surfaces or by inhibiting anodic or cathodic reactions, or a combination of both. The mechanism of action often involves surface adsorption and the formation of precipitate films [63]. Organic molecules with a high affinity for metals generally exhibit good corrosion inhibition.

Corrosion inhibitors are often categorised as anodic and cathodic inhibitors [64]. Anodic inhibitors slow down the anodic metal dissolution process and produce sparingly soluble reaction products that form protective films over anodic sites. This behaviour can be determined from potentiodynamic polarisation plots. Based on the mixed potential theory, the retarded anodic half-cell reactions would suppress measured anodic curves to lower current densities. If cathodic reaction rates are not affected, the anodic and cathodic curves would now intersect at higher potential values compared to the original situation, shifting the corrosion potentials to more positive values. By modifying the anodic reaction, these inhibitors effectively reduce the rate of metal degradation. Cathodic inhibitors, on the other hand, disrupt cathodic oxygen reduction reactions and promote the formation of reaction products that selectively precipitate at cathodic sites. Polarisation curves of metals treated with cathodic inhibitors show a shift in the corrosion potential to more negative values. Mixed inhibitors, which suppress both anodic and cathodic reactions, maintain the corrosion potential while significantly reducing the corrosion current.

Mechanism of inhibition

Organic molecules inhibit corrosion by adhering to the oxide or metal surface through physisorption or chemisorption, often forming chelates. These adsorbed layers act as protective barriers for metal surfaces. Physisorption occurs through electrostatic interactions between partially charged regions of the molecules and the charged surface, as well as through hydrogen bonding and Van der Waals forces. Chemisorption, on the other hand, involves charge sharing or charge transfer from the molecule to the surface or vice versa. In some cases, electrostatic interactions contribute significantly to the adsorption energy, similar to chemisorption [65].

Although some inhibitors predominantly operate via a single mechanism, many organic inhibitors exhibit multiple modes of interaction with metal surfaces. The inclusion of heteroatoms such as sulfur (S), nitrogen (N), and oxygen (O) into organic ring structures and branching ligands enhances their binding capacity. These atoms, with lone electron pairs, facilitate chemisorption through their interaction with the electronic structure of the inhibitor molecule. The electron density around these heteroatoms and overall electronic configuration of the molecular structure significantly influence the adsorption behaviour [66].

The nature of chemisorption in metals differs based on their electron configuration. Aluminium, with a vacant p-orbital, is electron-deficient and can act as a Lewis acid during chemisorption. Lone pair donation from heteroatoms such as S and N can result in strong covalent bonds between the Al surface and organic molecule [66]. S may form stronger bonds with Al surfaces due to its higher atomic polarisability (S = 2.90, N = 1.10 [67]). In transition metals, the vacant d-band allows parallel chemisorption via strong π-d orbital hybridisation or perpendicular chemisorption through the σ-molecular orbitals of unsaturated heteroatoms. For copper, which has fully occupied d-orbitals, the expected bonding is a comparatively weaker chemisorption through σ-molecular orbitals [68]. Chemisorption between inhibitors and Cu can be through donation of free electron pairs from the inhibitor to unoccupied orbitals of Cu, or through π-backbonding where

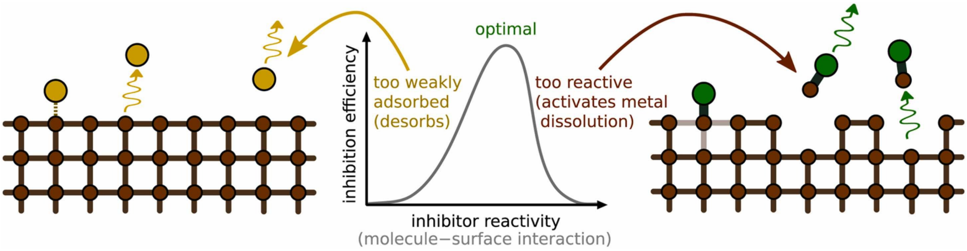

optimal
inhibition efficiency
too weakly adsorbed (desorbs)
too reactive (activates metal dissolution)
inhibitor reactivity
(molecule–surface interaction)

Figure 1.9: Illustration of a corrosion inhibition concept analogous to the Sabatier principle of heterogenous catalysis [76]. If the interaction between the inhibitor and the surface is too weak, the inhibitor molecules desorb quickly, failing to provide effective protection. Conversely, overly strong interactions can destabilise adjacent metal–metal or metal–oxygen bonds, instead facilitating metal dissolution. Thus, the optimal inhibitor–surface interaction is a balanced one: neither too weak nor too strong.

filled d-orbitals of Cu can donate electrons to bond to the vacant orbitals on the adjacent molecule. Functional groups such as carbonyls and their analogues with C double bonded heteroatoms can act as retrodonation ligands [69], which is a common structural trend in high performing inhibitors. Compared to Al, due to larger d-electron cloud and therefore increased van der Waals forces of Cu, increased physisorption would be expected. Considering not metallic surfaces but their (hydr)oxides, the nature of interactions would change significantly due to the ionic/covalent nature of the oxides and the availability of surface bonding sites. Surface hydroxyls can engage in hydrogen bonding or proton exchange, and oxygen vacancies can act as active sites, enhancing adsorption through charge redistribution [70–73]. One big difference is that Cu oxides are redox active, but Al oxides are not (see Pourbaix diagrams of both [74, 75]). The redox activity of Cu can facilitate oxidation or reduction reactions with adsorbed molecules through Cu^{2+}/Cu^{+} redox reactions. For example, in the presence of Cu^{2+} reduction, sulfur compounds may oxidise to form sulfates, and nitrogen compounds may undergo oxidation to form nitrates.

One key idea to highlight is that a stronger molecule–substrate bonding does not automatically result in better inhibition - otherwise corrosion inducing species such as chloride ions, which interact very strongly with metals, would act as corrosion inhibitors. The inhibitor should adsorb strongly enough to persist on the surface, but not too strongly or else it can promote metal dissolution because too strong molecule–metal interaction can weaken the neighbouring lattice metal–metal and/or metal–oxygen bonds [76]. Figure 1.9 demonstrates this principle.

In aqueous environments, adsorption can be understood as a substitution reaction involving competing interactions between molecules, water, metal/oxide, corrosive ions, and contaminants. The adsorbed inhibitor layer physically blocks the metal surface or increases ionic resistivity, preventing aggressive ions from initiating corrosion. Additionally, this layer electrochemically inhibits corrosion by decelerating anodic and/or cathodic reactions. Adsorption can result in a continuous anodic passivation layer or selective coverage of cathodic zones, effectively mitigating corrosion processes. A commonly proposed corrosion inhibition reaction mechanism for Cl^{-} containing environments progresses by the displacement of adsorbed Cl^{-} from the surface of the inhibited metal. Metallic cation-adsorbed chloride complexes react with inhibitor molecules present in the electrolyte to form metal-inhibitor complexes[47]:

$$M(Cl)_{n_{ads}} + Inh_{sol} \rightarrow M\text{-}Inh + nCl^{-}_{sol} \tag{1.15}$$

For high-performing inhibitor molecules, a compact water displacing barrier is formed as a result of this reaction. The separation of metal surface from the aggressive environment inhibits further corrosion.

Environmental factors and inhibitor concentration play critical roles in determining the performance and mechanism of organic inhibitors. The pH of the environment affects the speciation of ionisable organic molecules, with low pH conditions often hindering adsorption due to protonation [68]. In contrast, higher pH values result in deprotonated molecules, which results in stronger adsorption and improved protection [77]. Inhibition efficiency typically increases with concentration up to a critical threshold, beyond which it plateaus or in some cases may decline. This phenomenon is often attributed to oligomerisation of the inhibitors in solution, reducing their availability for adsorption, or the formation of oligomers that desorb from the surface [78]. Furthermore, exothermic adsorption reactions often result in decreased inhibition efficiencies at elevated temperatures. Time-dependent changes in the environment-interface interactions also influence the effectiveness of inhibitors [79].

For a more detailed analysis of various specific corrosion inhibition mechanisms, please refer to our recent review.[1]

1.3. Research Aim and Approach

As discussed in the previous sections, we need a hexavalent chromium replacement, and fast. The aim of this dissertation was to bring the fourth paradigm of science to this quest of corrosion inhibition discovery. In these last four years I have worked with Li, Ce, and Cr based systems as well, but the majority of my time was focused on the organic compounds. I tried to understand why they work, the common structural motifs in the compounds that do work, and tried to disentangle multiple free parameters from one another.

I started by creating a molecule library, as there were inconsistencies between different papers reporting the corrosion inhibition performance of organic molecules. **Chapter 2** presents the screening experiments, where I collected data on corrosion inhibition properties of more than 100 molecules (which are also visualised in the Mol-dex of the appendix) on AA2024-T3 substrates, and quantified the inhibitor performance using time-resolved electrochemical measurement methods of linear polarisation resistance (LPR), electrochemical impedance spectroscopy (EIS), and potentiodynamic polarisation (PDP/LSV). After identifying the best electrochemical target for training machine learning models on, a dummy model was trained (by my colleague) to show that predictive features can be captured from a smaller piece of this dataset (around 50 molecules).

[1] Winkler, D.A., Hughes, A.E., Özkan, C., Mol, A., Würger, T., Feiler, C., Zhang, D. and Lamaka, S., 2024. *Impact of inhibition mechanisms, automation, and computational models on the discovery of organic corrosion inhibitors.* Progress in Materials Science, p.101392. [1].

The screening experiments were performed at identical conditions: same inhibitor concentration, salt concentration, among other parameters. In **Chapter 3**, I wanted to capture the most important inhibitor-related factors that can impact electrochemical experiments. I believed this would increase trust in future corrosion inhibitor discovery experiments of our field, as right now every week there is a new chromate-vanquisher compound coming out that just doesn't work - possibly because the researchers ignore the effect of time, concentration, or other factors. I focused on five factors: influence of inhibitor concentration, exposure time to inhibitors, differences in inhibitor performance in the presence and following absence of inhibitors in the environment, inhibitor performance change with changing external potentials, and synergy - the combined corrosion inhibition effect of multiple molecules that surpass their individual performance. A comparison of electrochemical performance of selected highly-inhibiting organic molecules with sodium dichromate turned out to be interesting, where at the end a system that can rival the electrochemical performance is presented with a higher confidence.

I diagnosed sustaining inhibition in the absence of organic molecules with **Chapter 4**, which focused on the molecule 3-amino-1,2,4-triazole-5-thiol, and its unique ability of creating quasi-irreversible bonds. Attenuated total reflectance Fourier transform spectroscopy (ATR-FTIR), shell-isolated nanoparticle-enhanced Raman spectroscopy (SHINERS), X-ray photoelectron spectroscopy (XPS), time-of-flight secondary ion mass spectrometry (ToF-SIMS), atomic force microscopy (AFM), and scanning Kelvin probe force microscopy (SKPFM) was combined to understand the root-cause of the sustained inhibition behaviour during the absence of molecule in the environment. Through combining the results from different spectroscopies I proposed a molecule mechanism that might cause this behaviour, to be used as a molecular fragment that would enable sustained inhibition.

With **Chapter 5**, I explored whether we can gain scientific insight from small data. After training more than 12 thousand machine learning model configurations of different feature and target representations based on the previous screening data, I identified different ways to extract scientific insight from algorithmic feature selection of the most predictive models, by: *(i)* visualisation of selected features into molecule fragments, *(ii)* using Bayesian optimisation as a tool to extract molecules "thought" to be best by the model from the black-box model decision-making process, *(iii)* combining SHAP-analysis on models with different featurisations to find common structural motifs, which were used to come up with a corrosion inhibition template.

Chapter 6 combined the overall conclusions and outlook.

2

Causing Inhibition

Nature does not 'know' what experiment a scientist is trying to do.
God loves noise as much as the signal.

Lew Branscomb

Meten is weten.

Dutch proverb

Creating durable, eco-friendly coatings for long-term corrosion protection requires innovative strategies to streamline design and development processes, conserve resources, and decrease maintenance costs. In this pursuit, machine learning emerges as a promising catalyst, despite the challenges presented by scarcity of high-quality datasets in the field of corrosion inhibition research. To address this obstacle, we have created an extensive electrochemical library of around 80 inhibitor candidates. The electrochemical behaviour of inhibitor exposed AA2024-T3 substrates was captured using linear polarisation resistance, electrochemical impedance spectroscopy, and potentiodynamic polarisation techniques at different exposure times to obtain the most comprehensive electrochemical picture of the corrosion inhibition over a 24 hour period. The experimental results yield target parameters and additional input features that can be combined with computational descriptors to develop quantitative structure-property relationship (QSPR) models augmented by mechanistic input features.

This chapter has been published with some adjustments as *Laying the experimental foundation for corrosion inhibitor discovery through machine learning*, npj Materials Degradation (2024) [80].

2.1. Introduction

Corrosion inhibition research has come far since Chyżewski and Evans first categorised sparingly soluble corrosion decreasing substances as anodic and cathodic inhibitors [81]. Thanks to the advances in computational power and methods, we are observing a paradigm shift in how science is done, and this is also affecting corrosion inhibition research.

There are four contemporary paradigms of science [82, 83]. The first is empirical evidence, leading to general laws through 'trial and error'. The second involves theoretical models based on those laws. The third is defined by computational power offered by Moore's law, application of theoratical models to more complex and specific problems. This results in a data explosion, leading to the fourth paradigm: data-driven scientific discovery - such as using machine learning for categorisation and prediction.

We see examples of this paradigm shift in corrosion inhibitor research in two broad categories: mechanistic and statistical research. Lately, advances in surface analysis, electrochemical characterisation and computational methods have been complementing each other to facilitate the inhibitor discovery process for both of these categories.

On the mechanistic end, a deeper scientific understanding is obtained by controlled experiments and computational models. The critical need for the protection of aerospace aluminium alloys has driven the research that would uncover AA2024-T3 corrosion inhibition of many compounds. Throughout the years, AA2024-T3 corrosion inhibition mechanisms were experimentally uncovered for inorganic compounds such as chromates [62, 84–86], rare-earths [87–90], molybdate [91] and cobalt ions [92], magnesium-based pigments [93–95], lithium salts [96–98], and a vast variety of organic compounds such as imidazole [99, 100], triazole/thiazole[101, 102], quinoline [103, 104], carbamate [105], thiosemicarbazone [106] derivatives, among others [103, 107–111]. In addition to uncovering the mechanisms for specific inhibitor species, the physical features of inhibition mechanisms such as the importance of time [79, 112] and irreversibility [113] have been investigated.

The pressing demand for novel chromate-free corrosion inhibitors has created the need for high-throughput inhibitor screening methodologies. The approaches inspired by pharmaceutical drug discovery research spanned optical image analysis [114, 115], fluorometric detection [116], multi-electrode electrochemical evaluation [114, 117–119], surface copper enrichment analysis [120], hydrogen evolution detection [121, 122], weight-loss measurements [107, 123], and spectroscopic element analysis through multi-channels [124]. These methods rapidly created large datasets, but with the trade-off of losing mechanistic information.

The third paradigm supported the mechanistic understanding gained from experiments with computational models that span continuum to atomistic scales. Finite element method (FEM) models produced previously unattainable information - such as mechanical strains observed for inhibitor dissolution and leaching from coatings [125], local critical pH criteria for pit repassivation [126], and the effect of surface geometry on electrochemical behaviour [127]. Density functional theory (DFT) and molecular dynamics (MD) simulations have introduced a vast amount of quantum mechanical/-chemical information that is not directly available from empirical methods, such as density of states, band gap, and other physicochemical electronic properties [128]. The ease of investigation of atomistic properties offered by software/hardware advances has allowed corrosion scientists to replace costly and time consuming experiments. Molecular modelling was used as a computational microscope to expose the underlying mechanisms of inhibitor structure-substrate sorption phenomena [71, 128–132]. Recent papers [133–135] have reported on how experimental and computational methods are catalysing one another to combine the strength of empirical and theoretical methods, in which researchers have analysed the influence of type and length of backbone chains and anchor groups on inhibitor performance by combining carefully controlled experiments with DFT modelling.

The accumulated mechanistic understanding of inhibitors, high-throughput methodologies and FEM/DFT-MD computational approaches generated previously unavailable large datasets about mechanical and physicochemical behaviour of inhibitors, which paved the road for data-driven statistical investigations. This involved classification and predictive analytics of inhibitors. Properties of inhibitor molecules obtained from DFT calculations, and experimental inhibitor efficiencies gathered from high-throughput methods have been combined to build correlations using machine learning based quantitative structure-property relationships (QSPR). Winkler et al. [136] used QSPR to reveal empirical molecular descriptors most relevant for AA2024 and AA7075 inhibition, and identified that chemical descriptors solely using input features obtained from *in vacuo* DFT did not contain sufficient information to generate predictive models. Würger et al. [32, 137] have demonstrated a data-driven inhibitor prediction workflow for magnesium alloys, which combined the results of atomistic simulations and high-throughput experiments with unsupervised machine learning clustering algorithms and supervised learning approaches to predict the behaviour of untested inhibitors. Feiler et al. [122] have demonstrated that the combination of structural information with input features derived from DFT lead to robust predictive models for corrosion inhibition responses of small organic molecules based on an artificial neural network for pure magnesium, as well as Mg-based alloys [138]. The optimisation of machine learning approaches is an ongoing process, whether it is coming up with better methods of identifying the most relevant molecular descriptors [138], or analysis of different inhibitor classification algorithms and creation of new descriptors with intrinsic mechanistic meanings [139].

All in all, *in silico* inhibitor screening combined with smart high-throughput testing has enabled overcoming physical limitations of previous paradigms. However, a complete jump to the fourth paradigm will require a strong empirical foundation. A recent review of Coelho et al. [35] has identified the main challenge of utilising machine learning for corrosion research as the lack of high-quality datasets. Corrosion datasets are found to be typically noisy, rarely shared in a systematic machine-readable way, and lacking in time-dependent multidimensional input, which was shown to increase the accuracy of studied models. On the one hand, recent inhibitor data management initiatives such as CORDATA database [140] introduced open-source philosophies to the inhibitor discovery and selection - however although database contains hundreds of entries, inhomogeneous data is still a problem. The database contains data acquired on different raw batches of alloys, different or poorly controlled ambient temperatures, and different experimental methods and conditions. On the other hand, dedicated state-of-the-art high-throughput datasets for aluminium alloys have created data for hundreds of organic compounds [107, 114, 136, 139]. However, lack of multidimensional input is a distinctive shortcoming of high-throughput methods, where only one parameter is collected to represent the inhibition performance. For an alloy prone to localised degradation, such as pitting corrosion of AA2024-T3, a data creation procedure that obtains information on both the open circuit state as well as behaviour under applied potentials is crucial to get the full mechanistic picture.

We aim to address the need for a robust multidimensional time-dependent electrochemical database with this study. We also show the best practices for applying this multidimensional data to train a predictive machine learning model. AA2024-T3 samples exposed to around 80 small organic molecule containing electrolytes are electrochemically characterised through linear polarisation resistance, electrochemical impedance spectroscopy and potentiodynamic polarisation. The goal of this brute force 'high-throughput' approach that combines proven electrochemical methods is to demonstrate a methodology to create robust data that contains mechanistic time-dependent information. Gained mechanistic information spans double layer capacitance, charge transfer resistance, diffusion of corrosive ions through a protective inhibitor layer from electrochemical impedance spectroscopy, time-resolved corrosion resistance response from linear polarisation resistance, and corrosion rate, potential, breakdown potential, the kinetics of the electrochemical reactions and nature of anodic and cathodic reactions at biased electrical potentials from potentiodynamic polarisation. The obtained experimental parameters can be employed directly as target parameters for training a machine learning model that is predictive of the performance of untested compounds to create a shortlist of promising candidates. Moreover, the experimental investigation yields additional input features that can be combined with molecular descriptors derived from the molecular structure and atomistic simulations. These input features exhibit great potential to develop augmented quantitative structure-property relationships as they allow the direct inclusion of information of the underlying mechanisms in the model training. The results of this study are expected to support the development of faster inhibitor screening techniques in the future, which can leverage the link between the molecular structure of the inhibitor and its corrosion inhibition activity.

2.2. Methods

2.2.1. Sample preparation

Aluminum alloy 2024 with a T3 temper (AA2024-T3) in the form of 2 mm thick sheets is purchased (from Salomon's Metalen B.V., the Netherlands) to perform the electrochemical experiments. The chemical composition of the alloy measured by the supplier in accordance with the ASTM-E1251 standard is provided in Supplementary Table 2.4.

The sheets were cut with an automatic shearing machine to dimensions of 20 mm x 20 mm samples. The samples were mechanically ground on a rotating plate polisher under a stream of water using Struers waterproof SiC sandpapers with progressively finer grits of 320, 800, 1200, 2000 and 4000. Subsequently, the samples were polished using a fine diamond suspension (Struers DiaDuo-2) with 3µ and 1µ particle sizes. After the polishing procedure, samples were cleaned with isopropanol in an ultrasonic bath (EMAG-EMMI 30HC) for 15 minutes and dried with compressed air. Sample preparation resulted in a mirror-like surface finish.

2.2.2. INHIBITORS & ELECTROLYTES

THE salt solutions without the addition of inhibitors (pH 5.9) were prepared with NaCl powder with Milli-Q pure water (15.0 MΩ cm resistance at 25 °C). For inhibitor containing solutions, inhibitors in quantities corresponding to 1 mM concentrations were also added during the mixture step. No additional compounds were added to modify the pH and/or increase the solubility of inhibitors. 78 small organic molecules were tested as corrosion inhibitors, resulting in 0.1 M NaCl - 1 mM inhibitor electrolytes.

Initial organic molecule choice was based on previous inhibitor screening studies [107, 136]. Tested organic molecules had both aromatic/aliphatic moieties of thiol, amino, carboxyl and hydroxyl groups. CAS numbers and common names of the compounds are presented in the section appendix B. All chemicals were purchased from Sigma-Aldrich, with the exception of sodium chloride (J.T. Baker), 3-amino-5-mercapto-1,2,4-triazole, lithium nitrate, cerium carbonate hydrate (Alfa Aesar), cerium chloride heptahydrate, sodium acetate (Fluka), 2-mercaptobenzoate (Thermo Fisher Scientific), 5-mercapto-1-phenyl-1H-tetrazole (TCI Chemicals) and sodium mercaptobenzothiazole (Apollo Scientific). Almost all inhibitors dissolved fully in 1 mM concentrations, with the exception of thiosalycylic acid, 2-mercaptobenzothiazole, α-benzoin oxime, 2,2'-dithiodibenzoic acid, 4-mercaptobenzoic acid, 2-(2-hydroxyphenyl)benzothiazole, quercetin hydrate, berberine chloride hydrate and 2-(2-hydroxyphenyl)benzoxazole. The solutions of these compounds were either murky, resulted in muddy suspensions/emulsions or had visible undissolved particles in the solution. The pH of the resulting solutions were measured with Metrohm 913 pH meter, before and after the electrochemical experiments.

2.2.3. ELECTROCHEMICAL EXPERIMENTS

ELECTROCHEMICAL measurements were conducted at room temperature in open-to-air 0.1M NaCl solutions, with (or without) the added 1mM inhibitor candidates. A conventional three-electrode electrochemical cell (flat corrosion cell, Corrtest Instruments, China) with the sample as the working electrode, platinum mesh as the counter electrode, and Ag|AgCl (saturated KCl) as the reference electrode were used to perform the experiments. The designated electrolyte volume was 300 ml and the exposed surface area was 0.785 cm^2 (1 cm diameter circle). Electrochemical measurements were controlled with Biologic VSP-300 multichannel potentiostats through EC-Lab software (version 11.33, Biologic, France).

The electrochemical measurements consisted of three different techniques commonly used in the field of corrosion science: linear polarisation resistance (LPR), electrochemical impedance spectroscopy (EIS) and potentiodynamic polarisation (PDP). The electrochemical investigations were initialised after observing the open circuit potential (OCP) for 10 minutes. LPR was measured over a potential range of ±10 mV with a scan rate of 0.5 mV s^{-1} every 10 minutes for 24 hours. The polarisation resistance (R_p) values were calculated by applying a linear fit to the observed linear region of potential vs. current density plots. EIS measurements were conducted at the 2nd and 24th hour. EIS measurements were conducted by applying a sinusoidal AC perturbation with a peak-to-peak amplitude of 10 mV in the 10 kHz - 10 mHz frequency range with 10 frequency point per logarithmic decade with 3 repetitions per frequency point. OCP was observed in between LPR and EIS measurements. After the EIS at the 24th hour, potentiodynamic polarisation curves are recorded in a single sweep with a scan rate of 0.5 mV s^{-1} from -250 mV cathodic to +250 mV anodic potentials with respect to open circuit potential. Corrosion potentials and current densities were calculated with Tafel extrapolation, by obtaining the intersection of tangents from linear parts of anodic and cathodic curves of the log|current density|-potential polarisation curves. Visual summary of electrochemical experiments is presented in Supplementary Figure 2.8.

All electrochemical experiments were repeated at least three times per inhibitor to ensure the reproducibility of the experiments.

2.2.4. Molecular descriptor generation, feature selection and evaluation of random forest models

The molecular descriptors based on the structure of the molecules for the input to the random forest (RF) model, e.g. the molecular weight or the number of certain functional groups, have been generated using the open source chemoinformatic software package RDKit [141]. Additionally, DFT computations have been carried out to determine electronic key properties like frontier orbital energy levels using the commercial software package Turbomole [142] resulting in a pool of 216 molecular descriptors (208 structural, 7 derived from DFT simulations and 1 experimental parameter (the average pH, average of before and after electrochemical measurements). The aim of the recursive feature elimination (RFE) was, to reduce this number to five or ten input features. Furthermore, experimental parameters, especially the average pH, which were obtained from the experiments, were used as additional input to the ML model. To determine the influence of DFT and experimental parameters, the RF has been trained on different sets of input features: on the structural features only, on the structural features complimented by DFT or experimental parameters or both.

Prior to training, RFE, a sparse feature selection approach based on RF, has been carried out to select the most pertinent input features. The purpose is to select *n*-tuples of features that perform well together. Features that have low or no relevance to the modelled property would degrade the model and using too many input features will ultimately lead to overfitting on the training data. Therefore, the five and ten most relevant features in each of the four groups have been determined with RFE and subsequently used as the input to the RF model.

RF is a supervised learning method where the output is obtained by averaging the results of a set of decision trees. The RF model can use both the IE and IP as targets. Examining the data distribution for IE and IP (see Supplementary Figure 2.13), it can be observed that there is no uniform distribution in either case which may lead to an unintentional bias in the training data. Preprocessing step consisted of removal of minimally varying and highly correlated features, and scaling the rest. Features with variance lower than 0.1 have been removed with the VarianceThreshold function of scikit-learn.

Features with correlations higher than 0.8 to rest of the features are dropped. All features have been scaled using MinMaxScaler of scikit-learn. For the implementation of RF models in this work, the default parameters provided by scikit-learn have been utilised.

To evaluate performance of the models, the coefficient of determination (R^2) and the root mean squared error (RMSE) have been employed. The first step was to divide the data into a training and test set, with the test set containing ten molecules, or roughly 17 % of the total number of molecules in the dataset. To be more confident in the models performance, in the next step a CV approach has been used to assess the models robustness. For this purpose, the dataset was split into six different folds using the KFold function of sci-kit learn and all folds but one are used for training the models; this fold is held back and used as the test set. Each fold also contained roughly 10 % of the total number of molecules in the dataset. In total, the models are trained six times and the average of the errors is calculated to assess their robustness.

Unless otherwise stated, the error bars and bracketed values (±e.g.) presented throughout the study represent the standard error.

2.3. RESULTS AND DISCUSSION

2.3.1. EXPERIMENTAL RESULTS

FIGURE 2.1 plots the potentiodynamic polarisation (PDP), electrochemical impedance spectroscopy (EIS), and linear polarisation resistance (LPR) measurements of AA2024-T3 samples exposed to 0.1 M NaCl solution with and without the presence of 1 mM inhibitor candidates of benzotriazole, 2,5-dimercapto-1,3,4-thiadiazole, 2-mercaptobenz-imidazole, 2-mercaptobenzoate, sodium acetate, sodium mercaptoacetate, or ammonium pyrollidinedithiocarbamate. The summary of values obtained from the experiments is presented in Table 2.1. In order to showcase the broad spectrum of behaviours observed in the electrochemical experiments, inhibitor candidates with contrasting characteristics were selected.

Figure 2.1 (a) presents polarisation curves of AA2024-T3 samples recorded after 24 hours of immersion in inhibitor containing solutions. Polarisation curves show that the addition of small organic molecules results in corrosion current densities varying up to 2 orders of magnitude . It is noteworthy that the best inhibitor candidates reduced the corrosion current densities more than 10-fold compared to the uninhibited samples. Analysis of corrosion potentials shows that inhibitors act as mixed or anodic inhibitors. Anodic inhibitors reduce the current densities of partial oxidation reactions without affecting the partial reduction reactions, causing the shift of the corrosion potential in the positive direction (vice versa for cathodic inhibitors) [65]. Albeit small, addition of organic molecules shift the corrosion potentials to more positive values, with the exception of ammonium pyrollidinedithiocarbamate. However when breakdown potentials (potentials where a sudden increase in current for the anodic curves) are observed it is seen that the introduction of molecules resulted in negligible shifts with the exception of 2,5-dimercapto-1,3,4-thiadiazole. The distribution of electrochemical potentials among all inhibitor candidates is analysed more deeply in section 2.3.5.

Figure 2.1 (b) shows the EIS impedance Bode modulus plots after 24 hours of immersion in inhibitor-containing solutions. The impedance modulus |z| values observed at

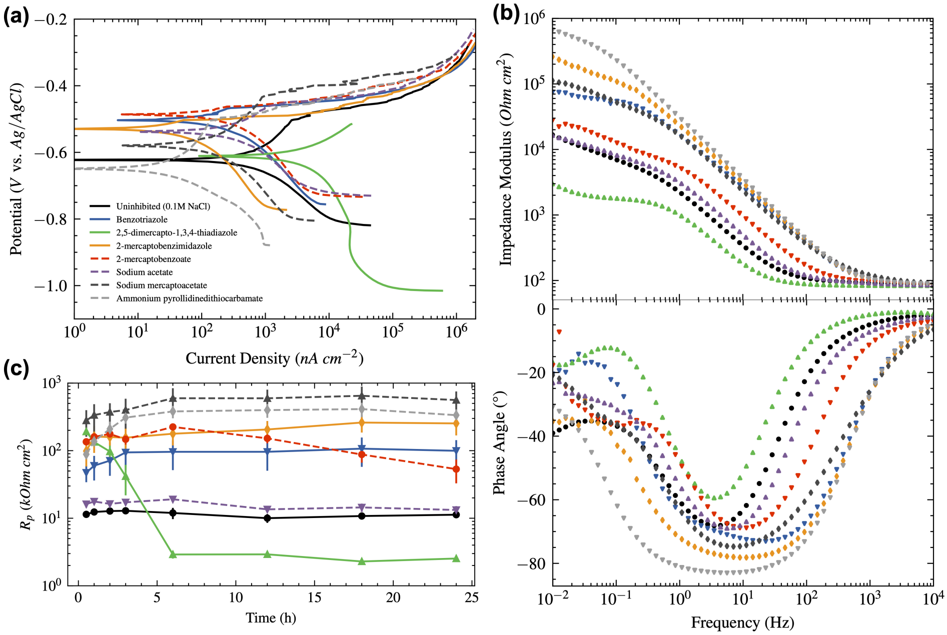


Figure 2.1: AA2024-T3 samples exposed to 0.1 M NaCl solution in presence and absence of 1 mM inhibitors. (a) Potentiodynamic polarisation curves and (b) electrochemical impedance spectroscopy Bode modulus and phase angle plots recorded after 24 hours of immersion, (c) linear polarisation resistance R_p values as functions of exposure time.

10^{-2} Hz frequency are treated as the R_p values calculated from EIS, as it was shown that it reflects the corrosion resistance of the inhibitor-substrate interface [143]. This approach is based on a simplification, since the low frequency impedance modulus includes contributions from the oxide film resistance, the charge transfer resistance, and often from the diffusion controlled processes. Moreover in addition to the real component, it includes the imaginary part. |z| values show more than a 2-orders of magnitude range as it was seen for corrosion current density measurements. Corrosion resistance with respect to the uninhibited samples showed more than a 30-fold increase. A comparison of low frequency impedance modulus values observed at 2nd and 24th hours presented in Table 2.1 show significant variation in inhibitor behaviour. This change from 2nd to 24th hour is more clearly observed in LPR plots, which correspond well with EIS results.

Figure 2.1 (c) shows estimated R_p results calculated from the LPR measurements conducted throughout 24 hours. The instantaneous corrosion resistance of a system can be indirectly assessed by measuring the polarisation resistance R_p. A higher R_p indicates a more resistive interface between the electrode and the electrolyte. The resistive interface hinders the flow of electrons and ions, increasing the corrosion resistance [144]. From the LPR measurements it is clear that the action of inhibitor species is highly time- and species-dependent. In some cases such as sodium acetate, there is negligible change in behaviour compared to the uninhibited solution. However in most cases, it

Table 2.1: Electrochemical information obtained from potentiodynamic polarisation, electrochemical impedance spectroscopy, and linear polarisation resistance measurements of AA2024-T3 samples exposed to inhibitor containing solutions. Corrosion current density j_{corr}, corrosion E_{corr} and breakdown E_{br} potentials vs. Ag|AgCl, impedance modulus values $|z|$ observed at 10^{-2} Hz evaluated for 2 and 24 hours, linear polarisation resistance R_p evaluated at 24 hours and the time-weighted average of the measurements $\langle R_p \rangle$ are presented.

Inhibitor	j_{corr}	E_{corr}	E_{br}	$\|z\|_{2h}$	$\|z\|_{24h}$	$R_p\|_{24h}$	$\langle R_p \rangle$
	(nA cm^{-2})	(mV)	(mV)	(kΩ cm^2)	(kΩ cm^2)	(kΩ cm^2)	(kΩ cm^2)
Uninhibited (0.1M NaCl)	604 (±108)	-620 (±12)	-486 (±2)	14 (±0)	14 (±3)	11 (±1)	11 (±1)
Benzotriazole	216 (±38)	-500 (±3)	-475 (±4)	79 (±31)	107 (±48)	100 (±44)	94 (±43)
2,5-dimercapto-1,3,4 thiadiazole	3822 (±399)	-604 (±6)	-479 (±6)	51 (±18)	3 (±0)	3 (±0)	16 (±4)
2-mercaptobenzimidazole	79 (±18)	-523 (±6)	-496 (±8)	80 (±30)	265 (±80)	253 (±85)	207 (±66)
2-mercaptobenzoate	261 (±65)	-527 (±16)	-472 (±19)	130 (±50)	38 (±16)	53 (±21)	135 (±7)
Sodium acetate	396 (±54)	-563 (±14)	-473 (±13)	16 (±2)	16 (±1)	13 (±2)	15 (±1)
Sodium mercaptoacetate	57 (±13)	-572 (±25)	-435 (±34)	203 (±47)	203 (±64)	561 (±191)	555 (±205)
Ammonium pyrollidinedithiocarbamate	38 (±4)	-636 (±14)	-488 (±14)	346 (±34)	480 (±106)	335 (±73)	356 (±173)

was observed that instead of having a constant behaviour, R_p values evolve with time. In cases such as benzotriazole, 2-mercaptobenzimidazole, sodium mercaptoacetate and ammonium pyrollidinedithiocarbamate, there is an initial increase in R_p, and further development of corrosion protection until the 6th hour, and stable corrosion protection after that. For 2-mercaptobenzoate it was seen that after an initial increase and a gradual development of corrosion resistance, the protection started to decrease to lower than initial values. For 2,5-dimercapto-1,3,4-thiadiazole it was seen that after the initial, more than an order of magnitude increase in R_p, the protection starts to decrease. This decline continues until the 6th hour and signifies stable active corrosion behaviour afterwards.

In the specific case of 2,5-dimercapto-1,3,4-thiadiazole, we conclude that this accelerated corrosion was caused by the pH change of the electrolyte after the introduction of the inhibitor. Analysis of pH measurements of the electrolytes prior to the electrochemical experiments shows that compared to the pH value of 6 of the uninhibited 0.1 M NaCl solution, 2,5-dimercapto-1,3,4-thiadiazole containing solution had an acidic pH value of 3. This is at the boundary of the thermodynamically stable region of Al at 1M Al^{3+}, but in the region of preferential stability of Al^{3+} at lower than 1M contraception of Al^{3+} [145], which is expected for OCP corrosion of AA2024. Therefore the considerable decrease in pH must have disrupted the stable aluminium (hydr)oxide layer, and lead to active corrosion of the samples.

Due to this dynamic corrosion and inhibition behaviour, it is vital to capture the performance during the whole time-span. One method to achieve this is to estimate the mean value of R_p through a trapezoidal integration over time:

$$\langle R_p \rangle = \frac{1}{t_f - t_0} \int_{t_0}^{t_f} R_p(t) dt \tag{2.1}$$

$$\approx \frac{1}{t_f - t_0} \sum_{k=1}^{N} \frac{R_p(t_{k-1}) + R_p(t_k)}{2} (t_k - t_{k-1}) \tag{2.2}$$

where t_f is the final measurement time, t_0 is the initial measurement time, and k is the indice for the performed discrete measurements. The mean estimated this way can be used as a screening metric that contains all time-dependent information in one number. The power of this approach as an inhibitor screening tool was recently shown for pure copper substrates exposed to small organic molecules [79].

2.3.2. QUANTIFYING INHIBITOR PERFORMANCE

THE electrochemical information obtained from the techniques PDP, EIS and LPR can be used to compare the performance of inhibitors. However, it is not possible to directly compare the electrochemical information obtained from different measurement techniques. To enable a more direct comparison between techniques, the results can be converted into relative protection values by comparing the results obtained from the inhibited solutions to the uninhibited ones.

The most widely used metric for comparing the inhibitor performance in the literature is the *inhibition efficiency (IE)*. The inhibition efficiencies are calculated from polarisation resistances R_p obtained from LPR or EIS, the cases when the inhibitor value is higher than blank:

$$\eta = \frac{{R_p}^{inh} - {R_p}^{blank}}{{R_p}^{inh}} = (1 - \frac{{R_p}^{blank}}{{R_p}^{inh}}) \times 100\% \tag{2.3}$$

and corrosion current densities j_{corr} obtained from PDP, the cases when the inhibitor value is lower than blank:

$$\eta = \frac{{j_{corr}}^{blank} - {j_{corr}}^{inh}}{{j_{corr}}^{blank}} = (1 - \frac{{j_{corr}}^{inh}}{{j_{corr}}^{blank}}) \times 100\% \tag{2.4}$$

where superscripts inh and blank stand for inhibited and uninhibited samples, respectively.

Inhibition efficiency is used widely because it is an easy to understand comparison tool. For inhibition, it has values between 0 (no protection at all) to 100% (complete prevention of corrosion). Negative values indicate acceleration of corrosion compared to the uninhibited case. It is also favoured as under simplifying assumptions it can directly be correlated to the surface coverage by the inhibitor molecules. However, this ease of use obscures the fact that as a mathematical function this mapping introduces a mathematical bias and as a result is highly non-linear. Due to its form, $(1 - \frac{a}{b})$, inhibition efficiency introduces an arbitrary 1 next to the relative values $(\frac{a}{b})$ that is of actual interest. As a result, minor differences in performance are seen as large jumps for the lower efficiencies (<90%), and major differences are hidden from view at higher efficiencies (>90%). This also causes researchers to wrongly conclude that good-performing inhibitors would also have lower standard deviations, since even major variations in electrochemical values are suppressed at the higher end of the inhibition efficiency metric. Therefore, it is not an optimal metric to compare the protection performance of strong inhibitors.

An alternative metric, *inhibition power (IP)*, has recently been proposed to address the limitations of inhibition efficiency [130]. It is the ratio of inhibited and uninhibited inhibition information presented in a logarithmic fashion. For polarisation resistance R_p it is defined as:

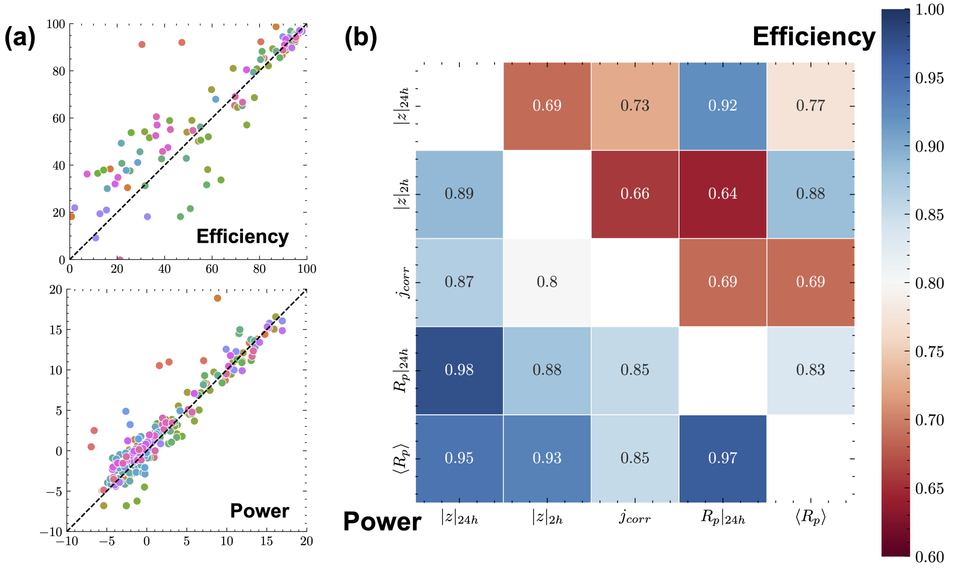


Figure 2.2: The correlation between different electrochemical measurement techniques: EIS performed at 2nd and 24th hour $|z|_{2h}$ and $|z|_{24h}$, potentiodynamic polarisation performed at 24th hour j_{corr}, LPR performed at 24th hour $R_p|_{24h}$ and the time-weighted average of LPR measurements $\langle R_p \rangle$. (a) Example correlation between $|z|_{24h}$ and $\langle R_p \rangle$, values from electrochemical measurements converted into top: inhibition efficiency, bottom: inhibition power. Each dot represents an individual measurement, categorised in colours with respect to their inhibitor species. (b) Pearson correlation coefficients between different electrochemical measurements, converted in top-right triangle: inhibition efficiency, bottom-left triangle: inhibition power metrics.

$$P_{inh} = 10\log_{10}\left(\frac{{R_p}^{inh}}{{R_p}^{blank}}\right) \tag{2.5}$$

and for corrosion current densities j_{corr} it is defined as:

$$P_{inh} = 10\log_{10}\left(\frac{{j_{corr}}^{blank}}{{j_{corr}}^{inh}}\right) \tag{2.6}$$

By taking only the ratio of electrochemical values into account, inhibitor power eliminates the influence of bias introduced by the arbitrary $(1-\frac{a}{b})$ form of inhibitor efficiency determination. In this form, an inhibition power increase of 10 from uninhibited condition corresponds to a corrosion resistance increase by 10-fold, while an increase of 20 corresponds to a 100-fold corrosion resistance increase.

2.3.3. Comparison of electrochemical techniques: Inhibition Efficiency vs. Inhibition Power

The comparison of electrochemical results converted into inhibition efficiency and inhibition power metrics is presented in Figure 2.2. Example correlations between

EIS measured at 24th hour with time-weighted LPR average $\langle R_p \rangle$ for individual experimental runs are visible in Figure 2.2 (a). Figure 2.2 (b) quantifies the correlation between different electrochemical measurement techniques in the form of Pearson correlations. P-values (value describing how likely it is that your data would have occurred under the null hypothesis of your statistical test) of the Pearson statistical test correlations were between 10^{-134} and 10^{-51}, much lower than the commonly used criteria 10^{-6}, indicating statistical significance.

The differences between inhibitor efficiency and power correlations are vividly seen in Figure 2.2 (a). For inhibition efficiency, the correlations are weak except for the top right part, the best-performing inhibitors. This might falsely lead to the impression that an increase in inhibitor performance results in a higher correlation between experiments. This impression is misleading, and is an artifact of the mathematical function used for converting raw electrochemical information into inhibition efficiency. When the correlations are visualised in the form of inhibitor power, higher correlations between the good-performing inhibitors are lost. All compounds behave in a similar way and cluster around the perfect correlation diagonal.

The only exceptions to the strong correlation seen for inhibitor power are the compounds that change their corrosion protection behaviour throughout time. Given that $|z|_{24h}$ measures the protective properties at the 24th hour, and $\langle R_p \rangle$ captures additional time-dependent information, this behaviour is completely expected.

Apart from being more consistent, inhibitor power facilitates discerning between better and best inhibitors. As more conceptually argued in previous section 2.3.2, inhibition efficiency metric squeezes the high-performing compounds together. This is clearly visible from the clustering of experiments for the efficiency metric, versus individually identifiable best-performing compounds for the power metric in Figure 2.2 (a).

The clustering seen for the inhibition efficiency metric also creates an issue for training a predictive model. Imbalanced data usually results in models that have poor predictive performance, especially for the minority class [146]. The homogeneous distribution of results is crucial in training an unbiased machine learning model, which is better provided with the inhibition power metric.

Figure 2.2 (b) presents the correlations between the electrochemical measurement techniques more quantitatively in the form of Pearson correlations. The top-right triangle shows the correlations between different electrochemical measurement technique results converted into inhibition efficiency, and the bottom-left triangle shows the same results converted into inhibition power.

Pearson's bivariate sample correlations quantify linear correlations between two sets of data with the following formula:

$$r_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.7}$$

where n is the total number of experiments (in this work ~300), i the index representing different experiments, x_i, y_i individual sample points from two different electrochemical measurement methods, $\bar{x}$, $\bar{y}$ sample means obtained from the different electrochemical methods. The correlation coefficient $r_{x,y}$ can take values from -1 to 1. 1 indicates a perfect linear relationship between x and y, where all data points lie on a line where x increases as y increases, and vice versa for -1. A value of 0 indicates that there is no linear relationship between the two variables. For time-invariant electrochemical behaviour, a correlation coefficient of 1 is expected between the different electrochemical measurement results [147].

A quick comparison of the inhibition efficiency and power correlation triangles shows that the correlations between measurements are consistently lower for the inhibition efficiency metric. For the inhibition efficiency, the correlations between different techniques are all below 0.9, with the exception of LPR and EIS measurements performed at the 24th hour. For the inhibition power, LPR and EIS measurements carried out at the 24th hour and time-weighted LPR average $\langle R_p \rangle$ show very high correlations. EIS performed at 2nd hour show the lowest correlations with the rest of the measurements. Trustworthy EIS measurements require the electrochemical system to be linear, causal, and time-invariant within the time-frame of the measurement [79, 112]. However for dynamic systems similar to the ones shown on Figure 2.1 (c), time-invariance would not be often observed at measurements done at 2nd hour, which would explain the low correlations. The highest correlation of EIS performed at 2nd hour was observed with a time-weighted parameter, $\langle R_p \rangle$. This again emphasises the time-variable inhibitor behaviour. For inhibition power, higher correlation was observed between $\langle R_p \rangle$ and EIS performed at 24th hour, compared to $\langle R_p \rangle$ and EIS performed at 2nd hour indicates that measurements at 24th hour were more representative of the time-dependent corrosion inhibition behaviour. Surprisingly for inhibition efficiency, the opposite is the case. This might be due to the volatility inherent to the inhibition efficiency transformation.

PDP measurements show lower correlations with the rest. This is most likely due to altered electrochemical behaviour caused by the high overpotentials (±250 mV) necessary for the PDP experiments. Due to the high overpotentials encountered during the potentiodynamic scans, the physicochemical properties of the surface are modified, potentially leading to an altered substrate surface chemistry [79, 148]. Another reason could be the increased user input during the Tafel slope analysis required for corrosion current density calculations, which is much higher than required for EIS or LPR. Specifically for the case of AA2024-T3, the use of Tafel approach is not straightforward. On one side, the cathodic behaviour is significantly influenced by oxygen diffusion limitations. On the other, the anodic processes are not solely governed by charge transfer, but rather occur at localised regions such as intermetallic particles and grain boundaries. Therefore, the conventional Tafel approach cannot be employed since it is applicable only under activation-controlled processes. Tafel analysis in such conditions is a very simplified approach, and can lead to deviations.

For the reasons presented above, we argue that inhibition power is a more 'efficient' way in discerning between better and best inhibitors, and a better approach to training an unbiased predictive machine learning model. Therefore, it is used to compare and rank the inhibitor performance in the next section.

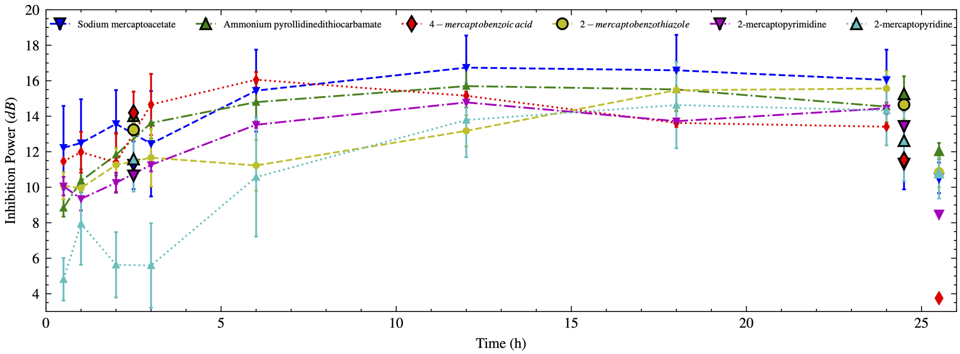


Figure 2.3: The representation of PDP, EIS, and LPR measurements converted into inhibition power for best performing among the tested inhibitors. Small line-scatter plots represent LPR, larger plots with black edges represent EIS at 2nd and 24th hours, and the final larger scatter plots represent the PDP measurement results converted into inhibition power. The solubility of inhibitors denoted with italics were less than 1 mM.

2.3.4. RANKING OF INHIBITORS

FIGURE 2.3 demonstrates the electrochemical measurement results converted into inhibition power for the best-performing inhibitors. LPR measurements are shown with line-scatter, EIS with black framed scatter, and PDP with the final individual scatter plots. The width of the EIS and PDP symbols were chosen so that it would convey the time it takes to perform the measurements.

The presented inhibitors show stable behaviour after 6 hours, with the exception of 2-mercaptobenzothiazole which develops inhibition until after 18 hour, and of 4-mercaptobenzoic acid which seems to show a minor decrease in inhibition performance with time. LPR and EIS results correlate strongly with each other (except for 2-mercaptopyridine EIS around 2 hours), as expected from the analysis in the previous section 2.3.3. PDP results demonstrated a lower inhibition power for all cases. This systematic difference was attributed to the destructive nature of PDP measurements.

Although a qualitative analysis is possible through such plots, the quantitative ranking of a high number of inhibitors is not feasible through such visualisations. To this end, the time-weighted LPR average $\langle R_p \rangle$ is advantageous as it captures the complete time-dependent behaviour in a single number. Additionally, it shows high correlation with other electrochemical techniques as seen from Figure 2.2 (b).

Figure 2.4 presents the ranking of inhibitor candidates in the form of a box-plot, created from the time-weighted LPR average $\langle R_p \rangle$ values converted into inhibition power through Equation 2.5. Inhibitor candidates are ranked with respect to their mean inhibition power values, and their medians are represented by horizontal bars. The box part shows the main portion of the data, the interquartile range. The edges of the box show the 25th and 75th percentile. Whiskers show the minimum and maximum measurement results.

The importance of heteroatom presence, aromatic ring and π-bond containing molecular structures on inhibitor performance have been consistently mentioned in the literature [47, 109, 149, 150]. It has been argued that the availability of non-bonded

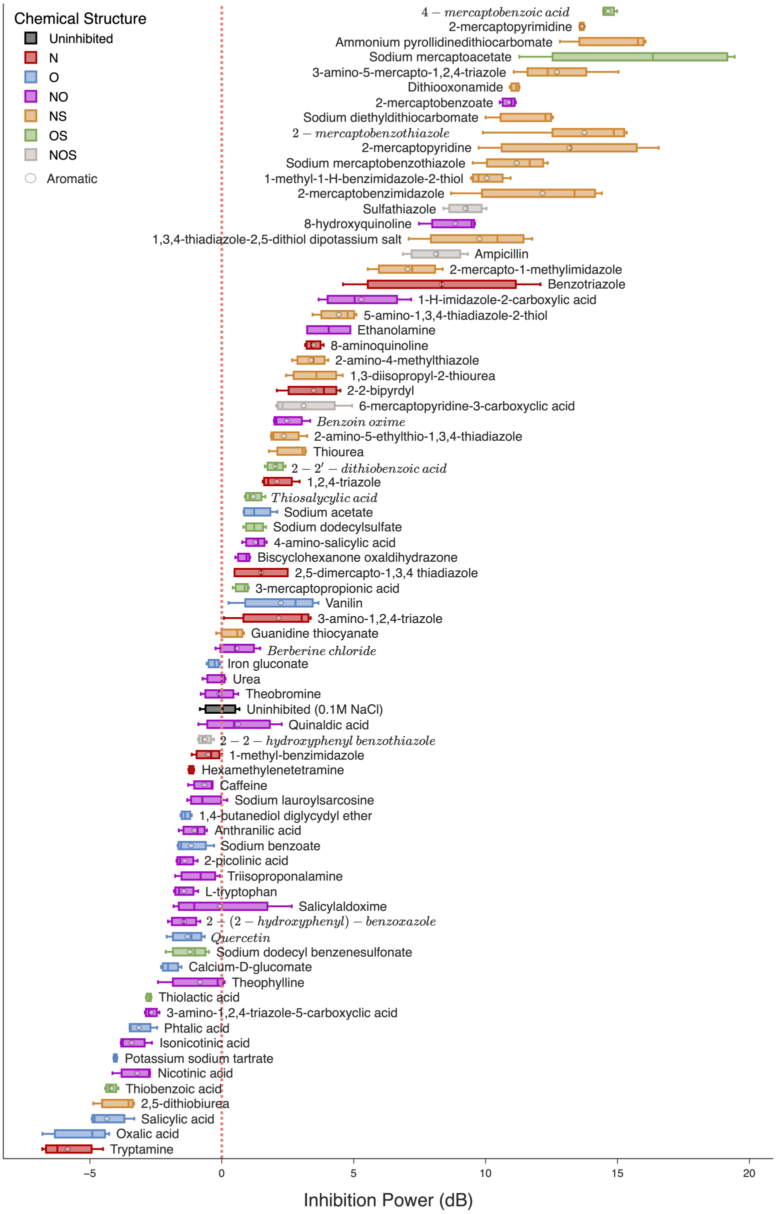


Figure 2.4: The inhibitor candidate ranking visualised as boxplots. Colours indicate the nitrogen (N), oxygen (O), sulphur (S) heteroatom content and presence/absence of aromatic ring structures. The solubility of molecules denoted with italics were less than 1 mM.

lone pair electrons of heteroatoms and π-electrons of double/triple bonds facilitate electron transfer from the inhibitor to the d-orbitals of the metal, acting as adsorption centers during metal-inhibitor interactions. To identify such trends in this experimental data set, the inhibitors have been categorised according to their molecular structures: the presence of N,O,S heteroatoms and their aromatic vs. aliphatic bond structures.

Almost half of the inhibitor candidates behaved as corrosion accelerators. This was in contrast to the findings of previous studies [107, 136], which was the basis of the inhibitor selection procedure of our paper. Non-adjusted pH could be one reason for this behaviour. 80% of sole O heteroatom containing compounds behaved as accelerators, with the exception of sodium acetate and vanillin. On the other hand, N and S heteroatom containing organic molecules performed consistently well. They had the highest inhibition power values with none of them performing as corrosion accelerators. Compounds that contained N, S and O together had in-between inhibition properties. This leads us to suggest that N, S heteroatoms grant inhibitive properties to the organic molecule, whereas O could potentially hinder inhibition. Specifically for AA2024-T3 corrosion inhibition, it was observed that functional groups with N and S heteroatoms form coordination complexes with Cu containing intermetallics, reducing the corrosion rate [151–153]. The heteroatom trend is generally in line with the previously suggested heteroatom electronegativity - inhibition effect, where heteroatoms provided inhibition with inverse order of their calculated individual electronegativity: $P > S > N > O$ [149]. It is proposed that lesser electronegativity results in increased charge transfer and provide inhibition. However the real situation in a small organic molecule is much more complex as the electronegativities of the heteroatoms will change depending on the molecular structure.

The discussion above addresses only one of the important molecular descriptors. Trends are not clear for the rest. The comparison of aromatic and aliphatic behaviour show no significant difference. This is most likely because the tested aliphatic molecules already contain excess π-bonds in their linear chain. The behaviour of N and O, and S and O containing molecules are most complex. The molecules are spread throughout the inhibitor/accelerator spectrum, seemingly without an underlying order and act as best and worst performing compounds, as seen in the behaviour of 4-mercaptobenzoic acid and thiobenzoic acid.

2.3.5. Understanding and predicting inhibition: experimental input features for the machine learning model

It is clear that any predictive corrosion inhibition model requires a more comprehensive description of the system than the presence or absence of heteroatoms. Compared to analysing individual properties like the presence of certain heteroatoms, π-bonds and functional moieties, quantitative structure–property relationship (QSPR) models have potential in exploring more complex physical phenomena [16, 34, 138, 154–161]. QSPR inhibition models relate predictor variables, which can be physicochemical properties and/or theoretical molecular descriptors of inhibitor compounds, to the experimentally measured inhibition performance. Quantified physicochemical properties or descriptors (obtained through theoretical calculations and molecular modelling techniques such as density functional theory and molecular dynamics) expressed in a mathematical relationship, a quantitative structure-property relationship, can be established to predict the performance of untested organic molecules.

The inclusion of experimental physicochemical descriptors is the next logical step to supplement the input feature pool and to concomitantly improve the robustness of the predicted values as well as the generalizability of QSPR models for small organic corrosion inhibitors . Some important physical and chemical experimental input features that are capable of increasing prediction quality are presented below.

Molecular Weight: Molecular weight - inhibitor power relationship can be found in the Supplementary Figure 2.2. It seems that most organic molecules cluster in the range of 100 to 200 g mol^{-1}, and after around 250 g mol^{-1} there seems to be a decrease in the inhibitor performance. This is most likely due to steric hindrance effects, where an increase in the size of the molecule would hamper the adsorption reactions with the substrate [162]. Based on this observation we suggest that as a rule of thumb, small organic molecules with molecular weights lower than 250 g mol^{-1} can hold more promise to be inhibitor candidates. This would limit the chemical space to be explored and facilitate the efficiency of novel inhibitor discovery.

Inhibitor Concentration: The influence of concentration is certainly important for inhibition behaviour. An exploratory comparison of 6 molecules at 0.1 and 1 mM concentrations show that with increasing concentration, inhibitor systems become more protective and accelerator systems become more corrosive (provided in Supplementary Figure 2.12). Typically as concentration increases, a corresponding increase in inhibition is observed until a critical concentration is reached. After this critical concentration the inhibition either reaches a plateau, or in certain cases it starts to decline [163]. It was previously argued that the decline in inhibition was related to the formation of oligomers: either the molecule concentration higher than the critical value causes adsorbed inhibitor molecules to desorb due to interaction with free molecules present in the solution, forming oligomers, or oligomers that form in the solution beforehand reduce the concentration of inhibitor available for adsorption [78]. Any analysis of an inhibition system has to be aware of such a behaviour when comparing inhibition performance at different conditions.

Electrochemical Potentials: Electrochemical information obtained from the experiments can serve as target parameters to be predicted (such as previously calculated inhibition power), and also can be utilised as descriptors. This can augment the molec-

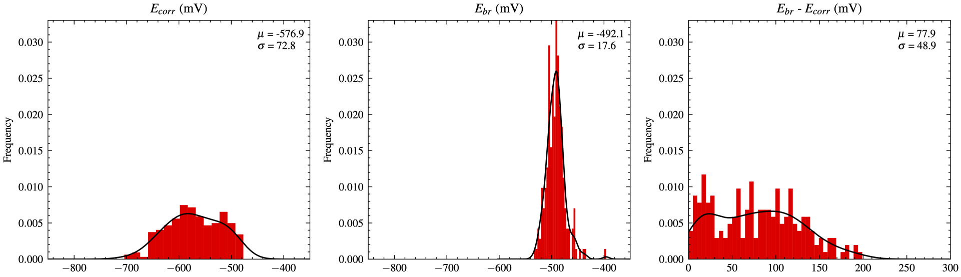


Figure 2.5: The distribution of corrosion potentials E_{corr}, pitting breakdown potentials E_{br} where a sudden jump in anodic current is observed, and the differences between the two. Histograms are shown as red bars, kernel density estimates of the probability functions are shown as black curves. μ and σ represent the mean and standard deviation, respectively.

ular descriptors of the model by adding mechanistic insights related to the electrolyte-electrode system, which were otherwise lacking from the statistical nature of machine learning models.

The dominant degradation mechanism of AA2024-T3 is pitting corrosion [42, 55]. Furthermore, the alloy is used in combination with composite structures in modern aeroplanes which triggers galvanic corrosion. For this reason, parameters that represent pitting and galvanic corrosion hold promise as either target parameters to be predicted, or as additional descriptors that provide mechanistic information to the models.

Figure 2.5 presents the distribution of corrosion potentials E_{corr}, pitting breakdown potentials E_{br} where an instantenous large increase in anodic current is observed [164], and the differences between the two. It is seen that inhibitors can modify E_{corr} significantly, as seen from the 200 mV range and high standard deviation of 72.8 mV. On the other hand, E_{br} values change negligibly, with a standard deviation of 17.6 mV, leading us to believe that this is an intrinsic property of the substrate. This is in line with previous dealloying studies, where an alloy dependent intrinsic critical potential was observed for activating porosity formation in an otherwise passive surface [165]. E_{br} acts as the threshold potential for preferential dealloying of the active phases, which in the case of AA2024-T3 is the potential for initiating stable pits resulting from active S (Al_2CuMg) and ϑ (Al_2Cu) phase intermetallics [39]. The difference between potentials E_{br} - E_{corr} describes the overpotentials required to reach this threshold, which was shown to be highly influenced by the introduction of inhibitors.

The influence of difference in potentials E_{br} - E_{corr} is denoted here as passive range, and plotted with respect to inhibitor power in Figure 2.6 to see whether there is a correlation between the two parameters. Different chemical groups are denoted with different colours. No significant correlation was observed between the passive range and inhibition performance. It was observed that apart from NS aliphatic and OS aliphatic/aromatic compounds, a weak negative correlation between the two parameters was observed. However, this behaviour was not statistically significant because of the high spread observed for the experiments. In any case, the seemingly unsystematic behaviour with low correlation highlights the need for further study. As the key parameter for localised electrochemical activity, passive range holds promise either as a target to be pre-

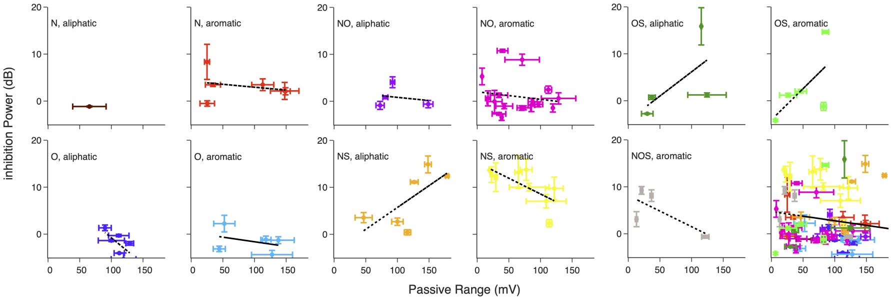


Figure 2.6: The correlation between inhibitor power and the passive range ($E_{corr} - E_{br}$).

dicted on its own, or as a descriptor to be used in combination with the molecular descriptors.

Bulk pH: Apart from 4-mercaptobenzoic acid, sodium diethyldithiocarbamate, and 1,3,4-thiadiazole-2,5-dithiol-dipotassium salt, the pH did not change in the presence of compounds with good inhibition performance (IP>10, IE>90%) and had neutral pH values around 6. On the other hand, IP was lower in the presence of compounds which caused the initial pH of the electrolyte to be out of the 4.5 to 8.5 Al stability window [166]. The clustering of lower pH values at the lower inhibition power segment suggests that this results in active corrosion becoming the dominant degradation mechanism.

It was seen that there is no correlation of inhibition power with either the average or the difference in pH (Supplementary Figures 2.10-2.11). It must be noticed that what is measured as bulk electrolyte pH and what the actual pH observed on the substrate surface can be very different, and bulk electrolyte measurements do not fully reflect local behaviour such as concentration gradients at the electrolyte-substrate interface and throughout the diffusion layer [167].

The lack of correlation does not mean that bulk pH information is useless as a machine learning model feature. It is very relevant for explaining the outlier behaviour, as the pH difference caused by the inhibitor molecule is not captured directly with computational descriptors. The addition of bulk pH as a feature can capture such pH-based behaviour, and can be used as a forensic analysis tool to explain outliers of the model.

2.3.6. EXPLORING EXPERIMENTAL DESCRIPTORS FOR MACHINE LEARNING

EXPERIMENTALLY measured pH shows the power of descriptors obtained from experiments. To produce a short list of compounds with possibly useful properties for further experimental testing.The selection of relevant input features is a crucial step in the development of QSPR models as features with low or no relevance to the target property will degrade the model. The recursive feature elimination (RFE) was carried out for the four distinct groups of input features: structural features only, structural features combined with DFT, structural features combined with average pH, and structural features combined with DFT and average pH. Feature elimination was performed for both IE and

Table 2.2: Results of one specific train test split.

Target	# Features	RMSE structural	RMSE structural + DFT	RMSE structural + pH	RMSE structural + DFT + pH	R2 structural	R2 structural + DFT	R2 structural + pH	R2 structural + DFT + pH
IE	10	0.24	0.23	0.18	0.18	0.17	0.19	0.49	0.51
	5	0.22	0.24	0.2	0.19	0.27	0.13	0.42	0.43
IP	10	0.19	0.19	0.18	0.18	0.25	0.31	0.32	0.38
	5	0.2	0.15	0.16	0.15	0.19	0.55	0.49	0.54

Table 2.3: Results of 6-fold cross validation.

Target	# Features	RMSE structural	RMSE structural + DFT	RMSE structural + pH	RMSE structural + DFT + pH	R2 structural	R2 structural + DFT	R2 structural + pH	R2 structural + DFT + pH
IE	10	0.22 (± 0.03)	0.19 (± 0.02)	0.14 (± 0.02)	0.14 (± 0.02)	-0.47 (± 0.33)	-0.12 (± 0.18)	0.3 (± 0.21)	0.35 (± 0.21)
	5	0.21 (± 0.03)	0.23 (± 0.02)	0.14 (± 0.02)	0.14 (± 0.02)	-0.46 (± 0.38)	-1.07 (± 0.67)	0.27 (± 0.22)	0.35 (± 0.22)
IP	10	0.2 (± 0.04)	0.19 (± 0.04)	0.2 (± 0.04)	0.18 (± 0.04)	0.3 (± 0.14)	0.35 (± 0.14)	0.31 (± 0.13)	0.41 (± 0.13)
	5	0.21 (± 0.04)	0.2 (± 0.03)	0.2 (± 0.03)	0.18 (± 0.03)	0.28 (± 0.13)	0.33 (± 0.14)	0.35 (± 0.11)	0.41 (± 0.11)

IP targets. The whole feature selection process was repeated 100 times with different random seeds and the *n*-tuples that were selected in the majority of the runs can be found in the Supplementary Tables 2.5-2.8.To use the same technique for the QSPR step that was employed for sparse feature selection, random forest (RF) models have been trained using the experimental database. By algorithmically eliminating the weakest features, it allows automatic feature selection without user bias or intervention [138]. Moreover, RF is an ensemble model that builds multiple decision trees and combines their predictions. Naturally, this ensemble approach helps reduce the risk of overfitting, which can be crucial when dealing with small datasets. Another advantage is robustness against outliers: outliers can have a significant impact on smaller datasets, whose influence again can be mitigated by aggregating predictions from multiple trees.

RF regression models predicting the quantitative inhibition performance values were trained to create an active material discovery loop to explore the vast chemical space for promising compounds in an efficient manner. Out of the 78 organic molecules that were tested, only 59 were fully dissolved in solutions. These molecules corresponded to a target concentration of 1 mM, and were used to train the ML models. As the input to these models molecular descriptors (MDs) based on the structure of the molecules, descriptors calculated by DFT as well as selected experimental parameters have been used. The accuracy and robustness of the trained models is assessed using a cross-validation (CV) approach.

In aqueous solutions, aluminium alloys have a protective passive (hydr)oxide layer preventing it from corrosion at a pH range roughly between 4 and 10 [166]. In this pH range, scratches or mechanical damage to the passive layer are quickly repaired but if the pH drops below or rises above the stable range, aluminium starts to corrode actively. As the oxide layer is no longer stable at such conditions, this influences the inhibitor-substrate interaction. As a result, the pH makes for an effective feature in a ML model because aluminum is typically more likely to corrode at very high or very low pH levels. The pH is selected by the RFE routine every time it is part of the set of input features. This demonstrates emphatically that the pH appears to be a key feature in the prediction of inhibition performance of organic molecules.

In addition to pH, several properties derived from DFT calculations are also selected by the RFE as soon as they are included to the set of input features. The DFT parameter

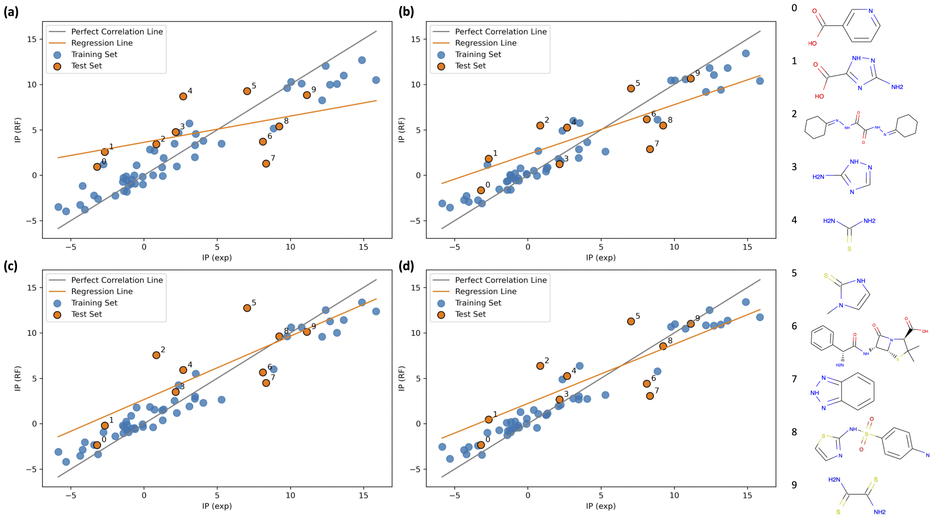


Figure 2.7: Prediction results for random forest models with 5 input features that uses the IP as target. Feature pool: (a) only structural features, (b) structural features and DFT parameters, (c) structural features and pH, (d) structural features, pH and DFT parameters.

that was selected most frequently, was the highest occupied molecular orbital (HOMO). Additionally, the lowest unoccupied molecular orbital (LUMO) and dipole were selected in at least half of the cases, with n = 10 for the RFE step. This contrasts with recent works, which have concluded that the correlation between DFT properties and the corrosion inhibiting effect of small molecules seems absent [130, 168]. However, neither of these works mixed the DFT features with molecular descriptors that encode the molecular structure. It is noteworthy that the correlation between the HOMO energy levels and IE/IP is essentially zero in this work as well, corroborating these prior works.

When examining the results for one specific train test split in Table 2.2, it is evident that at least for most of the cases where the DFT parameters and/or pH value are added to the set of input features, the R^2 increases and the RMSE decreases. This indicates that including these parameters enhances the prediction and increases the reliability and robustness of the models. The only case where this does not hold is the IE model with five input features combining structural features and DFT. A closer examination reveals that for IE ten input features allow for more accurate predictions than five, whereas for IP the reverse is true. Lowest RMSE and highest R^2 was achieved for the model that uses combined descriptors and IP as the target. In Figure 2.7 the measured IP is plotted against the IP predicted by the RF.

In order to perform CV, the dataset was divided into six folds and thus six RF models were trained. The average R^2 and RMSE and the corresponding standard deviation of these models are shown in Table 2.3. The evaluation of the models using different classes of input features indicate that adding DFT parameters and/or the pH value increases the prediction accuracy in most of the cases according to the determined mean values for R^2 and RMSE. The models with the lowest RMSE and highest R^2 include pH and DFT parameters as input in addition to the structural features, further supporting our claim that molecular descriptors derived from atomistic simulations can be helpful to generate QSPR models that predict the corrosion inhibition responses of small organic molecules to lightweight engineering metals such as aluminium and magnesium alloys. Unlike the specific train test split case, lowest RMSE was obtained for IE as target, and the highest R^2 was achieved for IP as the target. Unfortunately high variation among different folds makes it difficult to state with certainty whether IP or IE perform better targets for such models. However, the comparably low R^2 and RMSE values for all considered models and the high standard deviations of these metrics indicate that more training data is required to achieve better generalisation. Furthermore, they are highly sensitive to outliers in the blind test set.

2.4. Conclusion

In summary, we employed various standard electrochemical techniques at different intervals to investigate the electrochemical behaviour of around 80 small organic molecules. Our aim was to capture the most comprehensive electrochemical picture of AA2024-T3 immersed in inhibitor containing electrolytes. The performance of inhibitor candidates was quantified through statistical analysis of their electrochemical response. This highlighted the need for complementary information from different techniques to have a mechanistic understanding of an inhibition system. For initial inhibitor screening purposes, time-weighted LPR measurements showed very high correlations with other techniques and are a good substitute for representing the protective behaviour of the inhibitor. Time-dependent measurements showed that for the majority of organic molecules electrochemical measurements performed in less than 6 hours varied in time and were unstable. To understand the true inhibitive properties of inhibitor candidates, electrochemical studies should analyse the inhibition performance at least after 6 hours for more reliable results. Statistical analysis shows that inhibition efficiency is not an ’efficient’ way to distinguish between good inhibitors. Inhibition power is a more suitable metric for discerning between “better” and “best” inhibitors. Inhibition power eliminates clustering of data observed in higher efficiency range (>90%), which is an important condition for training an unbiased machine learning model. The need for more complicated predictive models with advanced descriptors was clear by categorising molecules based on heteroatom content and the presence of aromatic moieties. Compounds that contain both N and S heteroatoms performed consistently well, however the performance of compounds with other chemical structures was spread over a large range. Electrochemical information coming from corrosion potential and passive range bears no linear correlation to inhibition power and could be either a predictive descriptor in combination with other features for predicting corrosion resistance, or can be an important prediction target as it is a key parameter for localised corrosion. The machine learning model augmented with mechanistic information is key in exploring the complexity of corrosion phenomena, which was highlighted by the predictive power of pH. No linear relationship between bulk pH and inhibitor performance was observed, however information gained from pH assisted in describing the system better by including information about the environment not necessarily found in computational descriptors, which increased the prediction rate and assisted in outlier analysis of the random forest models.

At this stage rather than designing a final prediction system, we have explored the use of machine learning models to create an active learning loop for more efficient experimental discovery. The obtained experimental parameters can be employed directly as target parameter for training of a machine learning model that is predictive of the performance of untested compounds to create a shortlist of promising candidates. Moreover, the experimental investigation yielded additional input features like pH that can be combined with molecular descriptors derived from the molecular structure and atomistic simulations. These input features exhibit great potential to develop augmented quantitative structure-activity relationships as they allow the direct inclusion of information about the underlying mechanisms in training of the models. The results of this study are expected to support the development of i) faster inhibitor screening techniques which can capture the same high resolution electrochemical information on a shorter timescale, ii) more complex models that can leverage the link between the physicochemical nature of the inhibitor and its protective performance.

2.5. SUPPLEMENTARY INFORMATION

2.5.1. EXPERIMENTAL DETAILS

ALLOY COMPOSITION

Table 2.4: Chemical composition (wt.%) of AA2024-T3.

Element	Al	Si	Fe	Cu	Mn	Mg	Cr	Zn	Ti	V	Zr	Other
wt.%	93.1	0.08	0.19	4.6	0.56	1.3	0.01	0.11	0.02	0.01	0.00	0.05

GRAPHICAL SUMMARY OF EXPERIMENTS

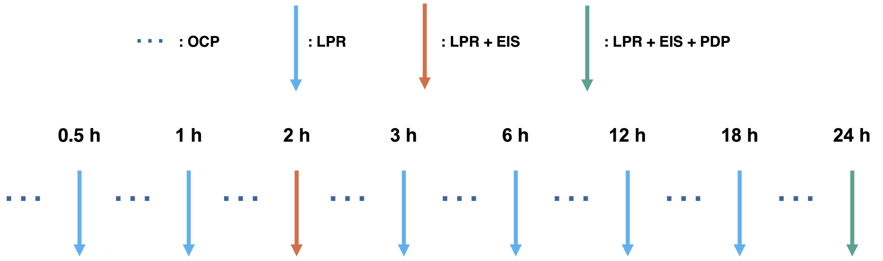


Figure 2.8: A visual summary of the electrochemical experiments. Acronyms correspond to: OCP - open circuit potential, LPR - linear polarisation resistance, EIS - electrochemical impedance spectroscopy, PDP - potentiodynamic polarisation. The provided times indicate when the respective measurements have been performed during the duration of the experiment.

2.5.2. EXPERIMENT FEATURE CORRELATIONS

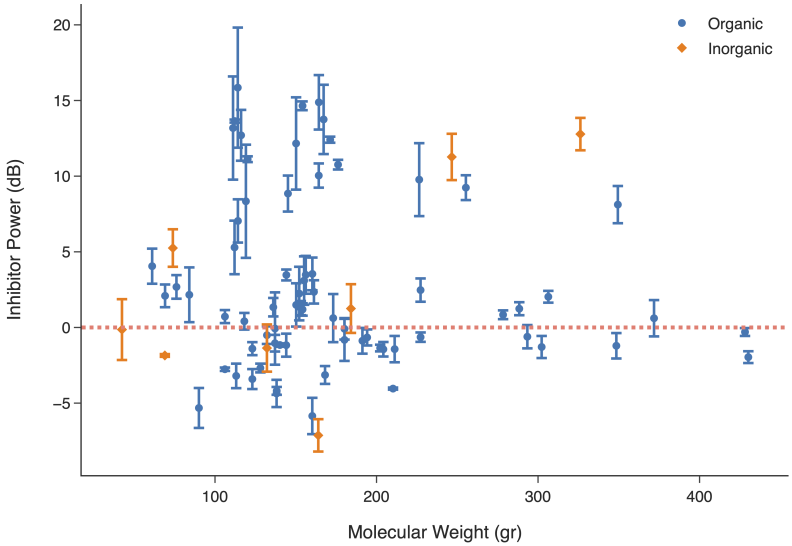


Figure 2.9: The correlation between inhibitor power and molecular weight of the inhibitor candidates.

*Inorganic compounds correspond to chemicals not explicitly mentioned in this paper. They were not the focus of the study as they cannot be used to train the statistical model in the same way as organic molecules, due to their different protection mechanisms among other reasons. They consist of Ce, Li, phosphate and sulphate salts.

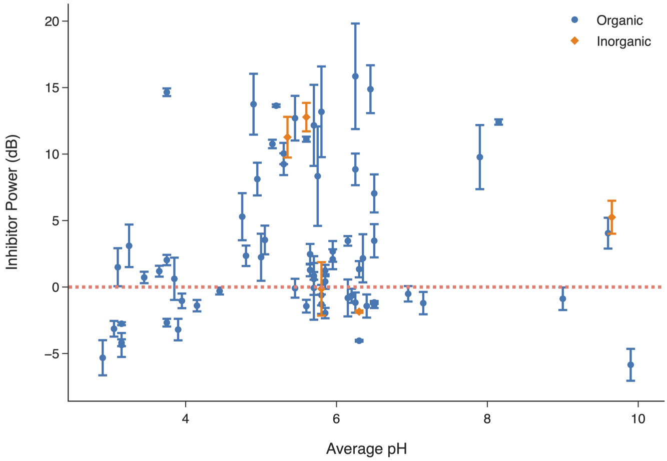


Figure 2.10: The correlation between inhibitor power and average pH of the electrolytes before and after the experiments.

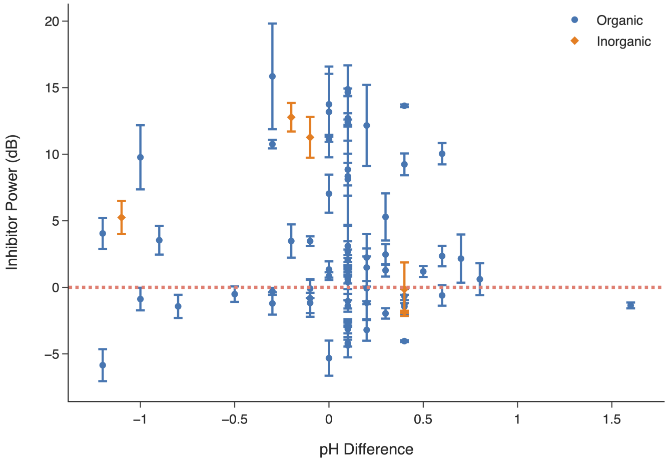


Figure 2.11: The correlation between inhibitor power and pH difference of the electrolytes before and after the experiments.

2.5.3. Electrochemical performance distributions

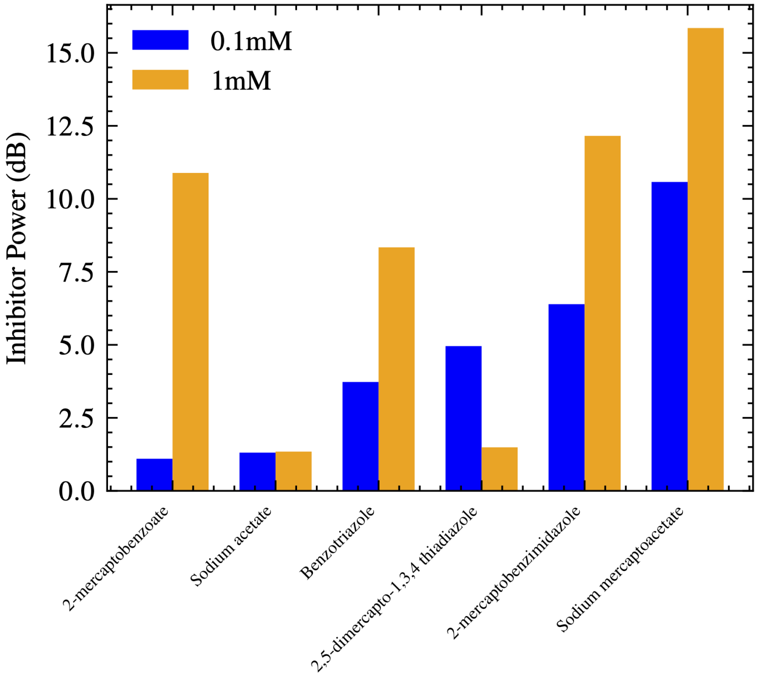


Figure 2.12: Comparison of electrochemical response of small molecules in 0.1 and 1 mM concentrations.

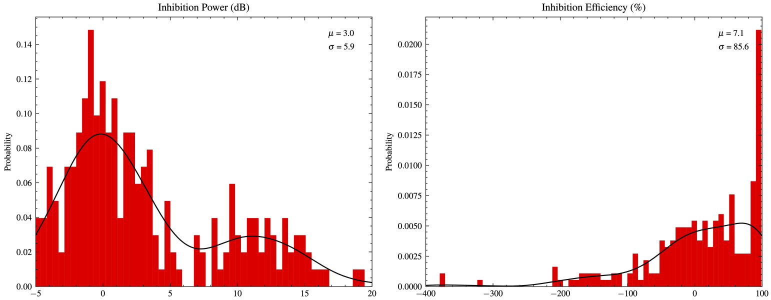


Figure 2.13: Comparison of the distribution of electrochemical data of $\langle R_p \rangle$ converted into inhibition power and efficiency. Red: histograms with bin size 60, black kernel density estimates of the distribution.

2.5.4. RECURSIVE FEATURE SELECTION

INHIBITION EFFICIENCY

Table 2.5: Selected features for IE (single train test split).

Type	Num of Features	Features
stuctural	5	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, Ipc, VSA_EState2
stuctural	10	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, BCUT2D_MRLOW, Ipc, PEOE_VSA14, SMR_VSA10, TPSA, VSA_EState2, VSA_EState4
stuctural + DFT	5	MaxEStateIndex, BCUT2D_MWHI, VSA_EState2, HOMO(eV), LUMO(eV)
stuctural + DFT	10	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, BCUT2D_MRLOW, Ipc, TPSA, VSA_EState2, VSA_EState4, HOMO(eV), LUMO(eV)
stuctural + pH	5	BCUT2D_MWHI, SMR_VSA10, TPSA, MolLogP, pH avg
stuctural + pH	10	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, BCUT2D_MWLOW, SMR_VSA10, TPSA, EState_VSA2, VSA_EState4, MolLogP, pH avg
stuctural + DFT + pH	5	SMR_VSA10, TPSA, MolLogP, HOMO(eV), pH avg
stuctural + DFT + pH	10	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, SMR_VSA10, TPSA, EState_VSA2, VSA_EState4, MolLogP, HOMO(eV), pH avg

Table 2.6: Selected features for IE (cross validation).

Type	Num of Features	Features
stuctural	5	MaxEStateIndex, MinAbsEStateIndex, Ipc, EState_VSA5, VSA_EState2
stuctural	10	MaxEStateIndex, MinAbsEStateIndex, FpDensityMorgan1, Ipc, PEOE_VSA14, SMR_VSA10, EState_VSA5, VSA_EState2, VSA_EState3, VSA_EState4
stuctural + DFT	5	MaxEStateIndex, BCUT2D_MWHI, SMR_VSA10, VSA_EState2, HOMO(eV)
stuctural + DFT	10	MaxEStateIndex, MolWt, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_MRLOW, SMR_VSA10, VSA_EState2, HOMO(eV), LUMO(eV), dipole(debye)
stuctural + pH	5	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, SMR_VSA10, pH avg
stuctural + pH	10	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, BCUT2D_MWLOW, SMR_VSA10, VSA_EState2, VSA_EState3, VSA_EState4, MolLogP, pH avg
stuctural + DFT + pH	5	MaxEStateIndex, MinAbsEStateIndex, SMR_VSA10, HOMO(eV), pH avg
stuctural + DFT + pH	10	MaxEStateIndex, MinAbsEStateIndex, BCUT2D_MWHI, BCUT2D_MWLOW, SMR_VSA10, VSA_EState2, MolLogP, HOMO(eV), LUMO(eV), pH avg

Inhibition Power

Table 2.7: Selected features for IP (single train test split).

Type	Num of Features	Features
stuctural	5	MaxEStateIndex, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_MRLOW, SMR_VSA10
stuctural	10	MaxEStateIndex, MinEStateIndex, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_MRLOW, SMR_VSA10, SlogP_VSA4, TPSA, VSA_EState2, MolLogP
stuctural + DFT	5	MaxEStateIndex, BCUT2D_MWHI, SMR_VSA10, HOMO(eV), dipole(debye)
stuctural + DFT	10	MaxEStateIndex, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_MWLOW, SMR_VSA10, TPSA, VSA_EState2, MolLogP, HOMO(eV), dipole(debye)
stuctural + pH	5	MaxEStateIndex, BCUT2D_MWHI, SMR_VSA10, VSA_EState2, pH avg
stuctural + pH	10	MaxEStateIndex, FpDensityMorgan1, BCUT2D_MWHI, Ipc, SMR_VSA10, SlogP_VSA4, TPSA, VSA_EState2, MolLogP, pH avg
stuctural + DFT + pH	5	MaxEStateIndex, BCUT2D_MWHI, SMR_VSA10, HOMO(eV), pH avg
stuctural + DFT + pH	10	MaxEStateIndex, FpDensityMorgan1, BCUT2D_MWHI, SMR_VSA10, SlogP_VSA4, TPSA, VSA_EState2, HOMO(eV), dipole(debye), pH avg

Table 2.8: Selected features for IP (cross validation).

Type	Num of Features	Features
stuctural	5	MaxEStateIndex, MolWt, BCUT2D_MWHI, SMR_VSA10, VSA_EState2
stuctural	10	MaxEStateIndex, MinEStateIndex, MolWt, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_MRLOW, BalabanJ, SMR_VSA10, VSA_EState2, VSA_EState3
stuctural + DFT	5	MaxEStateIndex, BCUT2D_MWHI, SMR_VSA10, VSA_EState2, HOMO(eV)
stuctural + DFT	10	MaxEStateIndex, MolWt, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_MRLOW, SMR_VSA10, VSA_EState2, HOMO(eV), LUMO(eV), dipole(debye)
stuctural + pH	5	MaxEStateIndex, BCUT2D_MWHI, SMR_VSA10, VSA_EState2, pH avg
stuctural + pH	10	MaxEStateIndex, MinEStateIndex, MolWt, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_MRLOW, SMR_VSA10, TPSA, VSA_EState2, pH avg
stuctural + DFT + pH	5	MaxEStateIndex, BCUT2D_MWHI, SMR_VSA10, HOMO(eV), pH avg
stuctural + DFT + pH	10	MaxEStateIndex, MolWt, BCUT2D_MWHI, BCUT2D_MRLOW, SMR_VSA10, VSA_EState2, HOMO(eV), LUMO(eV), dipole(debye), pH avg

3

Evaluating Inhibition

For every complex problem there is an answer that is clear, simple and wrong.

H. L. Mencken

The search for non-toxic alternatives to hexavalent chromium based corrosion inhibitors requires a comprehensive understanding of the factors critical to effective corrosion protection. Key considerations include the evolution of corrosion inhibition with inhibitor concentrations and exposure times, the inhibition efficacy in the presence and following absence of inhibitors, and the stability of inhibition upon polarisation. In our electrochemical comparison of promising organic molecules with sodium dichromate, we found that even top-performing candidates can lead to premature conclusions if such critical factors are overlooked. While organic molecules can match the inhibition performance of chromates under specific conditions, this can be misleading when considering concentration, time, and polarisation dependent behaviour. Initial high performance can also be deceptive in dynamic environments, as we observed that the inhibition provided by most organic molecules drastically decreases when the inhibitor is absent in the electrolyte. These observations call for broader comprehensive inhibitor robustness studies that take into account factors including time, concentration, stability, and polarisation effects in inhibitor efficacy analysis.

This chapter has been published with some adjustments as *Factors to consider in the quest for organic alternatives to hexavalent chromium based corrosion inhibitors*, Corrosion Science (2025) [169].

3.1. INTRODUCTION

THE use of chromates in corrosion protection for structural materials in aeronautics has been strictly regulated internationally for many years due to health and safety concerns. Despite significant advancements, academia and industry continue to explore various environmentally-friendly and sustainable alternatives to hexavalent chromium (Cr(VI)), which remains the benchmark corrosion inhibitor with a proven track record. According to the EU REACH (Registration, Evaluation, Authorisation, and Restriction of Chemicals) administered by the European Chemicals Agency (ECHA) [8], the use of most chromates was banned in Europe from January 2019 [170], unless an authorisation has been granted for a specific use by a specific chemical supply-chain or downstream user, for a limited time period, and for the specific cases where no suitable alternatives can be implemented. Given the substantial number of authorisation applications for the use of Cr(VI) substances, ECHA is now moving from authorisation (Annex XIV [171]) to restriction (Annex XVII [172]) of the usage of such compounds since authorisation is deemed no longer appropriate to control the risk to human health posed by these substances [173–175].

The development of Cr(VI) alternative corrosion inhibitors is an active area of research, with promising candidates including, but not limited to, lithium [113, 176, 177] and rare-earth [87, 178, 179] based systems. However, a one-to-one replacement of Cr(VI) pigments remains unlikely, as a recent review suggests [62]. This is due to the Cr(VI) pigments' unique ability to provide multi-functional corrosion protection, including passivation, self-healing, and environmental stability, which is challenging to replicate with a single non-toxic compound. Instead, synergistic systems combining multiple compounds are more viable, with each targeting specific aspects of corrosion prevention.

In this context, organic molecules have emerged as potentially suitable candidates due to their diverse structures and properties. Recent research have highlighted the potential of organic molecules as corrosion inhibitors, with significant progress in understanding structure-performance relationships through studies of related compounds [107], high-throughput screening using optical [114, 115, 136], electrochemical [117, 119], and spectroscopic methods [114, 120, 124], and machine-learning models to develop quantitative structure-property relationships [33, 114, 122, 136, 138, 180, 181]. Next to novel data generation, curation of open databases [1, 140, 181] and mining research papers through natural language processing [182] have been utilised to effectively search existing literature for potential Cr(VI) replacements.

Despite these advancements, many studies rely on single metrics such as inhibition efficiency or power captured at one timestep and concentration to evaluate inhibition performance. While convenient, this approach may not capture all the necessary aspects for identifying next-generation materials. The robustness of the inhibition performance under changing environmental conditions is necessary in transforming such molecules into actual problem-solving products, so the understanding of factors resulting in robust corrosion inhibition is most critical.

The influence of pH is historically the most well-studied environmental factor. Its effect is twofold: i) outside the stable pH window where the metal oxide is unstable, corrosion modes might change (e.g. from localised corrosion to a uniform one) and corrosion inhibitors with the right molecular structure and high inhibition efficacy at a mild pH may stop working at a harsh one, ii) depending on the isoelectric point and pKa values of the inhibitor molecule it might be positively/negatively charged and protonated/deprotonated, changing the surface binding mode and hence its inhibition efficacy [77, 119, 183, 184]. Despite its critical role, we've highlighted its importance in our previous work, so we focus on other factors in this one [80].

The influence of inhibitor concentration is also a relatively well-studied factor that affects robustness, where until a critical molecule-dependent concentration (often in 1-10 mM range) the inhibition increases, afterwards it plateaus or starts to decrease gradually [65, 100, 109]. Manipulating this critical inhibitor concentration is essential for incorporating inhibitors into coatings - when inhibitors leach from a coating matrix onto a defect of a specific size, they have to protect the largest defect area. This distance over which an inhibitor is able to protect a defect effectively is known as the "chemical throwing power", which is crucial in active protective coating design, but not studied for small organic molecules [185].

The influence of time is more sinister, easy to measure but tedious, and most likely for that reason often overlooked. Recent studies highlight that many corrosion inhibitors require a stabilisation period (which can take more than ten hours of exposure) where they gradually form protective layers or reach a stable state, impacting their effectiveness [80, 186]. Their efficiency may decline or even reverse over time due to interactions with environmental elements, leading to possible acceleration of corrosion at prolonged exposure. Therefore, continuous or repeated evaluations of inhibitors over time are essential for accurately determining their long-term performance in corrosion control applications [79]. Inhibition performance can also fluctuate due to transient electrochemical changes on metal surfaces, as revealed by Hilbert spectra analysis, where electrochemical noise patterns reflect the evolution of corrosion processes. Time-resolved electrochemical noise measurements can detect early surface transients that indicate the progression from active corrosion to a more inhibited state, emphasising that inhibitors can vary in effectiveness depending on the duration and characteristics of exposure [187].

There is scarcely any work either on influence of physicochemical or on electrochemical stability. A comparison between the benzotriazole and 2-mercaptobenzothiazole molecules with lithium carbonate has shown that all compounds result in effective corrosion inhibition, but the withdrawal of organics from the environment reverses the corrosion inhibition into an uninhibited case. This highlights the need to select the right organic molecules that can sustain inhibition in changing environmental conditions [113]. Polarisation is a widely used method also to analyse such phenomena, but not to understand how inhibited and uninhibited layers change with respect to overpotentials.

In light of these developments, in this study we have analyzed key factors that influence the effectiveness of corrosion inhibition, which are crucial for determining whether a molecule is a potentially effective and robust corrosion inhibitor. Building upon our previous work where we screened the electrochemical behaviour of AA2024-T3 substrate exposed to more than 100 organic molecules at 1 mM concentration throughout 24 hours [80], we have selected the top-performing non-toxic inhibitors for further study. To highlight the factors critical for corrosion inhibition, we have conducted an electrochemical comparison of ammonium pyrollidinedithiocarbamate and other non-toxic organic molecules with sodium dichromate from different perspectives: the influence of inhibitor concentration, the influence of inhibitor exposure time, the influence of physicochemical stability (inhibitor withdrawal from the environment), and the influence of electrochemical stability (polarisation of the substrate). The results of this study are expected to steer inhibition efficiency and robustness studies and facilitate the development of Cr(VI) replacement organic molecules by unveiling the nature of corrosion inhibition at different and varying conditions.

3.2. Methods

3.2.1. Sample preparation

2 mm thick AA2024-T3 sheets were purchased from Salomon's Metalen B.V., the Netherlands, to use as the substrates for the electrochemical experiments. After cut into 20 mm x 20 mm samples with an automatic shear cutting machine, samples were ground with progressively finer grits of 320, 800, 1200, 2000 and 4000 with a rotating plate sander under a running water. The resulting ground samples were then ultrasonically cleaned in isopropanol for 15 minutes and subsequently dried with compressed air.

3.2.2. Electrolytes

Initial organic molecule choice was based on our previous inhibitor screening study [80]. The chosen inhibitors were the non-toxic molecules with the highest corrosion inhibition efficiencies. Electrochemical measurements were conducted at room temperature in open-to-air 0.1M NaCl solutions, with (or without) the added 1mM inhibitor candidates: 3-amino-1,2,4-triazole-5-thiol, 2-mercaptopyridine, 2-mercaptopyrimidine, 4-mercaptobenzoic acid, ammonium pyrollidinedithiocarbamate, sodium diethyldithiocarbamate. Sodium dichromate dihydrate was used to prepare a 10 mM stock solution, and this solution was diluted with pure water and mixed with NaCl to prepare final solutions for electrochemical experiments. For concentration experiments a range of 0.05 to 10 mM concentration electrolytes were prepared. The basis salt solutions without the addition of inhibitors (pH 5.9) were prepared with NaCl powder with Milli-Q pure water (15.0 MΩ cm resistance at 25 °C). For cyclic voltammetry measurements, 0.1M Na_2SO_4 solution with 1 mM ammonium pyrollidinedithiocarbamate or 3-amino-1,2,4-triazole-5-thiol electrolytes were prepared.No additional compounds were added to modify the pH and/or increase the solubility of inhibitors. All chemicals were obtained from Sigma-Aldrich, except for sodium chloride (J.T. Baker) and 3-amino-5-mercapto-1,2,4-triazole (Alfa-Aesar).

3.2.3. Electrochemical experiments

The electrochemical measurements consisted of the following techniques: open circuit potential (OCP) observation, linear polarisation resistance (LPR), electrochemical impedance spectroscopy (EIS) potentiodynamic polarisation (PDP), and cyclic voltammetry (CV). The experiments were performed in a flat three-electrode electrochemical cell (Corrtest Instruments, China) where the sample was the working, platinum mesh was the counter, and Ag|AgCl (saturated KCl) was the reference electrode. The exposed surface area was 0.785 cm^2, covered by a 250 ml electrolyte volume. Biologic VSP-300 multichannel potentiostats were used to control the electrochemical measurements through EC-Lab software (version 11.33, Biologic, France). Only for sodium dichromate experiments Gamry E1010 potentiostats with Gamry software were used. The electrochemical behaviour of background uninhibited cases were compared to make sure the results between different potentiostats matched.

All electrochemical experiments were repeated at least three times per inhibitor to confirm the reproducibility of the experiments. All potentials presented in this work refer to the Ag|AgCl (saturated KCl) reference potentials unless mentioned differently.

Influence of inhibitor concentration

Inhibitor concentrations were varied from 0.05 to 10 mM for sodium dichromate or ammonium pyrollidinedithiocarbamate dissolved in 0.1M NaCl electrolytes. To check the influence of inhibitor concentration, separate anodic and cathodic potentiodynamic polarisation curves were recorded after 6 hours of OCP in a single sweep with a scan rate of 0.5 mV/s from -(+) 10 mV to +(-) 500 mV potentials with respect to the OCP values. Linear polarisation resistance values were calculated from the initial $\mp$ ($\pm$) 10 mV parts of the scans.

Influence of time

To check the influence of time, potentials were scanned from -10 mV vs. OCP to +10 mV vs. OCP at a rate of 0.5 mV/s every 10 minutes for 24 hours. OCP was observed in between the scans. A linear fit was applied to the observed potential vs. current density plots to obtain the polarisation resistance (R_p) values. At the 2nd and 24th hour, EIS measurements were conducted. A sinusoidal AC perturbation with a peak-to-peak amplitude of 10 mV was applied from 10 kHz to 10 mHz frequency range with 10 frequency point per logarithmic decade. Measurement was repeated 3 times per frequency point. 10 minutes of OCP was observed between LPR and EIS experiments.

Influence of physicochemical stability

To check the influence of physicochemical stability, electrochemical experiments were carried out in inhibitor-containing solutions for the first day, and inhibitor-absent solutions for the last 3 days. The first 24 hours of electrochemical experiments were conducted in 1 mM inhibitors dissolved in 0.1M NaCl solutions. Afterwards the electrolyte was poured out, the electrochemical cell was rinsed, and a new electrolyte containing only 0.1M NaCl solutions was used for the rest of the electrochemical experiments.

The electrochemical investigations were initialised after observing the OCP. LPR was measured over a potential range of ±10 mV with a scan rate of 0.5 mV/s after 1, 2, 6 hours and afterwards every 6 hours. OCP was observed in between LPR measurements. EIS measurements were conducted directly after LPR measurements every 6 hours,in the same manner as discussed previously. The selected data from EIS were quantified with equivalent electrical circuit fitting with the Zview software (v3.5h, Charlottesville, USA).

Influence of electrochemical stability

After observing OCP for 1 hour under exposure to 0.1M Na_2SO_4 with 1 mM ammonium pyrollidinedithiocarbamate or 3-amino-1,2,4-triazole-5-thiol, samples were

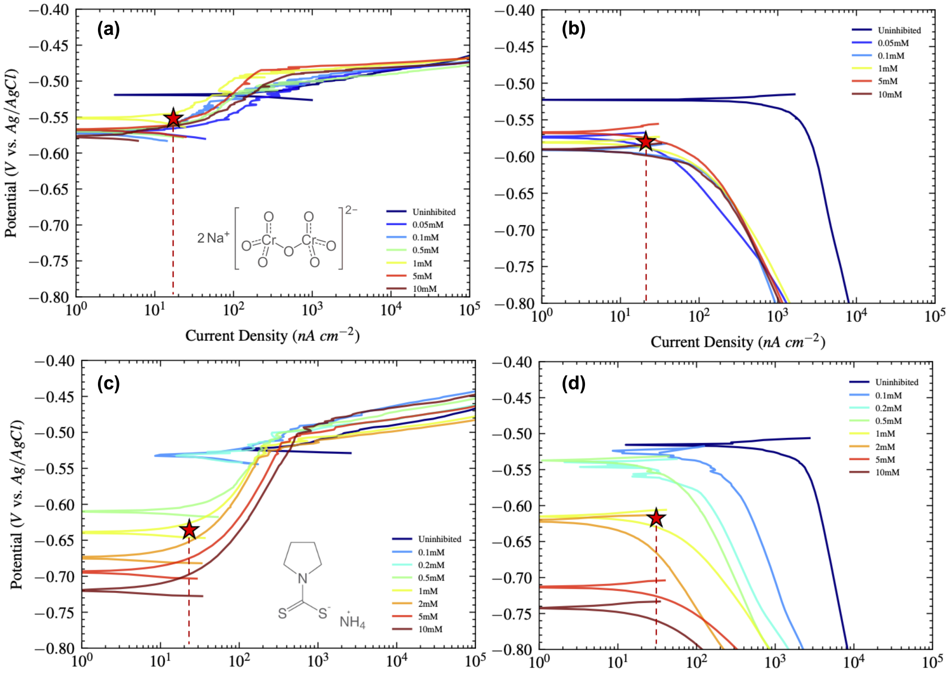


Figure 3.1: Potentiodynamic polarisation curves of AA2024-T3 exposed to 0.1M NaCl electrolytes with varying inhibitor concentrations: (a) sodium dichromate anodic polarisation, (b) sodium dichromate cathodic polarisation, (c) ammonium pyrollidinedithiocarbamate anodic polarisation, (d) ammonium pyrollidinedithiocarbamate cathodic polarisation. Red stars indicate corrosion potentials and currents at 1 mM concentrations.

scanned with 10 mV/s scan rate in a cyclic voltammetry fashion from 0.7 to -1.2 V vs. Ag/AgCl(saturated KCl). The scan was repeated 5 times, but only the first 2 are presented here as the last 4 cycles resulted in the same behaviour.

3.3. RESULTS AND DISCUSSION

3.3.1. INFLUENCE OF INHIBITOR CONCENTRATION

FIGURE 3.1 plots the anodic and cathodic polarisation curves of AA2024-T3 exposed to 0.1M NaCl electrolytes with sodium dichromate or ammonium pyrollidinedithiocarbamate at different concentrations. Figure 3.2 summarises the linear polarisation resistance (R_{LPR}) and corrosion potential (E_{corr}) values obtained from the scans of Figure 3.1.

From Figure 3.1 (a) and (b) and Figure 3.2 potential trends we observe that the addition of sodium dichromate, even as little as 0.05 mM, shifted corrosion potentials to more negative potential values of -50 to -70 mVs. Further concentration increases did not result in further changes in the corrosion potentials. As a result, active pitting behaviour (a rapid increase in anodic corrosion current densities at corrosion potentials) of uninhibited case changed with dichromate additions. For dichromate additions of less

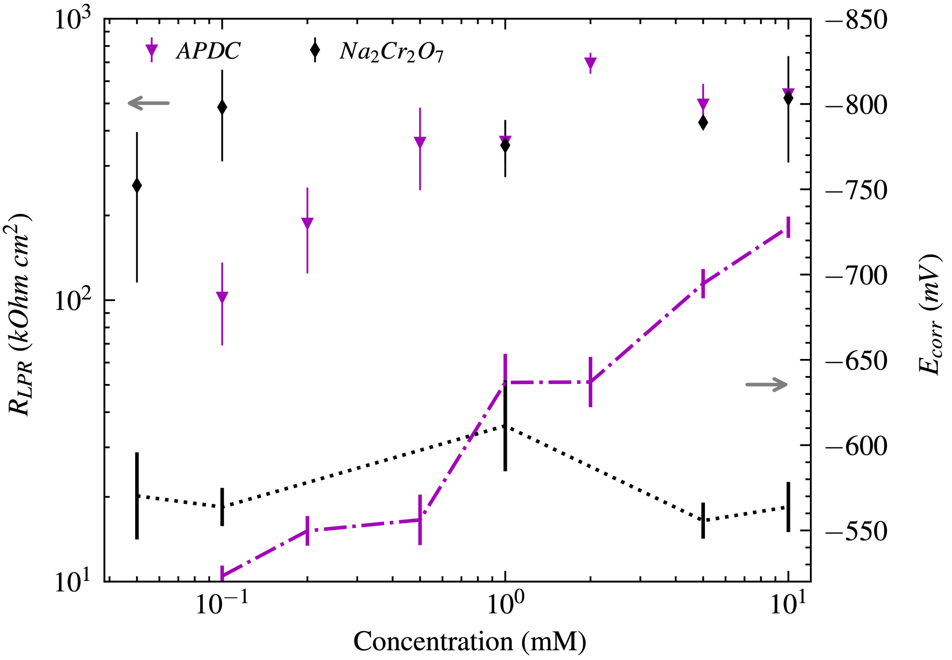


Figure 3.2: Influence of concentration on polarisation resistance and potential values for ammonium pyrollidinedithiocarbamate (APDC) and sodium dichromate ($Na_2Cr_2O_7$). Linear polarisation resistance (R_{LPR}) and corrosion potential (E_{corr}) values obtained from the initial parts of the anodic (cathodic) scans for ∓ (±) 10 mV range. Top markers correspond to the R_{LPR} scale shown on the left axis, bottom lines correspond to the E_{corr} values shown on the right axis. Values corresponding to the bottom plot border (10 kΩ cm^2 and -520 mV Ag/AgCl) correspond to the mean uninhibited case.

than 1mM, even though pitting occurred at uninhibited corrosion potentials, shift of corrosion potentials resulted in a larger stable potential range. For dichromate additions of more than 1mM, an approximately 20 mV extra range of Tafel behaviour appeared, due to the 20 mV positively shifted pitting potentials. This suggests further stabilisation of pits higher than this 1 mM critical concentration. Figure 3.1 (a) shows that increased dichromate concentrations decreased anodic current densities. 0.05 mM dichromate addition decreased corrosion current by an order of magnitude, and this decrease only continued with an increase in concentration. Cathodic curves of Figure 3.1 (b) showed more than 2 orders of magnitude current density decrease with the addition of 0.05 mM dichromate, however no additional decrease in current densities were observed with an increase in dichromate concentration. Cathodic curves showed similar behaviour rather independent of the dichromate concentration. This suggests cathodic inhibition was complete starting from as little as 0.05 mM dichromate concentration. This behaviour aligned with trends observed in R_{LPR} values plotted in Figure 3.2. The addition of sodium dichromate increased R_{LPR} (in kΩ cm^2) values to 256±210 for 0.05 mM, up to a maximum of 428±32 for 5 mM concentration. The standard deviation of measurements decreased with increasing concentrations with the exception of the maximum measured concentration 10 mM, and the mean values with their deviations overlapped between 0.1 - 10 mM concentrations.

Inhibition of corrosion by Cr(VI) compounds is due to their ability to adsorb onto the metal/oxide surfaces irreversibly, subsequently get reduced to form inert hydrophobic Cr(III) oxide barrier films, with retained releasable Cr(VI) reservoirs. Non-reduced Cr(VI) oxoanion adsorbtion on Al-oxides modifies the zeta potential, discouraging adsorption of corrosive ions such as chlorides that promote dissolution and destabilisation of the protective oxide films, further inhibiting pitting [85, 188, 189]. A 0.05 mM chromate concentration was found to be sufficient for the formation of such chemical chromate conversion films: Cr(VI) - Cr(III) mixed oxides which primarily suppress the cathodic oxygen reduction reaction (ORR) rate and inhibit localised corrosion initiation [84, 190]. For AA2024-T3 specifically, the chemical conversion layer thoroughly reduces corrosion activity at both the cathodic sites such as Cu-rich ϑ- and dealloyed S-phase intermetallic particles, and the Al matrix with the anodic intermetallics within [39, 42]. Based on this literature, we can infer that in our experiments 0.05 mM dichromate concentration was sufficient to form a barrier film that suppresses cathodic reactions, which did not change further with an increase in dichromate concentration. Meanwhile, further increase in dichromate concentration increased available Cr(VI) oxoanion ready to adsorb and suppress the localised pitting activity, which would explain the increased potentials required for pitting initiation, and consistent decrease of anodic current densities with increasing dichromate concentrations.

Addition of ammonium pyrollidinedithiocarbamate showed completely different trends. Despite the constant pH values around 6.0-6.5 at all concentrations, the electrochemical potentials changed significantly. Figures 3.1c and 3.1d show that corrosion potentials systematically shifted to more negative potentials with increases in concentration. This resulted in a passive range that became larger and larger with inhibitor concentration, which is expected to limit the localised electrochemical activity. The current densities of both anodic and cathodic curves decreased until 2 mM concentration, after which they started to increase. This is consistent with R_{LPR} values plotted in Figure 3.2, where ammonium pyrollidinedithiocarbamate addition increased R_{LPR} (in kΩ cm^2) values to 102±50 for 0.1 mM, up to a maximum of 697±89 for 2mM, and decreased to 498±132 for 5 mM concentrations. This plateauing or decrease in performance after a certain critical inhibitor concentration was likely due to more disordered self-assembled monolayers [191], which has previously been attributed to surface saturation with adsorbed molecules or self-micelle formation [109, 111, 192]. The decrease in current densities of both anodic and cathodic curves suggests an inhibitive film formed both on anodic Al matrix and anodic/cathodic intermetallics, the majority of which contain Cu [43]. Inhibiting the dealloying of Cu-rich intermetallics is critical in limiting the overall corrosion of AA2024-T3, as they are the main microgalvanic driving force for the electrochemical reactions. Previous studies confirm that ammonium pyrollidinedithiocarbamate can inhibit Cu by decreasing the active surface area and raising the charge transfer resistance through formation of an amorphous inhibitive film [193].

Initial comparison of concentration influence of both inhibitors seems to suggest that electrochemical performance of ammonium pyrollidinedithiocarbamate is on par with sodium dichromate for the 1-10 mM concentration range. It would seem as if our

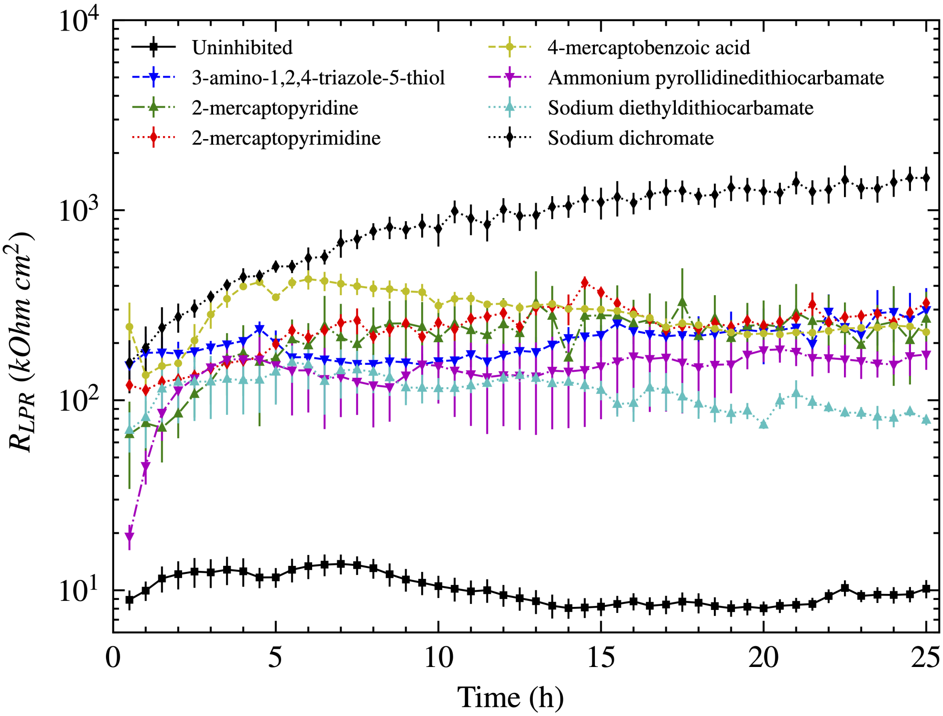


Figure 3.3: Influence of time on the linear polarisation resistance (R_{LPR}) of AA2024-T3 in the presence (and absence) of 1 mM corrosion inhibitors.

research has finally found the replacement for hexavalent chromium compounds. However is that really the case? For this end, the next section explores the behaviour of time on the electrochemical behaviour.

3.3.2. Influence of time

Figure 3.3 plots the polarisation resistance (R_{LPR}) values throughout time for the first 24 hours. In addition to ammonium pyrrolidinedithiocarbamate, we tested 5 other non-toxic organic molecules that had shown promising corrosion inhibition properties during our previous screening [80]. R_{LPR} values of the uninhibited case were relatively constant around 10±3 kΩ cm^2 throughout the first day. The organic molecules increased the R_{LPR} values in the range of 78±9 to 325±10 kΩ cm^2, but not immediately. It is observed that organic molecules require some time to stabilise and reach their peak polarisation resistance R_{LPR} values, which was around 6 hours. After that point, R_{LPR} values reached a plateau and did not change significantly anymore. On the other hand, polarisation resistance originating from the sodium dichromate kept increasing throughout the whole day, up to 1479±431 kΩ cm^2.

Looking back on the concentration experiments presented in the previous section, we can explain the comparable behaviour of sodium dichromate with the organic molecules observed after 6 hours. Whereas the ratio of polarisation resistance values of sodium dichromate and ammonium pyrollidinedithiocarbamate was around 2-3 around the 6-hour mark, which matches which trends from last section, this value increases to 8-10 at the 24th hour. This corresponds to inhibition efficiencies of 87-97% for organics, while sodium dichromate reached an inhibition efficiency of more than 99%. This shows that if we only look at the time-step of 6th hour, or any other single time-step for that matter, we come to the wrong conclusion about the behaviour. The gradual initial increase culminating in a plateau of polarisation resistance for the adsorption of organic molecules, and continuous development of the protective chromium oxide films highlight the time-sensitive nature of the corrosion inhibition, and the critical need for time-resolved measurements. The importance of time-resolved electrochemical measurements were highlighted before in a previous study [79]; here we once again underline that without time-resolved measurements, it is unlikely to have a correct efficacy assessment of the next-generation chromate replacement compounds.

3.3.3. INFLUENCE OF PHYSICOCHEMICAL STABILITY

SUSTAINING corrosion inhibition in changing environments is as important as sustaining corrosion inhibition throughout time. It is critical to keep corrosion inhibition going in dynamic conditions, especially in the widely-changing conditions observed for aerospace alloys: dry-wet cycles, temperature fluctuations, among others [194]. We name the sustained inhibition in the changing environmental conditions *physicochemical stability* of the inhibitor, which in previous papers were also called irreversibility of the inhibition [113].

To check the behaviour of physicochemical stability we have observed electrochemical impedance response of selected high performing organic corrosion inhibitors and sodium dichromate at 1 mM concentrations. Electrochemical impedance spectroscopy measurements were first performed in the presence of inhibitors after 12, 18 and 24 hours of exposure, afterwards substrates were exposed to electrolytes without any inhibitors and electrochemical impedance spectra were acquired for every 6 hours after the 12th hour for 3 days.

Figure 3.4 (a) shows the resulting impedance modulus and phase angle plots for the final measurements before and after electrolyte switch. Filled markers represent the case in the presence of inhibitors, and empty markers represent the case of the following absence of inhibitors. It is visible from the plots that the addition of inhibitors consistently increases the impedance modulus values. Sodium dichromate results in the largest impedance modulus increase. After changing the electrolytes, impedance modulus of all systems decrease significantly: the impedance modulus of all organic inhibitor systems except for 3-amino-1,2,4-triazole-5-thiol drop down to the uninhibited level, while sodium dichromate shows significantly higher modulus values. Even after the electrolyte exchange the impedance modulus values of sodium dichromate only drop down to the levels of organic inhibitor present systems. Figure 3.4 (b) shows the mean drop in impedance modulus values at 10^{-2} Hz converted into inhibition power. It is clear that apart from 3-amino-1,2,4-triazole-5-thiol, all organic molecules stop providing corrosion inhibition if they are not sustained in the environment. In this case 3-amino-1,2,4-triazole-5-thiol loses most of its inhibition as well - 74% of the initial inhibition

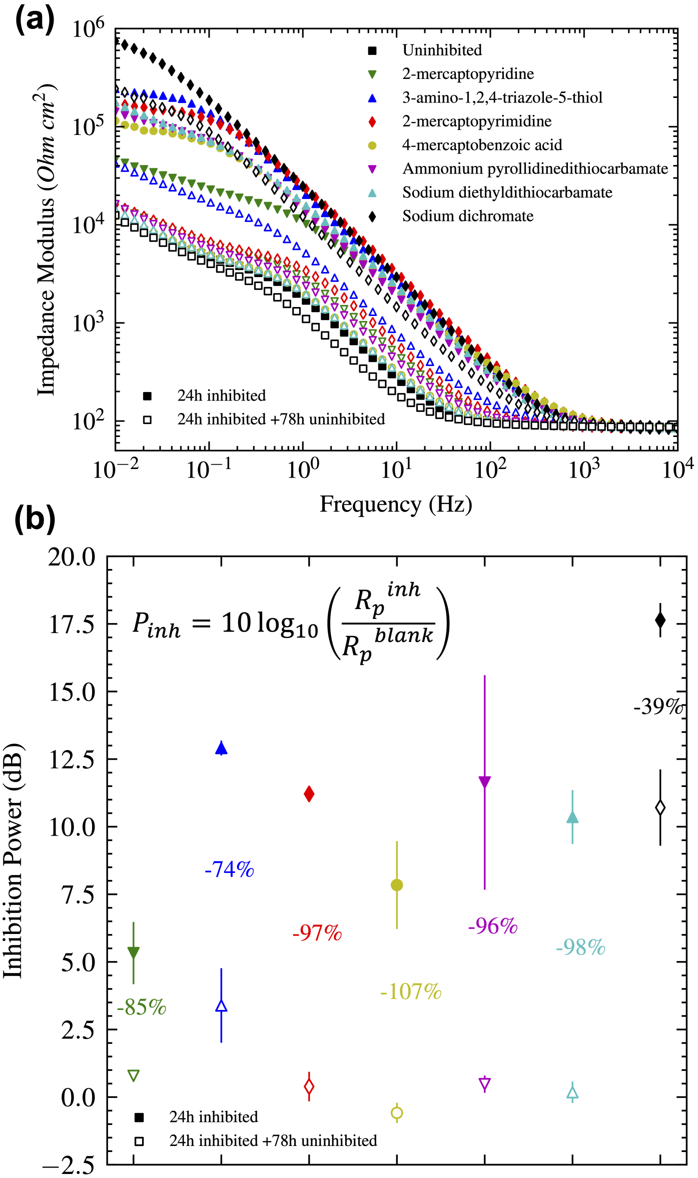


Figure 3.4: Influence of presence and subsequent absence of corrosion inhibitors on electrochemical behaviour. Filled markers represent the electrochemical impedance spectroscopy response before electrolyte exchange, while empty markers indicate the response after exchange. (a) Impedance modulus spectra for organic inhibitors. (b) Comparison of the impedance modulus at 10^{-2} Hz, converted to inhibition power (P_{inh}), before and after electrolyte exchange. The percentage reduction in original protection is indicated between the markers. Both subplots share the same legends.

power is lost - but it is not completely gone, resulting in a quasi-reversible corrosion inhibition behaviour. For comparison, sodium dichromate only loses 39% of its original inhibition power.

Figure 3.5 focuses on sodium dichromate, ammonium pyrollidinedithiocarbamate, and 3-amino-1,2,4-triazole-5-thiol. 3-amino-1,2,4-triazole-5-thiol still sustained some impedance modulus increase after the electrolyte switch. This is a key quality for stable and irreversible corrosion inhibition, as the low-frequency impedance determines the total retained corrosion resistance of the system [79]. As the frequency tends towards infinity, the impedance modulus magnitude tends towards the resistance of the electrolyte; as the frequency tends towards zero, capacitive contributions disappear and the impedance modulus magnitude tends toward the total impedance coming from the electrolyte, inhibitor, and charge transfer [195, 196]. Similar impedance behaviour between samples above 10^3 Hz stemmed from electrolyte impedances, resulting in similar impedance modulus values. The phase angle values became more negative as the frequency decreased: the more the corresponding impedance modulus, the steeper the phase angle decrease. This was the result of the capacitive dielectric formed on the substrate through oxide and/or adsorbed inhibitors. Related impedance modulus increase and more capacitive behaviour between 10^{-1} - 10^3 Hz stemmed from the electron transfer processes of the inhibitor-oxide/metal surface [197]. The behaviour in further lower frequencies corresponds to either resistive charge transfer processes where phase angle approaches 0, or otherwise mass-transfer limited diffusion processes where phase angle approaches -45 [195, 196]. The slowest time constant at lower frequencies ($\omega_{char} \sim 1/\tau = 1/RC$), which is the measure of how quickly the system responds to external changes in voltage and current [198], appeared at lower frequencies in the presence of inhibitors, which means the time constant has increased. This increase meant slowing down the electrochemical system, either through an increase in resistance or capacitance, through limiting charge transfer or diffusion. After the electrolyte switch time constants decreased again for all systems. The low frequency phase angles of uninhibited and ammonium pyrollidinedithiocarbamate samples approached towards -45° (also apparent as a high-frequency slope of 1 in Nyquist visualisation, not shown here), which suggests a diffusion-limited response. For others that was not the case. These suggest that a complete or quasi-reversal to the non-protected behaviour develops in the absence of a sustained inhibitor in the environment. The difference most likely originates from the different surface bonding behaviour of organic molecules. The inhibitors that maintained their effectiveness formed stable surface complexes or stabilised oxide layers that resisted their removal and/or dissolution of the substrate after the electrolyte switch.

To quantify inhibitors' electrochemical response, a modified Randles circuit shown in Figure 3.5 inset is used as an equivalent electrical circuit to fit the spectra. The chosen circuit with two time constants was used to model the physics of the metal electrode covered with an imperfect overlaying inhibitor layer. This is a widely used equivalent circuit fit used for modeling the impedance of an electrode coated by a thick dielectric layer with pores exposing the electrode to the electrolyte [98, 105, 195, 199]. In this fit R_s, R_1 and CPE_1, R_2, CPE_2 corresponded to the electrolyte resistance, protective film resistance (through adsorbed molecules and/or passive film) and its associated capac-

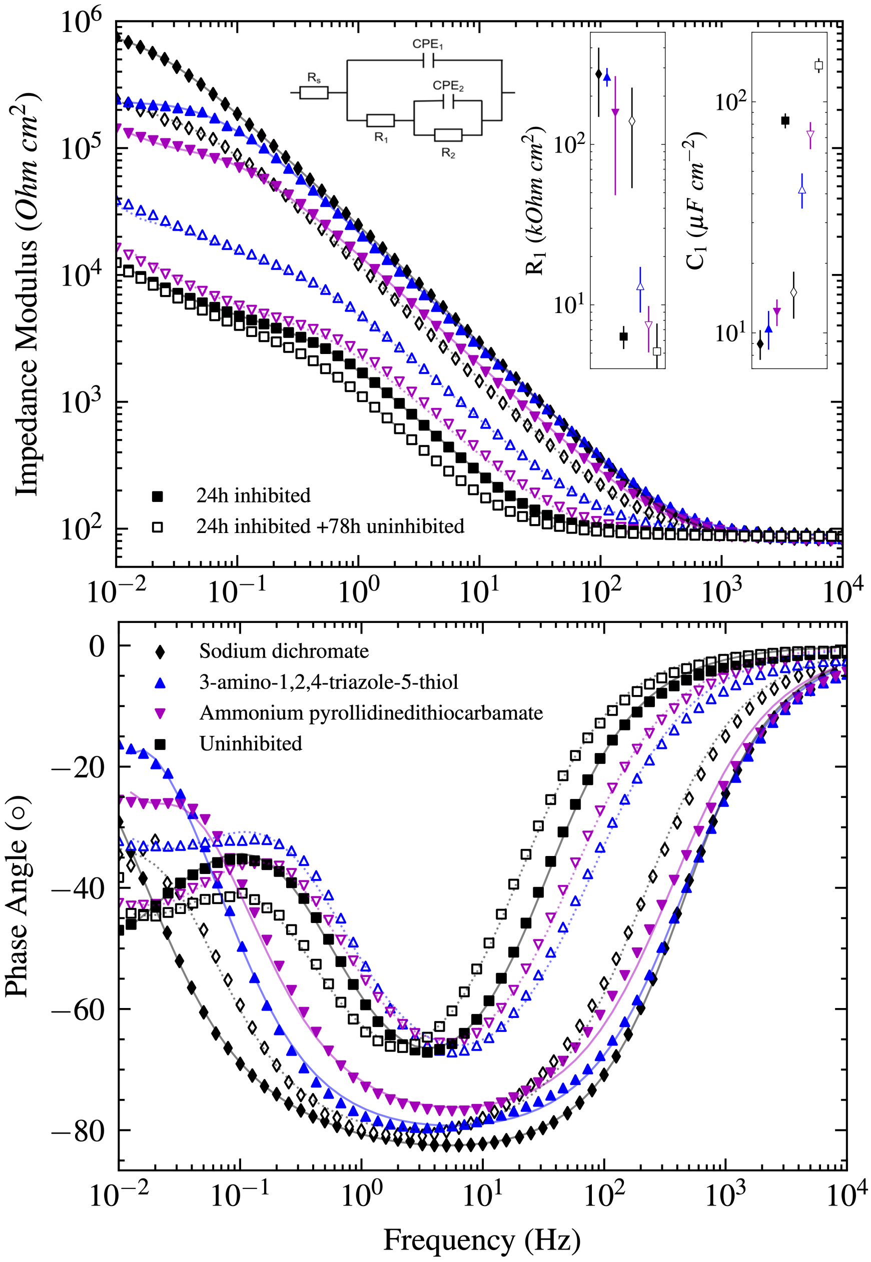


Figure 3.5: Impedance modulus and phase angle plots with equivalent circuit fits demonstrating the influence of presence and subsequent absence of corrosion inhibitors on electrochemical impedance spectroscopy response. Selected equivalent circuit and fit values relevant to the inhibition shown in the inset.

itance, charge transfer resistance and the double layer capacitance, respectively. Constant phase elements (CPE) are employed instead of capacitors due to the deviation from the ideal capacitive behaviour. Capacitance of the constant phase elements were calculated according to the Hsu-Mansfeld approach [200]:

$$C = R^{\frac{1-n}{n}} Q^{\frac{1}{n}} \tag{3.1}$$

where C is the capacitance, R the resistance, Q is the magnitude of the CPE associated with its capacitance, and n an empirical constant, taking values between 0 and 1 (1 represents the case for the ideal capacitor, 0 the ideal resistor, and in between values the non-ideal capacitive responses). The calculated equivalent resistance and capacitance values related to the presence/absence of inhibitors are plotted as an inset of figure 3.5. Resistance values increased up to 50-fold in the presence of inhibitors. The capacitance values showed an order of magnitude decrease in the presence of inhibitors as well. Through the relationship: [195]:

$$C = \frac{\varepsilon_0 \varepsilon}{d} \tag{3.2}$$

where C is the capacitance, ε_0 the vacuum permittivity, ε is the relative permittivity, and d the thickness of the dielectric field responsible for capacitive behaviour, it can be argued that corrosion inhibitors either created a steric hindrance through a thicker barrier film, or decreased the relative permittivity of the surface. After the electrolyte change the resistance values had a sharp decline: dichromate resistance dropped to half of its original value but stayed strongly inhibitive, 3-amino-1,2,4-triazole-5-thiol dropped to a fraction of its original gained inhibition and showed around 65% inhibition efficiency, whereas ammonium pyrollidinedithiocarbamate lost all inhibition. Trends were similar for capacitance: capacitance of dichromate doubled, 3-amino-1,2,4-triazole-5-thiol quadrupled, whereas ammonium pyrollidinedithiocarbamate returned to uninhibited values. Doubling of capacitances for the uninhibited case is most likely resulting from the growth of the Al-oxide under the 3-day electrolyte exposure.

Figure 3.6 plots the evolution of impedance modulus values measured at 10^{-2} Hz frequency to understand the stability of inhibitor systems through time. Although being based on a simplification since the low-frequency impedance modulus includes contributions from the oxide film resistance, the charge transfer resistance, and often from the diffusion-controlled processes - it has been shown that the impedance modulus values observed at 10^{-2} Hz frequency effectively represent the corrosion resistance of the inhibitor–substrate interface [143].

It is observed that the electrochemical behaviour of almost all samples returns to the uninhibited performance 60 hours after the inhibitor removal. 3-amino-1,2,4-triazole-5-thiol sustains its -albeit decreased - protection at least for 3 days after the electrolyte exchange, but all other organics completely lose their protection. Dichromate sustains its original protection for a long time, and even after 3 days measured impedance modulus is more than 5-fold the impedance modulus of the best organic corrosion inhibitor.

These observations suggest that corrosion inhibition gained through organic molecules is lost for almost every organic system if the molecule is not sustained in the environment. Despite the initial inhibition, the majority of the tested organic molecules

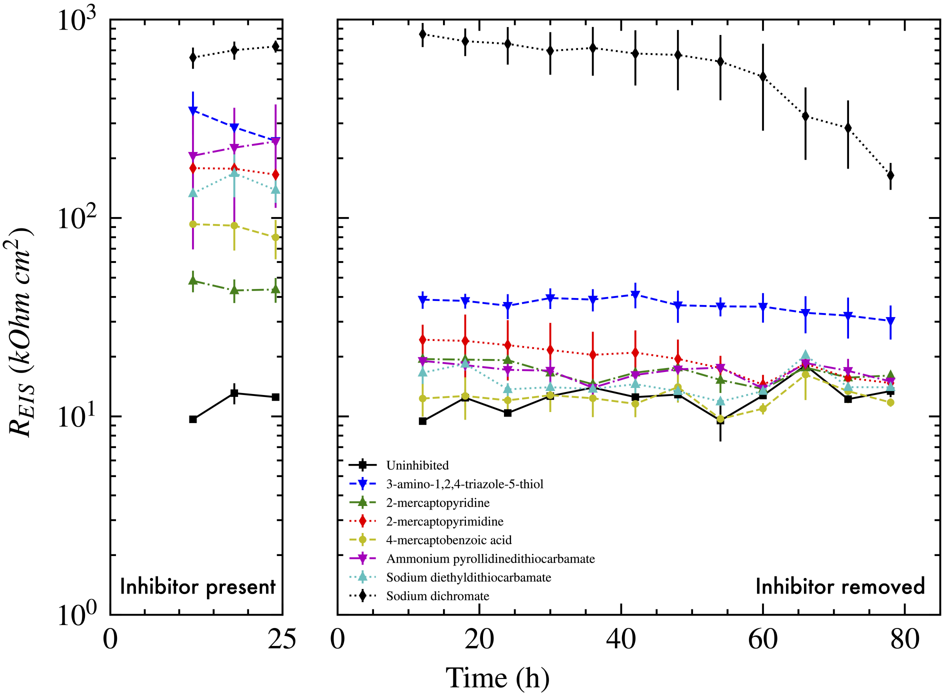


Figure 3.6: Evolution of impedance modulus values measured at 10^{-2} Hz frequency.

have reversible bonds that limit their inhibition performance and applicability in dynamic environments. Best-performing inhibitors were not necessarily more irreversible or had a higher performance after electrolyte change. 3-amino-1,2,4-triazole-5-thiol provides a quasi-reversible corrosion inhibition, possibly through more permanent bonds formed with some of the intermetallics instead of the Al substrate. Sodium dichromate showed the best inhibition performance before and after the electrolyte exchange, but it also showed a significant decrease in inhibition. However, even when the dichromate was absent afterwards, the inhibition was better than the best organic inhibitor tested in this study.

3.3.4. Influence of electrochemical stability

Corrosion inhibition must be sustained in a wide range of electrochemical potentials. In the vicinity of the localised galvanic couples, such as pitting corrosion cells of AA2024-T3 [42], open circuit potential is different because as the electrochemical corrosion reactions proceed anodic areas become more acidic while cathodic areas become more basic, both destabilising the oxide of aluminium alloys. Ensuring corrosion inhibition in a wide range of potentials minimises such microgalvanic interactions, which we here label as the *electrochemical stability*.

To check the behaviour of electrochemical stability we have performed cyclic voltam-

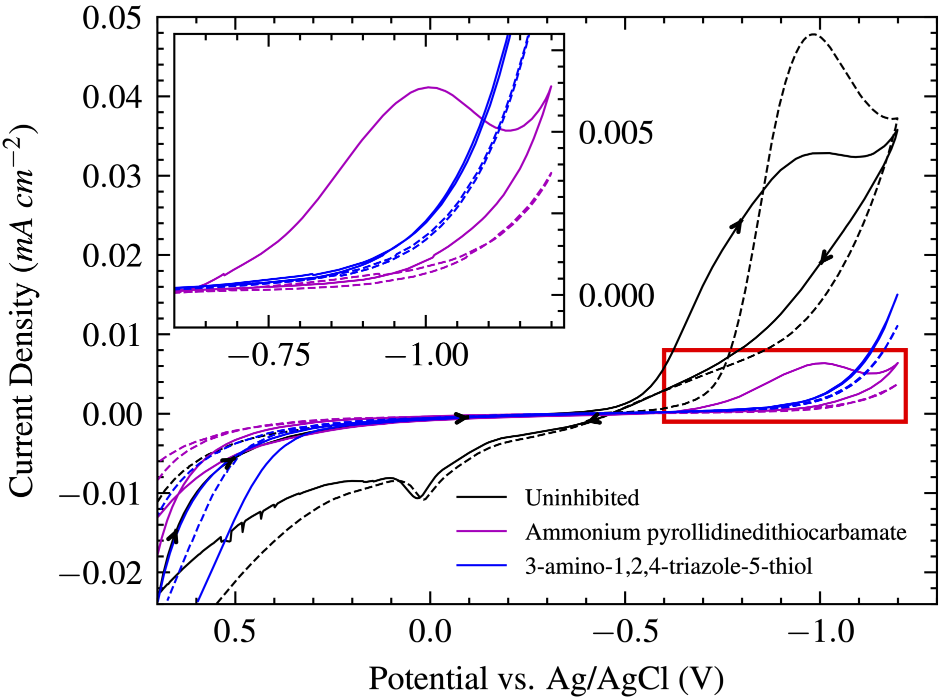


Figure 3.7: Cyclic voltammetry measurements of AA2024-T3 in the presence (and absence) of 1 mM corrosion inhibitors. Solid lines show the first, dashed lines show the second cycle. Inset figure is the close-up of the red framed area in the cathodic overpotential region.

metry measurements. In total 5 cycles were recorded, but no significant change in electrochemical behaviour is observed so only the first 2 cycles are plotted in figure 3.7. Cycles are initiated from positive towards the negative potentials, with the hypothesis that initially the inhibitors would form self-assembled monolayers in the first hour of exposure, then they would be forcibly desorbed throughout the scan to the negative potentials (assuming deprotonated negatively charged molecules, which is justified given the low pKa trend of mercaptans [201]), then electrosorbed [202] again to the aluminium surface with the positive scan. The first and second scan would show the difference between self assembled monolayer and electrosorption behaviour.

On the scan towards negative potentials, for the uninhibited case a peak appears around -0.55 V, which shifts to -0.65 V for the second cycle. The onset values of these peaks are typical for diffusion limited oxygen reduction reaction (ORR), which extend up to around -1.1 V where hydrogen evolution reaction appears as a sharp increase in cathodic current densities [203]. ORR is dependent on the surface properties and composition - because of the surface modifications to Al (hydr)oxide during the first scan, the ORR onset shifts to more negative potentials and result in higher peak current densities. Looking at the results for organic inhibitors during the first scan, ORR is partially suppressed for ammonium pyrollidinedithiocarbamate, and completely suppressed for

3-amino-1,2,4-triazole-5-thiol after the self-assembly process of the adsorbed organic layers. After the second cycle, the current densities decrease even more and the peak of ammonium pyrollidinedithiocarbamate present in the first scan disappears, suggesting increased inhibition through the electrosorption.

On the scan back towards positive potentials, for the uninhibited case a peak appears around -50 mV. The position and magnitude of this peak matches very well with literature where cyclic voltammetry and glow discharge mass spectrometry (GDMS) was used on AA2024 [204], with which this peak was attributed to surface enrichment with Cu due to the the anodic Cu oxidation reactions:

$$2Cu + H_2O \rightarrow Cu_2O + 2H^+ + 2e^- \tag{3.3}$$

$$Cu_2O + H_2O \rightarrow 2CuO + 2H^+ + 2e^- \tag{3.4}$$

For the uninhibited case both cycles had this peak, which was completely suppressed in the presence of organic inhibitors. This would mean that organic inhibitors are successful in preventing surface enrichment with Cu, in both self-assembled and electrosorbed form, which is a critical corrosion initiation mechanism for Al-Cu alloys. Both molecules conveyed stable inhibition in a wide potential range.

3.4. Conclusions

In order to support the quest for promising non-toxic alternatives to hexavalent chromium based inhibitors, we have to be aware of how the corrosion inhibition evolves with inhibitor concentration and measurement time, and whether it is stable in the presence/absence of the inhibitor molecule in a wide potential range. Here we show that even after screening more than 100 organic molecules experimentally, the best-performing molecules from the screening can tempt us to jump to premature conclusions. When not taking different corrosion inhibition critical factors such as time, concentration and physical/electrochemical stability into consideration, conclusive remarks about final performance cannot be drawn. We have observed that at 1 mM concentrations some organic compounds do offer comparable inhibition to sodium dichromate around the 6-hour mark, but afterwards performance of chromate keeps increasing throughout the first day whereas organics reach a stable plateau. Despite the initial inhibition, the majority of the tested organic molecules have reversible bonds that limit their inhibition performance and applicability in dynamic environments where a constant inhibitor reservoir is not present. Compared to weaker inhibitors, best-performing inhibitors were neither necessarily more physically stable and irreversible, nor had a higher performance after removal from the electrolyte. On the other hand, when the organic molecules are sustained in the environment they can offer corrosion inhibition away from the open circuit potential for a wide potential range, and can suppress both Cu oxidation and oxygen reduction reactions. Some minority molecules show that quasi-stable corrosion inhibition is possible with small organic molecules - meaning long-term (studied up to 3-days in this work) stable barrier properties are possible even when the molecules are withdrawn from the environment, albeit at a lower inhibition.

4

SUSTAINING INHIBITION

Insight must precede application.

Max Planck

The dream corrosion inhibitor would work for every substrate-environment combination, and the protection would be sustained indefinitely with an irreversible barrier layer when exposed to aggressive and changing environmental conditions. However our prior electrochemical experiments on AA2024-T3 have shown that despite the initial inhibition, all of the tested molecules had reversible bonds that limit their inhibition performance and applicability in dynamic environments, with the exception of 3-amino-1,2,4-triazole-5-thiol, which still showed 42% inhibition efficiency after being exposed to 0.1 M NaCl only for three days. Potentiodynamic polarisation, atomic force microscopy and scanning Kelvin probe force microscopy (AFM/SKPFM), X-ray photoelectron spectroscopy (XPS), attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR), shell-isolated nanoparticle-enhanced Raman spectroscopy (SHINERS), and time-of-flight secondary ion mass spectrometry (ToF-SIMS) complemented by density functional theory (DFT) calculations were used to identify the molecular mechanism responsible for the quasi-stable adsorption provided by 3-amino-1,2,4-triazole-5-thiol. Our findings suggest that a sulphatisation of the Al-(hydr)oxide is the key contributor to the quasi-sustained corrosion inhibition. Sustained molecule adsorption over intermetallics in trace amounts was also observed, but their presence was insufficient to inhibit corrosion.

This chapter has been published with some adjustments as *Quasi-stable adsorption as a stepping stone to stable corrosion inhibition*, Applied Surface Science (2025) [205].

4.1. INTRODUCTION

A corrosion inhibitor is a compound that reduces the corrosion rate of a metallic substrate exposed to an aggressive environment, when it is present in the environment in sufficient but minute amount [65]. By prolonging the service life of materials, corrosion inhibitors reduce maintenance costs and minimise process downtime in various industries. Over their service lifetime of about 30 years, aerospace components face harsh humidity, salt and temperature fluctuations - which in the absence of inhibitors may cause catastrophic failure through stress corrosion cracking and fatigue [194]. Without the application of corrosion inhibitors, metal contacts of photovoltaic solar cell elements degrade, heat exchanger piping severely corrode, energy storage capabilities of batteries decrease due to electrode materials interacting with highly conductive and aggressive electrolytes. The ubiquitous need for corrosion protection is becoming even more critical in an era where sustainable computation and renewable energy are predisposed to replace the traditional oil and gas-based economy. This transition drives the emergence of new industries and technologies in the areas of nuclear energy, carbon capture systems, and lightweight vehicle design - all of which introduce fresh demands for advanced corrosion protection which will only further fuel the growth of a corrosion inhibitor market already valued at over US$ 8 billion [206].

The ideal corrosion inhibitor would be universal: it would work in all aggressive media and substrates to be protected; and perpetual: it would keep working in changing environmental conditions. Organic molecules have demonstrated significant efficacy as corrosion inhibitors across various substrates [1, 109, 110, 207–212] - their endless structural versatility inspires optimism for one day identifying a universal approach that would inhibit corrosion in all electrode - electrolyte systems. Recent works have been conducting searches in chemical spaces for a limited version of this dream for optimising inhibitors for specific alloy - environment systems, many of which capitalising on the recent developments in machine learning [31, 34, 35, 80, 136, 139, 168, 180, 181, 213–217]. However, the search for the perpetual molecule is a solemn affair, as far as the authors' knowledge goes no work has been done to systematically evaluate the potential of sustained corrosion inhibition of organic molecules in changing and dynamic environments.

Regrettably on the contrary, previous studies have shown that corrosion inhibition provided by an organic molecule is often compromised when its continuous presence in the environment cannot be maintained. For an AA2024-T3 substrate, a complete loss of previously gained corrosion inhibition in the subsequent absence of molecules in the environment has been shown for 2-mercaptobenzothiazole, 1,2,4-triazole, 3-amino-1,2,4-triazole, benzotriazole, 4-mercaptobenzoic acid, 2-mercaptopyridine, 2-mercapto- pyrimidine, ammonium pyrollidinedithiocarbamate, and sodium diethyldithiocarbamate [101, 102, 113, 169, 218]. Ideally, a one-time corrosion inhibitor application should provide prolonged protection under dynamic and often harsh conditions. This requirement is especially important for ensuring the longevity and reliability of materials used in environments where maintenance or inhibitor reapplication is challenging or impractical. The fact that many inhibitors rely on a persistent supply to sustain their protective properties is disconcerting, which calls for a deeper understanding of how organic inhibitors can form stable, long-lasting layers without the need for constant replenishment.

In this work, we tackle this question of corrosion inhibition stability offered by organic molecules. Organic molecules mainly inhibit surfaces by forming an insoluble complex or polymeric film via physisorption or chemisorption, providing a steric and/or potential barrier for corrosive species [76, 113]. During our previous studies we have observed that this sort of interaction is often transient, and organic molecules lose previously gained corrosion inhibition efficacy when they are no longer sustained in the environment [169]. The only exception we identified previously was the particular molecule 3-amino-1,2,4-triazole-5-thiol, which continues to inhibit corrosion, albeit at a reduced efficacy when it is no longer supplied in the environment. In this work, we aimed to illuminate this phenomenon. We begin by characterising the electrochemical behaviour of AA2024-T3 substrate in the presence and subsequent absence of the molecule, establishing a foundation for understanding its interaction with the substrate. Subsequent local electrochemical analyses using AFM/SKPFM reveal that the protection is not confined solely to intermetallic zones but is rather uniformly distributed. To pinpoint the molecular features responsible for this quasi-stable adsorption, we employ surface spectroscopy techniques of XPS, ATR-FTIR, SHINERS, and ToF-SIMS. Complementing our experimental observations, molecular speciation and DFT calculations provide a theoretical framework to rationalise these findings. Ultimately, by deciphering the source of the quasi-sustained corrosion protection, our work aims to inform the design of more robust and enduring corrosion inhibition systems.

4.2. Materials and Methods

4.2.1. Sample preparation

2 mm thick AA2024-T3 sheets, 1 mm thick pure Cu, and 3 mm thick AA1050 alloys were used as the substrates for the experiments. The samples were first cut into 20 mm × 20 mm pieces using an automatic shear cutting machine. They were then ground with sandpapers (Struers waterproof SiC) with increasingly finer grits of 320, 800, 1200, 2000, and 4000 using a rotating plate sander under running water. After grinding, the samples underwent ultrasonic cleaning in isopropanol for 15 minutes and were then dried with compressed air.

4.2.2. Electrolyte exposure

3-amino-1,2,4-triazole-5-thiol was selected to be the corrosion inhibitor to be analysed based on the results from our previous work [169], which demonstrated a quasi-stable corrosion inhibition behaviour in the subsequent absence of inhibitor in the environment. The molecule was purchased from Alfa Aesar, which had >98% purity.

1 mM molecule containing solutions were prepared with Milli-Q pure water (15.0 MΩ cm resistance at 25 °C). Samples are first exposed to the molecule-containing solution for 24 hours, and afterwards exposed to only water-containing solutions. For the electrochemical measurements solutions also contained 0.1 M NaCl. For the subsequent molecule absence experiments the exposure environment conditions varied depending on the experimental technique. AFM/SKPFM, XPS, and ATR-FTIR experiment samples were exposed to 2 hours of only water exposure. For samples

analysed with SHINERS the process of water exposure was followed in-situ. Furthermore, to check the influence of solvent the experiments were also repeated for THF instead of water as solvent, which resulted in no significant differences. For ToF-SIMS measurements, after 24 hours of molecule exposure the samples are rinsed briefly (~1 min) or extensively (~1 hr), and dried under a nitrogen stream.

4.2.3. ELECTROCHEMICAL MEASUREMENTS

THE experiments were conducted in a flat three-electrode electrochemical cell (Corrtest Instruments, China), where the sample served as the working electrode, a platinum mesh was used as the counter electrode, and an Ag–AgCl (saturated KCl) electrode was used as the reference. The working electrode had an exposed surface area of 0.785 cm^2 and was immersed in 250 mL of electrolyte. Electrochemical measurements were controlled using Biologic VSP-300 multichannel potentiostats with EC-Lab software (version 11.33, Biologic, France).

To assess corrosion inhibition stability, experiments were performed on samples that were first exposed to inhibitor-containing solutions for one day, followed by exposure to inhibitor-free solutions for three days. For the inhibitor-containing experiments, 1 mM inhibitors were dissolved in 0.1 M NaCl. For the inhibitor-free experiments, after the initial one-day exposure to the inhibitor-containing electrolyte, the initial solution was poured out, the cell was rinsed, and a fresh 0.1 M NaCl solution (without inhibitor) was used for the subsequent three-day period. Potentiodynamic polarisation experiments were then conducted at the end of exposures with polarisation curves recorded in a single sweep at a scan rate of 0.5 mV s^{-1}, covering a potential range from -250 mV (cathodic) to +250 mV (anodic) relative to the open circuit potential. Corrosion potentials and current densities were calculated using Tafel extrapolation by finding the intersection of the potential where the lowest current density observed with the tangents from the linear portions of the anodic and cathodic sections of the log|current density|-potential curves.

All electrochemical experiments were repeated at least three times per inhibitor to ensure reproducibility. Unless stated otherwise, all potentials reported in this work are referenced to the Ag–AgCl (saturated KCl) electrode.

The inhibition efficiencies were calculated from corrosion current densities of inhibited and uninhibited samples with the equation:

$$\eta = \frac{j_{corr}{}^{uninh} - j_{corr}{}^{inh}}{j_{corr}{}^{uninh}} = (1 - \frac{j_{corr}{}^{inh}}{j_{corr}{}^{uninh}}) \times 100\% \tag{4.1}$$

where superscripts uninh and inh stand for uninhibited and inhibited samples, respectively.

4.2.4. ATOMIC FORCE MICROSCOPY (AFM) / SCANNING KELVIN PROBE FORCE MICROSCOPY (SKPFM)

TO gain a comprehensive understanding of the topographical features and electrical surface potential/charge distribution of the adsorbed layer of an organic molecule on the aluminium alloy surface, atomic force microscopy (AFM) and high-surface sensitive scanning Kelvin probe force microscopy (SKPFM) were performed. Each sample was half-submerged in an inhibitor-containing solution for 24 hours, so that one part of the alloy remained untreated. The sample in this state was referred to as the “molecule-present" case, where the boundary between the bare alloy and the region exposed to the inhibitor was analyzed. Following this exposure to inhibitor-containing solutions, the samples were half-submerged in a second solution without any inhibitor for 2 hours. The sample in this state was referred to as the “molecule-absent" case, where the boundary between the bare alloy and the region first exposed to the inhibitor and subsequently exposed to the inhibitor-free solution was analyzed.

AFM and SKPFM mappings were carried out using a Bruker Dimension Edge instrument, equipped with an antimony (n)-doped silicon pyramid single-crystal tip coated with PtIr5 (SCM-Pit-V2 probe). The probe featured a tip radius of 25 nm and a height of 10–15 µm. The surface potential/charge was mapped using a dual-scan approach. During the first scan, topographical data were recorded in dynamic (tapping) mode. In the subsequent scan, the tip was elevated by 50 nm to measure the surface potential, maintaining alignment with the topographical contour captured in the initial scan. All measurements were conducted ex-situ under controlled conditions (ambient air at 22 °C, relative humidity ~ 40 %). All AFM/SKPFM measurements were performed with a resolution of 512 × 512 pixels, a zero-DC bias voltage, and a scan frequency of 0.3 Hz. An AC voltage of 6 V was applied to the tip. To ensure accuracy and eliminate variability due to probe sensitivity, the same tip was used consistently for all measurements conducted within the same day.

4.2.5. X-RAY PHOTOELECTRON SPECTROSCOPY (XPS)

THE AA2024-T3 surfaces exposed to inhibitor-containing and subsequent inhibitor-absent solutions were studied using PHI 5400 ESCA system supplied by Physical Electronic, Inc. This system is equipped with a non-monochromatised aluminium (Al) Kα X-ray source (hv = 1486.7 eV), operated at 200 W power and 13.5 kV accelerating voltage, with an analyzer work function of 4.25 eV. During measurements, the pressure within the sample (analysis) chamber was maintained at 10^{-9} mbar.

For the full survey acquisition of the samples, the pass energy of the analyser was set at 89.45 eV (with 0.5 eV resolution), whereas during the high resolution scans, the pass energy was set at 71.55 eV (0.1 eV resolution for N1s and S2p, 0.2 eV for the rest). Importantly, the take-off angle during both the high-resolution and full-survey measurements was maintained at 45°. All specimens were studied on a circular scanning area with a diameter of 0.4 mm, and their theoretical depth of analysis was 3-5 nm. In order to compensate for the charging of the specimens during the XPS analysis, the high-resolution spectra were peak adjusted through the adventitious carbon shift, during which the reference C-C peak of the C1s spectrum was set to 284.8 eV, and other spectra were offset accordingly. All the processing of the XPS spectra was carried out using the MultiPak version 8.0 software from Physical Electronics, Inc. The curve fitting and decomposition were done by Shirley-type background removal. A constrained fitting procedure was used in which the mixed Gauss–Lorentz shapes for the different fit components in the peaks were allowed to change in the 80–100 % region. Only small variations in peak position and full widths at half-maximum (FWHM) were permitted.

4.2.6. Attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR)

ATR-FTIR was performed using a Thermo Nicolet Nexus 470 FTIR spectrometer equipped with a liquid nitrogen-cooled mercury cadmium telluride (MCT) detector. A Smart Golden Gate ATR accessory with a diamond crystal was employed for sample analysis. Prior to measurements, the stability of the MCT detector was monitored by checking background stability over time.

Samples were prepared by directly placing them onto the diamond ATR crystal, where a gentle pressure was applied using the built-in clamp to ensure optimal contact with the crystal. Prior to the exposure of samples to inhibitor-containing and subsequent inhibitor-absent solutions, background spectra were acquired from freshly prepared samples not exposed to any molecules to account for atmospheric, instrumental, and substrate related interferences.

Infrared spectra were collected in the mid-infrared region (4000–650 cm^{-1}) with a resolution of 4 cm^{-1} by reflection of a p-polarised incident beam at an angle of incidence of 45 °. Each spectrum was averaged over 128 scans to improve the signal-to-noise ratio.

Spectral data were processed using Thermo Fisher Scientific OMNIC software. Baseline correction was applied to minimise spectral artifacts. Peak identification was performed by comparing obtained spectra with reference databases and DFT calculated vibrational spectra for thiol and thione tautomers.

4.2.7. Shell-isolated nanoparticle-enhanced Raman spectroscopy (SHINERS)

Raman spectroscopy was performed using a WiTec alpha300R Raman Imaging Microscope. To enhance the Raman signal at the interface, shell-isolated nanoparticle-enhanced Raman spectroscopy (SHINERS) was employed. Gold nanospheres (40 nm in diameter, OD20) in aqueous sodium citrate solution were purchased (AUCR40, NanoComposix) and diluted 20 times prior to use.

The Au shell-isolated nanoparticles (Au-SHINs) were prepared following the method described [219]. Specifically, 0.4 mL of (3-aminopropyl) trimethoxysilane (APTMS) solution (1 mM) was mixed with 30 mL of the as-prepared gold colloid. Subsequently, 3.2 mL of sodium silicate solution (0.54 wt%) with a pH of ~10 was added. The mixture was then transferred to a water bath at 95 °C and stirred for approximately 30 minutes to facilitate the formation of a 2 nm silica shell. The synthesised Au-SHINs were centrifuged twice and washed with ultrapure water. Finally, the concentrated solution was diluted with ultra pure water before application.

The prepared Au-SHINs were drop-cast onto the sample surface and dried on a hot plate at 60 °C. The sample was subsequently exposed to an inhibitor solution for 24 hours. For ex-situ measurements, after exposure the sample was removed, and dried prior to Raman spectroscopy measurements. For in situ inhibitor desorption measurements, samples were immersed in only ultra pure water containing solutions after the first measurement.

For ex situ Raman measurements, a 633 nm wavelength laser was employed with a 50× Zeiss objective (working distance: 9.1 mm) and a laser power of 1 mW to prevent damage.

For in situ measurements, the same 633 nm laser was used with a 63× Zeiss water-dipping objective (working distance: 2.4 mm) and a laser power of 5 mW. Raman spectra were collected every 10 minutes with an integration time of 20 seconds and 10 accumulations per measurement.

4.2.8. Time-of-flight secondary ion mass spectrometry (ToF-SIMS)

The samples were examined using an ION-ToF (GmbH) ToF-SIMS IV equipped with a Bi cluster liquid metal ion source using a BiMn emitter. A pulsed 25 keV Bi^{3+} cluster primary ion beam was used to bombard the sample surface to generate secondary ions. Positive or negative secondary ions were extracted from the sample surface, mass separated and detected via a reflectron-type of time-of-flight analyser, allowing parallel detection of ion fragments having a mass/charge ratio (m/z) up to 900 within each cycle (100 µs). A pulsed, low energy electron flood was used to neutralise sample charging. This technique is extremely surface sensitive, probing only the top 1–3 nm of the sample. The detection limits are believed to be in the range of ppb - ppm, depending upon the ion yield of different elements or species. Note that ToF-SIMS is not a quantitative analytical technique because ion yields for different elements are very different and dependent on the chemical environment in which the elements exist (matrix effect).

At least three areas of 300 µm × 300 µm were measured on each of the samples. The positive secondary ion mass spectra were calibrated using NH^+, C_3H^+ and Cu^+, while the negative spectra were calibrated using CH^-, CN^- and SH^-. The mass resolutions of $C_3H_5{}^+$, $C_4H_9{}^+$, C_2H^- and CSN^- are 5100, 6200, 3400 and 4500, respectively. Fragments were assigned with respect to the theoretical reference values of H (1.0073 amu), C (11.9995 amu), N (14.0025 amu), O (15.9944 amu), Al (26.9815 amu), S (31.9715 amu), Cu (62.9291 amu), ^{65}Cu (64.9272 amu). Normalisation of spectra was performed by dividing the spectra by total counts for any given measurement.

As the secondary ion mass spectra were collected at 128 x 128 pixels over the scanned area, ions can be mapped by plotting their intensities against each pixel. The ion images are represented by a false colour scale, where a brighter colour corresponds to a stronger ion intensity.

4.2.9. Speciation calculations

Speciation calculations and prediction of pKa values were performed through the Chemicalize Instant Cheminformatics Solution software package of ChemAxon [220, 221]. The Chemaxon pKa calculator employs a computational methodology based on the analysis of partial charge distributions across molecular structures to predict ionisation constants. The algorithm calculates the partial charge of atoms, which are sensitive to protonation and deprotonation events, to determine the acidic and basic dissociation constants (pKa values) of ionisable functional groups. For multiprotic compounds, the tool distinguishes between micro and macro dissociation constants: micro constants derive from equilibrium concentrations of conjugated acid-base pairs, while macro constants are calculated using global mass and charge conservation principles, enabling prediction of complex ionisation equilibria [222].

4.2.10. DENSITY FUNCTIONAL THEORY (DFT) CALCULATIONS

ALL quantum chemical calculations were performed using the ORCA 6.0 software package [223, 224]. The electronic structures of the molecules were optimised using the hybrid B3LYP functional, which combines the Becke three-parameter exchange and the Lee-Yang-Parr correlation [225–228]. To account for dispersion interactions critical in non-covalent and adsorption phenomena, the Grimme's D3(BJ) empirical dispersion correction with Becke-Johnson damping was included [229]. Geometry optimisations were carried out with the def2-TZVPD basis set[230], a triple-ζ valence polarised basis with diffuse functions. Convergence criteria for the self-consistent field (SCF) procedure were tightened to ensure stringent convergence thresholds for geometry optimisation (max SCF energy change $\Delta E < 10^{-8}$ Eh).The RIJCOSX approximation (resolution of identity for Coulomb integrals (RI-J) and chain-of-spheres exchange (COSX)) was employed to accelerate computations without significant loss of accuracy [231].

Solvent effects were incorporated using the conductor-like polarisable continuum model (CPCM) [232] with the solvation model based on density (SMD) parameterisation [233] to simulate aqueous environments. The SMD model employs a universal solvation approach based on solute electron density and solvent-specific parameters (dielectric constant, surface tension, etc.), to cost-effectively predict solvation free energies in water. The solvent was defined as water ($\varepsilon = 78.36$).

Harmonic vibrational frequency calculations were performed to confirm that the optimised geometries correspond to true minima (no imaginary frequencies) and to compute thermal corrections to the Gibbs free energy. Base electronic and thermal contributions (enthalpy, entropy) were extracted from the frequency output to calculate temperature-dependent thermodynamic Gibbs free energy values at 298.15 K.

Simulated vibrational spectra were generated by representing each computed vibrational mode as a Gaussian peak centered at its corresponding frequency. The intensity of each peak was determined by the computed vibrational intensity, while the broadening was controlled by a fixed width parameter. The overall spectrum was constructed by summing these individual peaks over the relevant frequency range, producing a smooth vibrational profile. An empirical wavenumber scaling factor of 0.99 was applied to correct for systematic over/underestimations inherent to the chosen functional and basis set.

Calculations were performed on a desktop computer with an AMD Ryzen 7 7800X3D processor, which were parallelised over 8 cores to increase computational efficiency.

Dipole moment magnitude, E_{HOMO} and E_{LUMO} values were extracted from the simulations, and were used to calculate properties of the HOMO-LUMO bandgap, electronegativity, chemical hardness, and electrophilicity. The HOMO-LUMO gap ΔE is calculated as:

$$\Delta E = E_{LUMO} - E_{HOMO} \tag{4.2}$$

and is directly related to the reactivity of the molecule, where a smaller gap enhances electron transfer. The electronegativity χ of a molecule can be approximated by:

$$\chi = -\frac{E_{HOMO} + E_{LUMO}}{2} \tag{4.3}$$

which is related to the charge transfer tendency of molecules. Chemical hardness η is calculated as:

$$\eta = \frac{E_{HOMO} - E_{LUMO}}{2} \tag{4.4}$$

which measures resistance to electron cloud deformation. The electrophilicity index ω is derived as:

$$\omega = \frac{\chi^2}{\eta} \tag{4.5}$$

which is a measure of how susceptible a molecule is to electrophilic attack [234].

To evaluate the site-specific reactivity of the molecule, the electronic structure was further analyzed by computing the atomic Mulliken charges and by performing Fukui analysis. Atomic Mulliken charges are calculated by partitioning the electron density among atoms based on molecular orbital coefficients, which provides insight into the charge distribution and reactivity of a molecule. Fukui analysis allows prediction of the most electrophilic and nucleophilic sites of a molecule by quantifying the changes in electron density at specific positions in a molecule during a chemical reaction involving electron transfer, which is calculated as:

$$f(r) = \frac{\partial \rho(r)}{\partial N_{electron}} \tag{4.6}$$

where $\partial\rho(r)$ is the electron density. By adding or removing an electron from an optimised DFT calculation, and taking the difference between anion-neutral and neutral-cation electron density distributions, finite-difference approximations of the electron density response to changes in electron population can be obtained. Fukui functions are these finite-difference approximations to changes in electron densities, which can inform about the sites susceptible to an electrophilic or nucleophilic attack, which can be calculated using:

$$f_+(r) = \rho_{N+1}(r) - \rho_N(r) \tag{4.7}$$

$$f_-(r) = \rho_N(r) - \rho_{N-1}(r) \tag{4.8}$$

where $f_+(r)$ is the Fukui function for the addition of an electron to a molecule, and $f_-(r)$ the Fukui function for the removal of one electron from the molecule.

Chemcraft software was used to visualise Fukui functions, HOMO and LUMO to identify the most reactive sites for electrophilic and nucleophilic interactions.

4.3. Results and Discussion

4.3.1. Electrochemical response to molecule presence and subsequent absence

Figure 4.1 presents the potentiodynamic polarisation curves of AA2024-T3 in inhibited and uninhibited conditions. In the uninhibited case, where the samples were

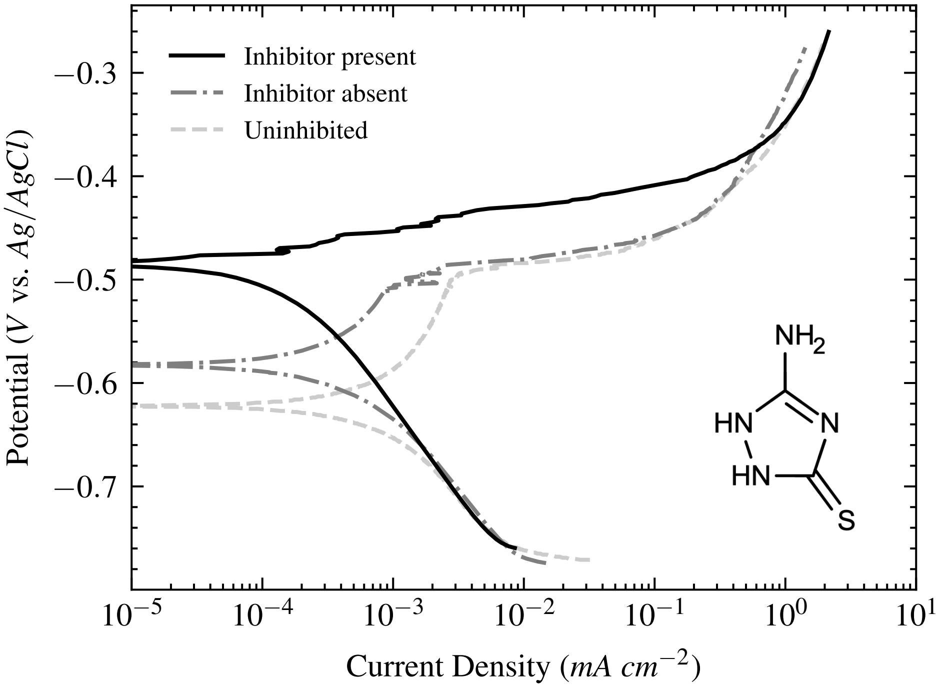


Figure 4.1: Potentiodynamic polarisation curves of AA2024-T3 exposed to a 0.1 M NaCl electrolyte with 1 mM 3-amino-1,2,4-triazole-5-thiol molecule for one day (inhibitor present), followed by three days in 0.1 M NaCl without the molecule (inhibitor absent), and sample exposed to 0.1 M NaCl alone for four days (uninhibited). Thione tautomer of the molecule shown as an inset. Potentials measured with respect to Ag–AgCl (saturated KCl) references.

exposed to only 0.1 M NaCl for four days, the resulting corrosion potential values were –668 ± 40 mV vs. Ag–AgCl (saturated KCl), and corrosion current densities were 36.82 ± 3.60 × 10^{-5} mA cm^{-2}. With the addition of 1 mM 3-amino-1,2,4-triazole-5-thiol molecule and exposure to this inhibitor present corrosive environment for one day, the corrosion potential values resulted in values of –485 ± 8 mV, and corrosion current densities were 3.82 ± 1.99 × 10^{-5} mA cm^{-2}, which corresponded to inhibition efficiencies 91.50 ± 4.42 %. With the subsequent exposure of three days in the absence of molecule and only 0.1 M NaCl, corrosion potential values were –613 ± 35 mV, and corrosion current densities were 25.94 ± 3.08 × 10^{-5} mA cm^{-2}, which corresponded to inhibition efficiencies 42.30 ± 6.85 %.

The initial exposure to the molecule caused the mean corrosion potential values to shift to 183 mV more positive potentials, and subsequent exposure to a molecule absent environment decreased this positive potential shift. However, at the end of the exposure the corrosion potential of the inhibitor absent samples were still 128 mV more positive than the completely uninhibited case. In a similar manner, corrosion current densities also dropped an order of magnitude in the presence of the molecule, which climbed back up to the uninhibited values for inhibitor absent case, but not completely: from 91 % to

42 % inhibition efficiency. This *is* the sustained quasi-inhibition behaviour offered by the molecule 3-amino-1,2,4-triazole-5-thiol: whereas other molecules lose their gained corrosion inhibition completely if the molecule is not sustained in the environment, this particular molecule somehow still sustains corrosion inhibition, albeit with a reduced efficacy.

Based on mixed potential theory, if the cathodic half reactions remain the same while the corrosion potential shifts to more positive values and corrosion current densities drop to lower values, the cause of this shift must be the inhibition of anodic half-reactions. In light of the cathodic parts of the plots, which seem to be completely unaffected by the molecule presence and overlap in all conditions, and the pitting potentials which overlap in the inhibitor absent and uninhibited cases, the primary mechanism that causes quasi-stable corrosion inhibition has to be through the suppression of anodic reactions.

4.3.2. Surface topography and potential distributions

To determine whether this sustained inhibition occurs across the entire sample surface or is primarily due to molecule interaction with the intermetallics, the self-assembly of the molecules were analyzed through the topography and surface potential investigations. The AFM coupled with SKPFM was utilised to examine the influence of the adsorbing layer of inhibitor molecules on the nanoscale surface morphology and electrical potential distribution of the aluminium alloy matrix and its intermetallic particles. Special emphasis was placed on variations in the electrical surface potential of the layer, particularly after exposure to an inhibitor-free electrolyte. It is important to note that the physicochemical interactions between these inhibitor molecules and intermetallic particles - whether relative to the matrix anodic (leading to their own dissolution) or cathodic (accelerating matrix dissolution through micro-galvanic interactions) [42, 43] - play a crucial role in controlling or inhibiting corrosion processes.

The presence of a thin overlayer of organic and inorganic materials on the sample surface can alter the work function (WF) due to electron transfer and structural relaxation at the interface [235]. Similarly, in doped semiconductors, band bending in the subsurface depletion layer can induce comparable changes [236]. It is important to note that, in SKPFM analysis, the electrical forces between the AFM tip and the substrate can be categorised into two main components: capacitance forces, which arise from surface potential and dielectric screening, and Coulombic forces, which result from static charges and multiples [236]. For self-assembled monolayers (SAMs) adsorbed on metallic surfaces, a new energy level arrangement forms at the SAM/oxide film interface (here, aluminium native oxide film). SAM adsorption on an aluminium oxide film affects electrostatic interactions and capacitance by modifying the local WF or contact potential difference (ΔCPD) between the AFM tip-apex and the SAM-covered Al oxide layer. These changes arise from band bending (Δ_{bb}), the perpendicular SAM dipole moment (μ_{SAM}), and interfacial bonding (Δ_{bond}), leading to a new local surface potential ($\mathrm{SP}_{\mathrm{SAM}}$) on the SAM-covered aluminium oxide [237]:

$$\mathrm{SP}_{\mathrm{SAM}} = \mathrm{SP}_{\mathrm{Al\,oxide}} + \mu_{\mathrm{SAM}}/\mathrm{e} + \Delta_{\mathrm{bond}} \tag{4.9}$$

where e is the elementary charge.

Figure 4.2 presents the results of the surface examination of topography and potentials. The left side of each map represents surfaces initially exposed to the inhibitor molecules, while the right side shows the bare surfaces used as control. The top row (Figures 4.2 (a) – 4.2 (d) depicts surface segments exposed to the molecule-containing electrolyte. In contrast, the bottom row (Figures 4.2 (a') –4.2 (d')) illustrates surface segments initially exposed to the inhibitor molecule-containing electrolyte, followed by exposure to an inhibitor-free electrolyte. Topography maps in Figures 4.2 (a) and 4.2 (a') clearly reveal a morphological transformation of the adsorbed inhibitor nanolayer, shifting from a finely agglomerated structure to a larger domain-agglomerated form after exposure to the inhibitor-free electrolyte. Moreover, in both conditions, the topography maps indicate a sustained surface coverage by the organic molecules.

As observed in the AFM line scans in Figures 4.2 (c) and 4.2 (c'), the topographical values decline relative to the bare surface was approximately 50 nm for the inhibitor-exposed surface and around 10–20 nm for the subsequent inhibitor-absent surface. This suggests thinning of the previously formed film due to the desorption of weak bonds and other less stable adsorption configurations in the absence of the inhibitor molecules. The SKPFM map in Figure 4.2 (b) and line scans in Figure 4.2 (c) indicate that the presence of the inhibitor molecules increased the electrical surface potential and charge by approximately 60 mV compared to the bare surface, resulting in a significantly higher value. However, after the removal of the inhibitor molecules (Figures 4.2 (b') and 4.2 (c')), the surface potential and charge dropped to values lower than those of the bare surface, with a significantly larger potential difference of approximately 140 mV.

Considering the complementary spectroscopy findings discussed in the following sections, this phenomenon can be attributed to the orientation and dipole moment of the stabilised molecules on the surface. The reduction in organic layer thickness indicates partial desorption of the adsorbed molecules when they are no longer sustained in the environment. This implies that certain bonding interactions or adsorption configurations exhibit greater stability than others, which may correlate with the quasi-stable corrosion inhibition behaviour. According to the previous studies, the self-assembled monolayers of molecules oriented with the positive pole upwards decrease the work function of the surface, whereas monolayers oriented with the negative pole upward increase it [238, 239]. Assuming a single dominant stable bonding configuration for the surface, a decrease in surface potential and charge should result from the molecule in thione form with the sulfur functional group adjacent to the surface, or in thiol form with amino group adjacent to the surface (see section 4.3.6 - the strong dipole moment of thione bonded to surface through sulfur supports this).

The opposite trends seen in molecule-present case could be due to the other additional, less stable bonding configurations of the molecule, which bonds to the surface with dipole moments in opposite direction (e.g. through electron donation via amino groups in thione form). The formation of a thicker electrically insulating multilayer inhibits electron transfer, acting as a barrier and increasing the measured surface potential. Additionally, the presence of the inhibitor layer can raise the work function by altering the local electrostatic interactions, depending on the molecular orientation and dipole interactions [236]. This behaviour is consistent with the formation of a multilayer, where the initial monolayer induces a dipole via Pauli repulsion, compressing the

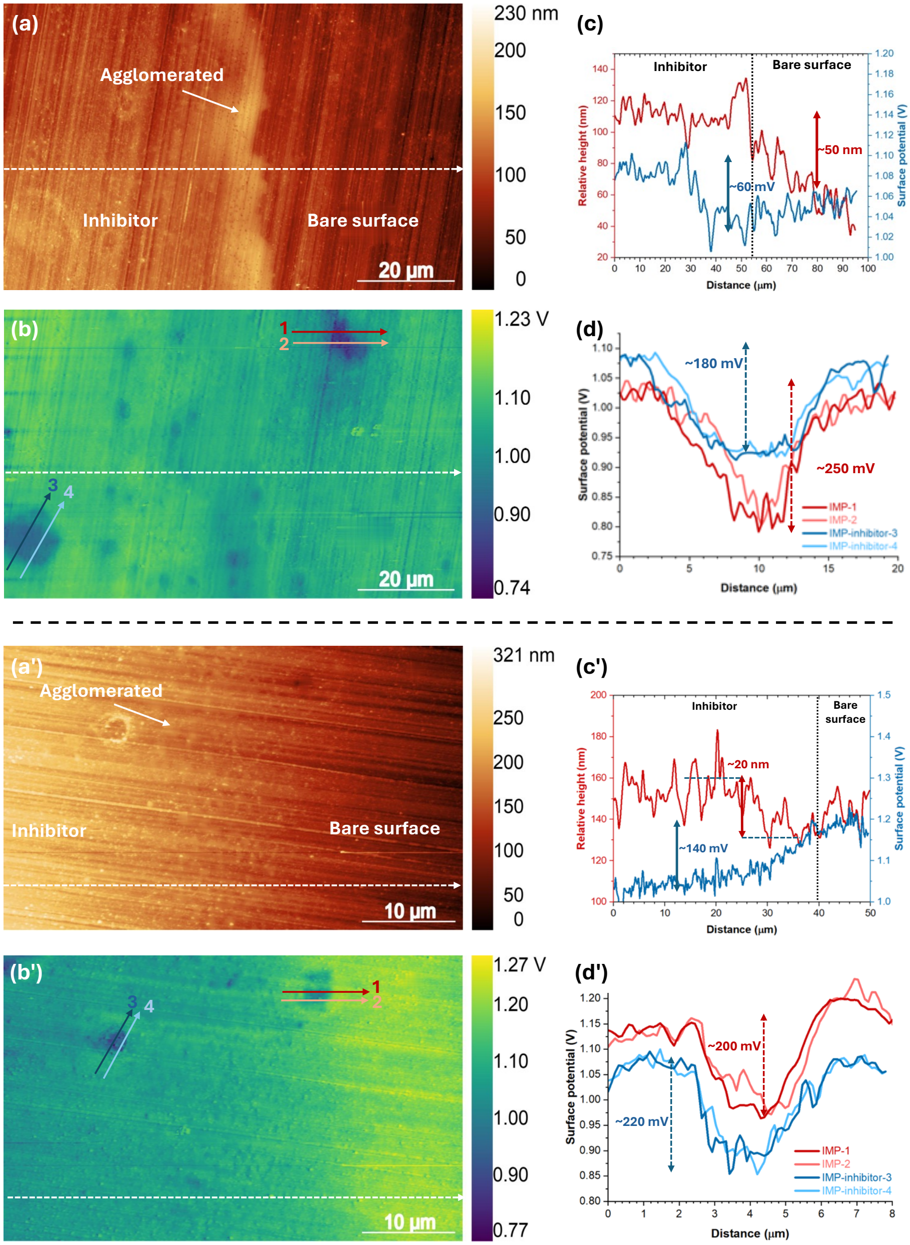


Figure 4.2: (a) AFM topography and corresponding (b) surface potential maps, and associated line profiles for (c) matrix and (d) intermetallics. Subfigures above the dashed line denote the molecule-present case, and below the subsequent molecule-absent case. “Inhibitor" represented the regions initially exposed to the inhibitors, as opposed to the “bare surface" side that experienced no electrolyte exposure.

surface electron cloud and lowering the vacuum level. Additional layers contribute minimally due to their lack of direct interaction with the substrate and partial cancellation of molecular dipoles. These results align with previously reported studies of metal–organic interfaces, where the electronic structure and dipole formation are governed by the nature of bonding at the interface rather than bulk molecular properties alone [240].

The surface potential line profiles of the intermetallic particles are shown in Figures 4.2 (d) and 4.2 (d'). The particles were determined based on their shapes and surface potentials, which were clearly different than those for the matrix [239, 241]. Given the resolution of the scans, larger intermetallics were chosen for the line scans for a more accurate interpretation. The trends between smaller and larger intermetallics were similar, as seen by the same colour contrast (z-scale bar) with the matrix.

The predominant intermetallic phases on the AA2024-T3 alloy surface are Cu-rich ϑ-phase and S-phase, which play a crucial role in localised dissolution and pitting corrosion in the presence of chloride ions. This corrosion susceptibility arises from the intrinsic dealloying behaviour of intermetallic particles [42, 55], and the surface potential differences between the intermetallic particles and the surrounding aluminium matrix [43, 242, 243]. The presence of inhibitor molecules reduced the surface potential difference between the matrix and intermetallic particles from approximately 250 mV to 180 mV, thereby mitigating the driving force for localised corrosion. However, upon the subsequent absence of the inhibitor, this effect is lost, as the surface potential of both the bare alloy and the inhibitor-covered surface converge to similar values (approximately 200 mV and 220 mV, respectively), increasing the driving force for galvanic coupling in comparison to the inhibited molecule-present case. This suggests that corrosion protection of the intermetallics is not sustained in the absence of the inhibitor. This finding is further supported by the pitting potentials shown in Figure 4.1, where, as compared to reference, the inhibitor presence shifts the potential around 20 mV positive values, but the subsequent absence of the inhibitor deprives the sample of this effect.

4.3.3. Persisting chemical signatures after molecule withdrawal

To understand the chemical states responsible for the stable bonding configurations, AA2024-T3 surfaces were observed with XPS in the presence and subsequent absence of the inhibitor molecule. Figure 4.3 presents the N1s (4.3 (a) -4.3 (a')) and S2p (4.3 (b) - 4.3 (b')) high resolution XPS spectra. The rest of the high resolution and survey spectra can be found in supplementary information. The binding energy peaks for the N and S atoms in different chemical environments of adsorbed molecules corresponding to the red, blue and green fits are collected in Table 4.1. No peaks for the values presented herein were observed for the control measurements with pristine samples.

The N1s spectrum was best fitted with three components in the presence of the inhibitor molecule, and two components in the subsequent absence of the molecule, as visible in Figure 4.3 (a) and 4.3 (a'), respectively. The peak at 399.4 eV was attributed to the amino functional group, and 400.6 eV the triazole, which was based on the previous spectra obtained for 3-amino-1,2,4-triazole and 1,2,4-triazole [244]. The peak at 401.6 eV was assigned for protonated N. These assigned values were in line with the reference spectra and other studies in which amino groups and aromatic azoles were studied [245–

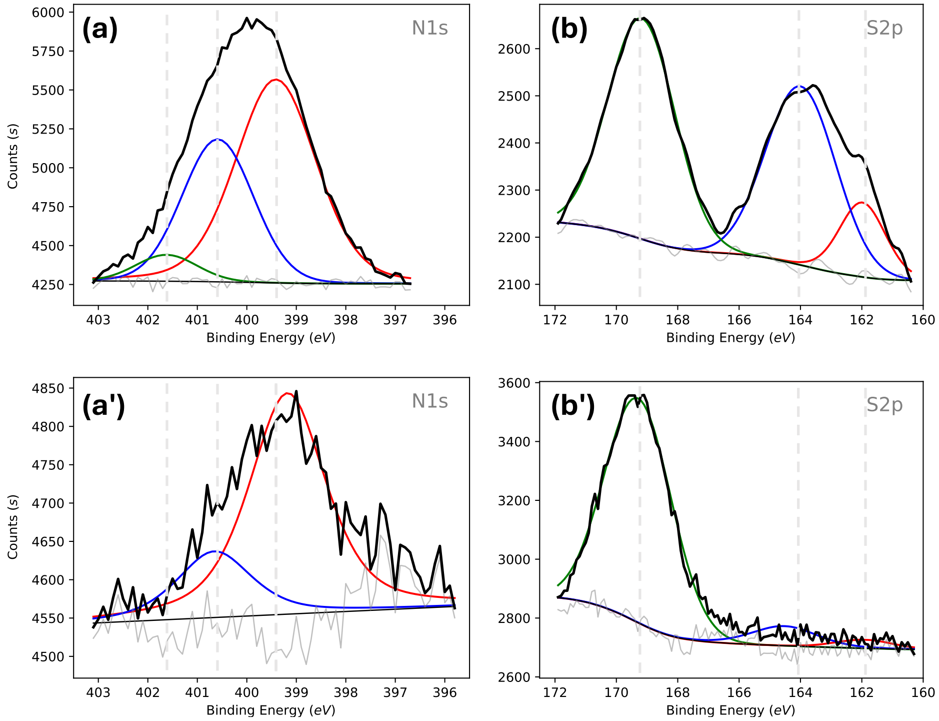


Figure 4.3: High resolution XPS N1s and S2p spectra for (a) - (b) inhibitor presence, (a') - (b') subsequent inhibitor absence.

	Binding Energy (eV)					
Molecule	-NH_2	N-ring	N-protonated	$\cdots$S-C	$\cdots$S=C	O=S=
Presence	399.4	400.6	401.6	162.0	164.0	169.2
Absence	399.2	400.6				169.3

Table 4.1: Binding energy peaks of N1s and S2p for 3-amino-1,2,4-triazole-5-thiol in the initial presence and subsequent absence from the environment.

248]. The signals that would result from nitride- and nitrite/nitrate-like bonds to metal and oxides were missing from either spectra, which would appear below 397.5 eV above 404 eV [248, 249]. The stability of the 400.6 eV peak suggests the triazole ring remains intact during adsorption/desorption. The 0.2 eV decrease in the amino group binding energy could be related to an increase in the electron density around the nitrogen, but is hard to say for certain due to low signal-to-noise ratio and multiple possible peaks shifting at the same time. If that is the case, this might indicate that amino nitrogen, which was previously weakly bonded, has desorbed and returned to its normal, less electron-deficient state.

The S2p spectrum was best fitted with three components in the presence of the inhibitor molecule, out of which only one remained in the subsequent absence of the molecule, as visible in Figures 4.3 (b) and 4.3 (b'), respectively. The S2p spectrum was fitted as a single peak model neglecting spin-orbit coupling effects, due to peak broadening from complex interactions between the alloy surface and the adsorbing/desorbing organic molecule, making accurate spin-orbit resolution challenging. The peak at 162.0 eV was attributed to the thiol, 164.0 eV the thione form of the molecule, and 169.2 eV to the oxidised sulphate-like structure. The assignment was based on the reference work [248], and prior study on the adsorption of 3-amino-1,2,4-triazole-5-thiol molecules on Ag and Au surfaces [250].

The thiol and thione signals demonstrate that both tautomers interact with the surface when the molecule is present in the environment. Thione seems to be the dominant tautomer, (this is further validated in section 4.3.6). There appears to be a transient interaction, likely through physisorption of the molecule to the surface through protonated nitrogen, and the sulfur atoms of thiol and thione. However, when the inhibitor is no longer present in the environment, all these peaks disappear. Instead, the sulfate-like peak becomes more intense in the absence of the inhibitor. This is likely due to the removal of excess weakly-bonded molecules, which would otherwise shield the interface signal more, which is also in line with decreasing carbon and nitrogen signals in the subsequent absence of the molecule.

This implies sulfate signal coming from the interface. The persistence of nitrogen signals, and the increased sulfate-like peak suggest surface functionalisation via sulfur. This sort of functionalisation of aluminium (hydr)oxide surfaces by thiols have been observed before [251]. In this stable state, triazole nitrogen does not protonate or deprotonate, as indicated by unchanged nitrogen binding energies. This implies that thione complexation with aluminium (hydr)oxide is key to quasi-stable corrosion inhibition.

A similar behaviour is also observed from the ATR-FTIR spectra of the molecule-surface interactions. Figure 4.4 presents the ATR-FTIR spectra of the AA2024-T3 surface (a) exposed to the inhibitor molecule, (b) spectra when the same surface is shortly rinsed-off with ultra-pure water, (c) following exposure to inhibitor absent environment, and simulated vibrational spectra of (d) thione and (e) thiol tautomers through DFT computations. The vibrational modes that might be relevant for analyzing the experimental results are summarised in Table 4.2.

A comparison of Figures 4.4 (a) and 4.4 (d), and calculated peaks closest to the experimentally measured spectra shows that simulated thione spectrum better overlaps with the experimental spectrum, once again showing that majority of the molecules are present in the environment in their thione tautomers. The -SH peak was missing (nothing at 2640 cm^{-1}), indicating either thiol form of molecule is in trace amount or all thiol tautomers are found in deprotonated mercapto forms. However it seems that there indeed might be lesser contributions coming from the thiol vibrations when molecule is present in the environment, as seen in Figure 4.4 (a) and Figure 4.4 (e) partial peak overlap around 1600 cm^{-1}.

The simulations reveal that experimentally observed peaks around 1645, 1484, 1346 and 1278 cm^{-1} were related to the various stretching and rocking vibrations of amino and triazole protons. All these vibrations disappeared with absence of molecule in the

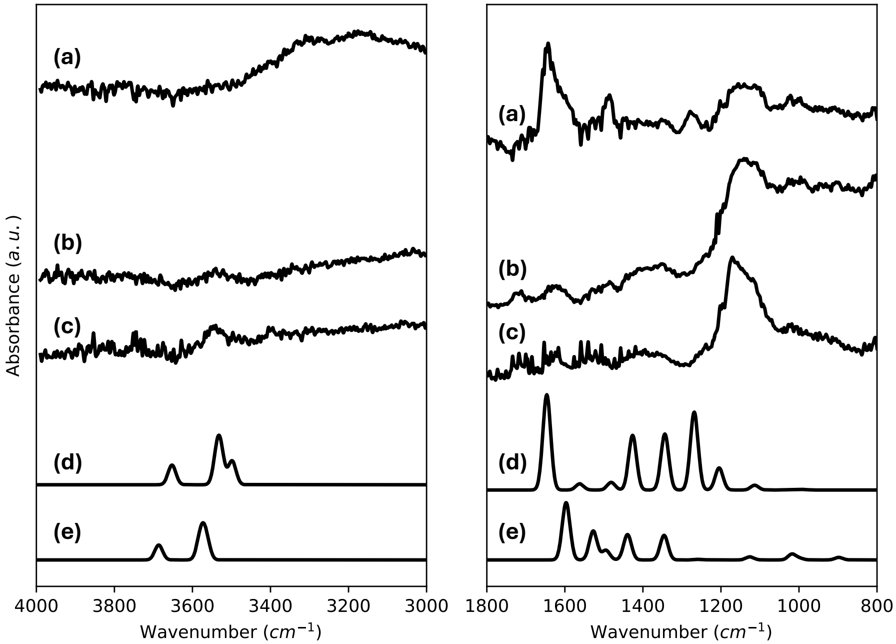


Figure 4.4: ATR-FTIR spectra related to the molecule adsorption. ATR-FTIR experiments of (a) inhibitor presence, (b) rinsed-off, (c) subsequent absence, and vibrational spectra of the molecule calculated with DFT in (d) thione, and (e) thiol tautomers.

environment. Weak but consistent peaks at 1200 and 1115 cm^{-1} were present in all conditions, which corresponded to the vibrations for C-NH stretch for NH near sulfur, and NH-NH stretch, respectively. These peaks may suggest a stable bonding configuration via the thione tautomer, where the NH group nearest to the sulfur atom does not deprotonate. A new broad weak peak appeared at 3520-3560 cm^{-1}, with possible contributions from -NH_2 or proton stretch for NH near S. If indeed it is originating from the amino group (which is likely, as the signal for proton group from NH near S should have been present for all conditions, but it was not), it might be related to a bonding configuration where a hydrogen bonding between -NH_2 and the surface is released. This would also agree with the 0.2 eV binding energy decrease observed for N2s -NH_2 peak. A very weak peak with increased presence for subsequent molecule absence was observed for 1071 cm^{-1}, calculated to be related to C=N stretch. This might be related to a type of original π-bonding state recovered from a transient surface interactions.

The main prominent consistent peak observed between 1170-1090 cm^{-1} was present through all conditions, and was not observed throughout the vibration calculations. The peak displays quite an interesting behaviour, becoming only more prominent and sharper in the absence of the molecule in the environment. The values for this peak

Table 4.2: Relevant calculated vibrational frequencies for thiol and thione tautomers of 3-amino-1,2,4-triazole-5-thiol with their main contributions, and experimental results. Peaks close to the experimental results (<10 cm^{-1} difference) indicated in bold. Peaks ↑ : increase, ∘ : constant, ↓ : decrease when environment is changed from inhibitor-containing to inhibitor-absent environment.

Thiol		**Thione**		**Experiment**
Wavenumber (cm^{-1})	**Main Contribution**	**Wavenumber (cm^{-1})**	**Main Contribution**	**Wavenumber (cm^{-1})**
3686	Asymmetric -NH_2 stretch	3652	Asymmetric -NH_2 stretch	
3578	Proton stretch (NH near S)	**3536**	Symmetric -NH_2 stretch	↑ 3520 – 3560
3565	Symmetric -NH_2 stretch	**3528**	Proton stretch (NH near S)	
1597	C-NH_2 stretch	**1646**	C-NH_2 stretch	↓ 1650 – 1590
1495	Proton rocking (N near S)	**1481**	Proton rocking (NH near S)	↓ 1484
1346	C=N stretch (N near S)	**1344**	Asymmetric proton rocking (triazole)	↓ 1346
1255	Proton rocking (NH near S)	**1268**	C-N stretch (N near S)	↓ 1278
		1205	C-NH stretch (NH near S)	∘ 1200
				↑ 1170 – 1090
1126	N-NH stretch	**1113**	NH-NH stretch	∘ 1115
		1077	C=N stretch	↑ 1071

corresponded to the the S=O vibrations observed for asymmetric and symmetric sulfate stretch [252–256]. The fact that the measurement background was the original sample prior to molecule treatment, suggests that this is a newly formed sulfate-like layer on the surface. Supported by the previous XPS analysis, the fact that this peak is absent in the simulated spectra based only on the molecule, but strongly present in experimental spectra calls attention to the role of sulfatisation on the quasi-stable corrosion inhibition behaviour.

4.3.4. TEMPORAL EVOLUTION OF SURFACE-BOUND SPECIES IN THE SUBSEQUENT ABSENCE OF INHIBITOR MOLECULES

To elucidate the evolution of stable bonding mechanisms through time and untangle the effect of intermetallic and the Al matrix, in-situ molecule desorption experiments were performed for pure aluminium, copper, and AA2024-T3 surfaces. After exposing surfaces to inhibitor molecule for one day, their desorption is followed in-situ with SHINERS spectroscopy. Additionally, ex-situ Raman spectra were collected for molecule in powder form, molecule in aqueous solution, AA2024-T3 alloy surface without nanoparticles, alloy surface only exposed to water, and alloy, pure aluminium, pure copper surfaces exposed to inhibitor-containing solutions. The collected information was used for analyzing in-situ spectra, and can be found in supplementary information.

Figure 4.5 presents the results of the in-situ Raman experiments. Figures 4.5 (a) - 4.5 (b) correspond to aluminium, 4.5 (c) - 4.5 (d) correspond to copper, 4.5 (e) - 4.5 (f) correspond to alloy spectra and heatmaps, respectively. Heatmaps present the square root of the intensities to help with the identification of weaker signal trends. Upon initial observation it is clear that pure Al and AA2024-T3 heatmaps are more similar to one another.

The weak peaks present at lower wavenumber 200-300 cm^{-1} likely result from a mixed contribution of the signals from the substrate (measured peaks around 290 cm^{-1}), aqueous molecule (measured peaks around 273 and 360 cm^{-1}), and metal sulphide bonds [257], and disappear in the first 30 minutes. The shoulders at 430-450 and 650 cm^{-1}, and peaks at 700-800 cm^{-1}) range correspond to the signals associated with

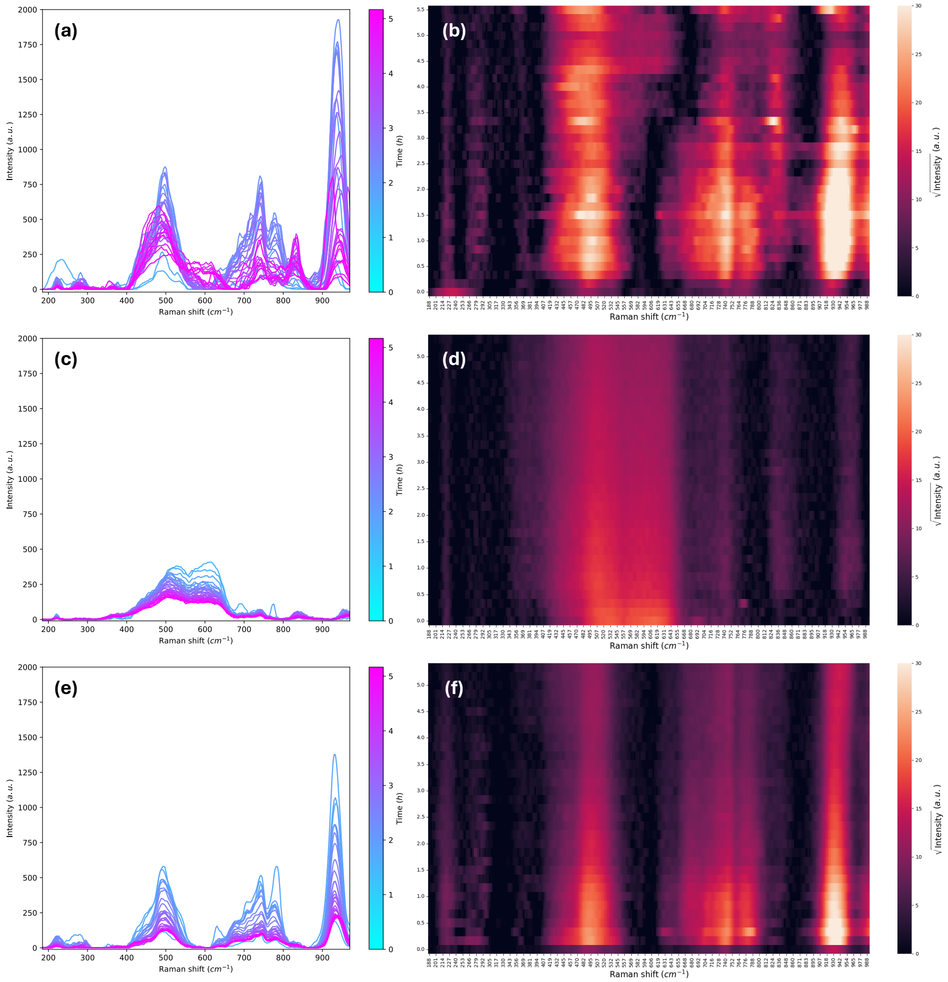


Figure 4.5: SHINERS Raman spectra and heatmaps of in-situ molecule desorption phenomena from (a)-(b) Al, (c)-(d) Cu, (e)-(f) AA2024-T3 surfaces.

the 3-amino-1,2,4-triazole fragment stretch, bend and torsion modes [258]. These signals were more much present with pure Al and alloy substrates and their signals decreased with time. Copper nitride-like peaks are potentially contributing to the signals observed for 610-640 cm^{-1} [259, 260], but it is difficult to isolate their effect from the signals originating from Cu_2O and $Cu(OH)_2$ as their previously observed peaks correspond to 523 and 623 cm^{-1}, and 490 cm^{-1}, respectively [261]. However, in the case of a copper oxide growth, corresponding peak signals should increase or stay the same. This, in addition to the faster signal decrease at higher wavenumbers of ~630 cm^{-1} makes us believe that they are related to the bonding between N of the molecule and Cu surface, which depreciate through time. For the molecule in aqueous solution, a sharp strong peak at 500 cm^{-1} was observed, which was previously attributed to C=S vibrations [250]. This peak shifted to lower values in the 480-500 range for Al and alloy, and to higher values of 510-520 cm^{-1}) for Cu - but again it is difficult to disentangle this from the overlapping potential signals from Cu_2O.

In the ex-situ spectra of the molecule dissolved in aqueous solution, no peaks are present between 550-740 and 800-960 cm^{-1} range. In the ex-situ spectra of the alloy, no peaks are present between 320-910 cm^{-1} range. However, peaks appear for molecule exposed surfaces in these ranges. Aluminium nitride-like peaks were observed between 610-660 cm^{-1} for Al and alloy surfaces [262], which appeared at later exposure times for Al and was relatively constant, but disappeared almost immediately from the alloy surfaces. The peak around 840 cm^{-1} that developed with time for Al and weak but almost constant for Cu, was previously attributed to C-H out-of-plane bending [263] and a vibration at 829 cm^{-1} was previously calculated to be related to 3-amino-1,2,4-triazole molecule coupled out-of-plane rocking of the amino group with triazole ring torsion [258]. The strongest peak for Al and alloy however was the one observed at 936 cm^{-1}, which was attributed to the symmetric sulfate stretch [264–266].

For the Cu and alloy substrates, all observed peaks exhibited a periodic decrease in intensity over time. Despite this decline, in the case of the alloy a significant signal remained for the molecule related peaks at 492 and 744 cm^{-1}, as well as the sulfate peak at 940 cm^{-1}. This persistence suggests that the formation of a sulfate-functionalised surface is associated with the quasi-stable corrosion inhibition behaviour. This behaviour was evident for Al and the alloy, but was not present for Cu.

On the Cu surface, peaks corresponding to Cu-nitride Raman shift values were observed. While nitride-related features also emerged on the Al surface at later stages, they were absent for the alloy. In fact the peak at 630 cm^{-1}, which is potentially linked to metal-nitride formation, disappeared within the first 20 minutes for the alloy substrate.

These findings highlight significant differences in adsorption mechanisms, and in comparison to the transient nature of sulfide- and nitride-like peaks, the sulfate bonding configuration developed on aluminium oxide appears to provide a more robust and stable molecule-substrate interaction on the alloy surface.

4.3.5. DETECTION OF PERSISTENT MOLECULAR FRAGMENTS ON THE SURFACE

ToF-SIMS was used to analyze the strongly-bound surface molecular fragments. The positive and negative ion spectra of 3-amino-1,2,4-triazole-5-thiol, with its formula M = $C_2H_4N_4S$, was collected after drying from its aqueous solution on an aluminium weighing boat, which can be found in supplementary information. The major positive ions included the hydrogenated molecular ion $[M+H]^+$, and fragment ions $CH_3N_2^+$, $CH_6N_3^+$ and $C_2H_5N_4^+$. The major negative ions included molecular ion M^-, dehydrogenated molecular ion $[M\text{-}H]^-$ and fragment ions CN^-, S^-, CHN_2^-, CSN^- and $C_2N_3^-$.

The inhibitor-treated substrates of AA2024-T3, Cu, and Al were rinsed briefly (~1 min) and extensively (~1 h). No significant differences in ToF-SIMS spectra were observed between rinsing conditions, suggesting that the inhibitor forms a robust layer on the substrates. Therefore, only the inhibitor-treated samples that received an extensive rinse are elaborated further to focus on rather strongly adsorbed species. Figure 4.6

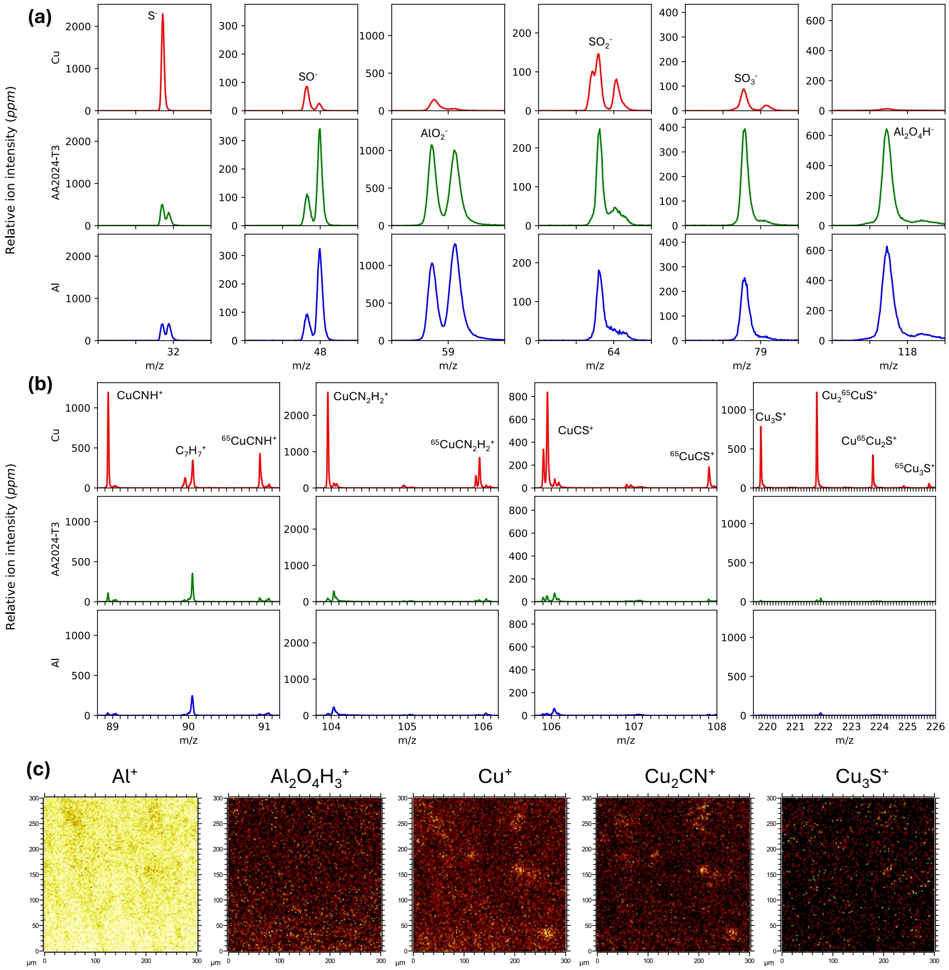


Figure 4.6: Selected (a) negative and (b) positive ion spectra for Cu, AA2024-T3, Al samples; and (c) ion maps of the AA2024-T3 sample. One tick distance on x-axis denotes 0.1 m/z.

shows some selected peaks relevant for analysis. For the complete spectra of the samples refer to the supplementary information.

While no aluminium was detected on the treated Cu sample, Cu was detected for all three samples, with a greatly reduced Cu^+ intensity detected on AA2024-T3 sample and an even weaker Cu^+ intensity on sample Al. No Al-containing fragments (e.g. Al-S, Al-N or Al-CN) were observed for AA2024-T3 and Al samples. As seen in Figure 4.6 (a) SO^-, SO_2^-, and SO_3^- peaks were present for all samples, with a stronger SO_3^- signal observed for AA2024-T3. However SO_2^- peaks were also present for the control pristine surfaces (without molecule application) as well, which makes a conclusive analysis about sulfur-oxygen not possible using ToF-SIMS. Thus, the following text focused on the Cu-related phenomena.

Copper-inhibitor complexes $CuC_2H_4N_4S^+$ and $^{65}CuC_2H_4N_4S^+$, are detected on the Cu sample, confirming direct copper-inhibitor bonding. The complex, $CuC_2H_4N_4S^+$, is also detected on AA2024-T3 sample but with a greatly reduced abundance. There were even weaker $CuC_2H_4N_4S^+$ signals detected on the Al reference sample (most likely originating from Cu impurities). The reduced complex signal in AA2024-T3 and Al correlated with Cu abundance in these samples.

On the other hand, no $^{65}CuC_2H_4N_4S^+$ (180.946 amu) could be confirmed for samples AA2024-T3 and Al since there is another peak at the very similar m/z value (180.941 amu). It was confirmed that there is a peak at this m/z value, which may be assigned to $Al_4F_2O_2H_3{}^+$ by mass matching (it is worth mentioning that decent F^- signals were detected on the AA2024-T3 and Al samples). This peak thus covers the rather weaker $^{65}CuC_2H_4N_4S^+$ signals of samples AA2024-T3 and Al. By contrast, there was no such interfering peak at this m/z value for the pristine Cu sample.

This observation that the copper-inhibitor complex ion is too weak for AA2024-T3 and Al samples, and the fact that the detection of $CuC_2H_4N_4S^+$ does not clarify whether copper interacts with the amine, the thione, or both, calls for examining the copper-containing fragment ions. It was confirmed that there are various, rather abundant copper-containing ions indicating interactions between copper and the inhibitor.

Shown in Figure 4.6 (b) are the positive spectra for $CuCNH^+$, $CuCN_2H_2{}^+$, $CuCS^+$ and Cu_3S^+. The first two ions and the last two ions are interpreted to represent the interaction of copper with the inhibitor at the amine site and the thione site, respectively. Other copper-containing positive ions include Cu_2S^+, Cu_2SH^+, Cu_2CN^+ and Cu_2CNS^+. These ions further corroborate affinity of copper for both amine and thione sites.

As seen in Figure 4.6 (b), the copper-inhibitor complexation signals for AA2024-T3 are scaled with its copper content compared to those for the pure copper substrate. This is evident in the images of Cu^+ and Cu_2CN^+, where stronger Cu^+ and Cu_2CN^+ signals are observed over the aggregates (10-20 µm across) compared to the more homogeneously distributed background of these ions. These aggregates are most likely Cu-containing intermetallics, such as the commonly found Al_2Cu and Al_2CuMg, or the larger AlFeCuMn constituents. Figure 4.6 (c) also shows the image of Cu_2S^+, which is less abundant than Cu_2CN^+. The Al^+ image presenting the substrate shows contrast corresponding to the copper-rich aggregates. The image of the aluminium oxide cluster ion $Al_2O_4H_3{}^+$ also shows this trend, though to a lesser degree due to its weaker signals compared to Al^+. Therefore, as evidenced by the ToF-SIMS results, the inhibitor molecules displayed stable bonding to the intermetallics, with a greater portion through their nitrogen sites and a smaller portion through their sulfur sites.

4.3.6. THEORETICAL INSIGHTS INTO MOLECULE STABILITY AND CHEMISTRY

FIGURE 4.7 illustrates the speciation analysis of the inhibitor molecule at various pH values. Under the electrochemical conditions studied (1 mM 3-amino-1,2,4-triazole-5-thiol in 0.1 M NaCl), the pH is approximately 5.8. Around such pH values, the molecule is expected to predominantly exist in a mixture of its protonated, zwitterionic, and neutral thione forms, with neutral form being the majority species.

To study how such a species distribution might result in different forms with dif-

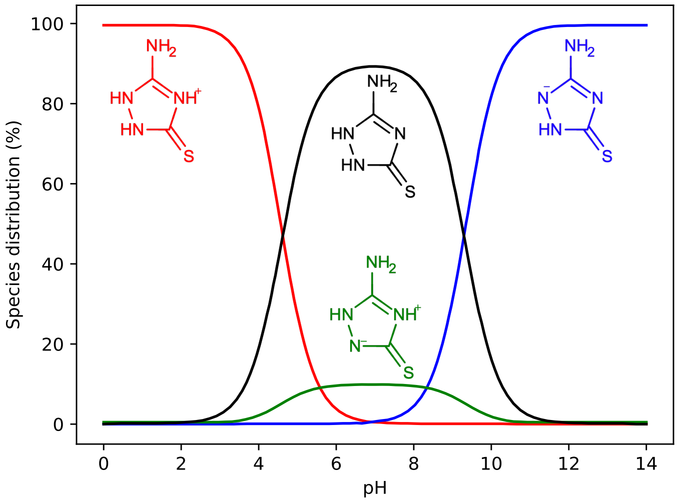


Figure 4.7: Speciation analysis of 3-amino-1,2,4-triazole-5-thiol for different pH values.

ferent adsorption-related properties of the molecule, deprotonated, neutral, zwitterionic and protonated forms of the molecule in thiol and thione tautomers were studied through DFT. Figure 4.8 visualises the key quantum chemical properties of these different species. Figure 4.8 (a) summarises the calculated quantum chemical parameters potentially relevant to the stable bonding configurations. Figures 4.8 (b) through 4.8 (e) present the dipole moment and Mulliken charges, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), and Fukui function surface maps, respectively. It is important to note that DFT analysis is undertaken here only to give an idea of where bonds might form. Adsorption of the molecule is not the only factor that determines corrosion inhibition, and other effects, such as intermolecular forces between inhibitor molecules, how well the formed organic film blocks corrosive species, and changes in the surface's electronic properties due to the adsorbed inhibitor, can be just as or more important in influencing the overall inhibition process [71].

The electronic properties of the HOMO, LUMO, and the HOMO-LUMO gap are helpful for understanding how a molecule interacts with a surface. The HOMO energy (E_{HOMO}) reflects the molecule's ability to donate electrons, with higher E_{HOMO} values indicating a greater likelihood of donating electrons to a surface with lower-energy empty orbitals. Similarly, the LUMO energy (E_{LUMO}) shows the molecule's ability to accept electrons, with a lower E_{LUMO} making it easier to accept electrons from donors. While the HOMO-LUMO gap provides insight into a molecule's reactivity, it is not directly correlated to corrosion inhibition or adsorption, as previously discussed [130]. More important for adsorption behaviour is the alignment of the molecule's HOMO and LUMO energies with the Fermi energy (and effective Fermi energy of the localised d-orbitals) of the surface [71, 267]. The Fermi energy represents the highest occupied

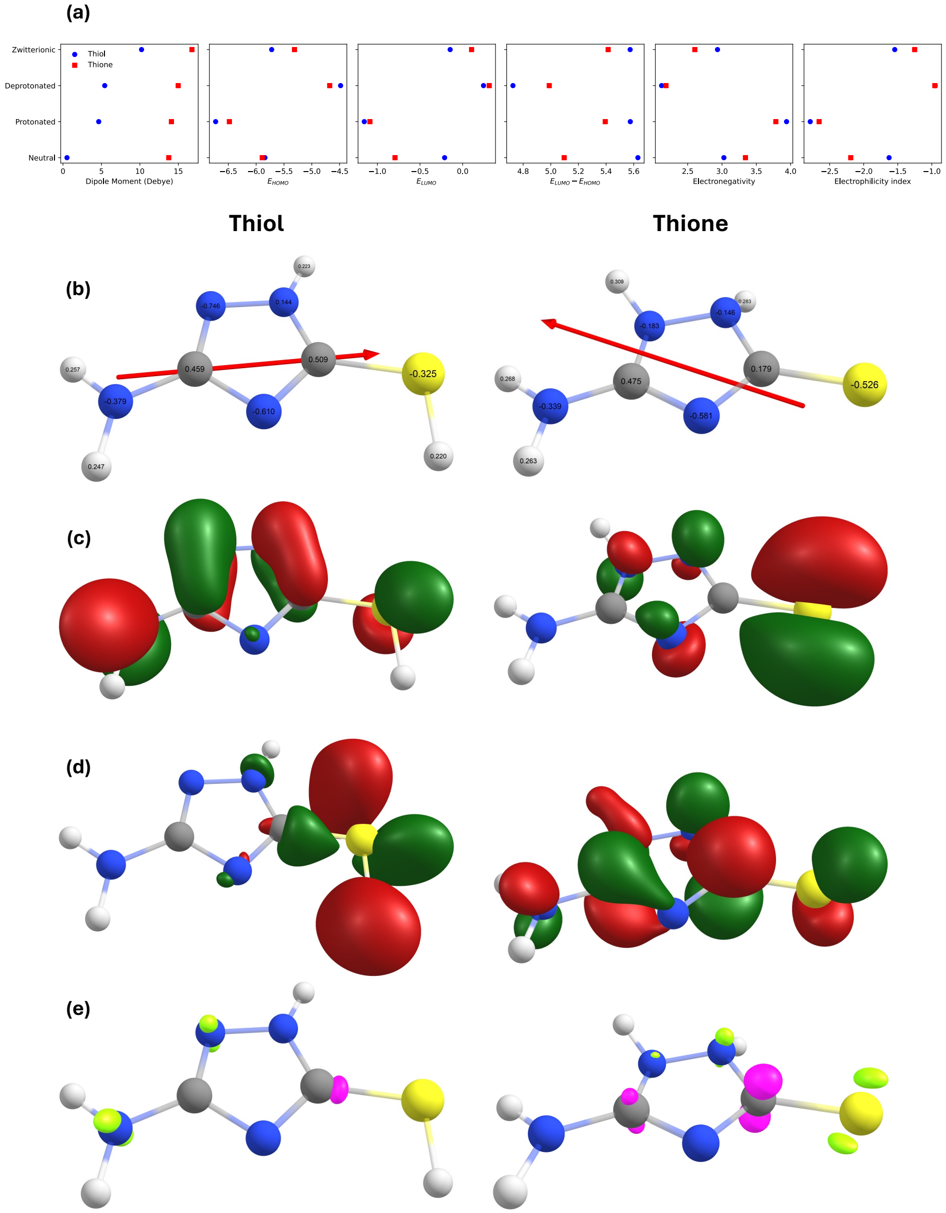


Figure 4.8: (a) Calculated quantum chemical parameters of different molecule species; for thiol and thione species (b) dipole moment and Mulliken charges, (c) HOMO, (d) LUMO, (e) Fukui function surface maps (lime electrophilic, pink nucleophilic attack sites).

electron state on the surface, and effective electron transfer during adsorption depends on the overlap of the molecule's electronic states with the surface's Fermi energy. A higher E_{HOMO} typically leads to an emptier anti-bonding state, while a lower E_{LUMO} results in a fuller bonding state, both of which enhance molecule-surface interactions. These interactions determine the strength and nature of adsorption. However, while strong molecule-surface bonding is important for corrosion inhibition, it is not sufficient on its own. Inhibitors must adsorb strongly enough to remain on the surface, but not excessively, as too strong bonding can weaken metal-lattice interactions and promote corrosion [76]. In summary, while the HOMO-LUMO gap indicates overall reactivity, and the alignment of the molecule's HOMO and LUMO with the surface's Fermi energy plays a central role in governing adsorption strength and electron transfer, they do not directly determine the corrosion inhibition effectiveness of the chemisorbed layer.

In addition to these electronic properties, the dipole moment informs about the molecule's polarity. A larger dipole moment generally means stronger interactions with polar surfaces which can help the molecule adsorb to a surface, whereas a smaller dipole moment may enhance molecular accumulation on the surface [71]. The molecule's electronegativity also plays a role, as more electronegative molecules tend to have stronger interactions with surfaces, especially those that are electron-deficient. Lastly, the electrophilicity index quantifies the molecule electron acceptance tendency, providing extra insight into its reactivity. Whereas state-of-the-art simplified DFT calculations fail to predict reasonable values, especially for quantities based on accurate HOMO-LUMO gap values, the approximation was proven to be still useful to ascertain trends of similar molecules [71].

These trends of the aforementioned quantum chemical properties dipole moment, E_{HOMO}, E_{LUMO}, molecular orbital energy gap E_{LUMO} - E_{HOMO}, electronegativity and electrophilicity index for the zwitterionic, deprotonated, protonated and neutral triazole forms can be observed in Figure 4.8 (a). In all forms, thione tautomer had a larger dipole moment than the thiol. The difference was on average around three times, and specifically for the neutral forms more than an order of magnitude. This huge difference between neutral thiol and thione forms, combined with their dipole moment vector pointing at opposite directions as visible on Figure 4.8 (b) would definitely influence the adsorption mechanisms. For both tautomers dipole moment values increased in the order of forms: neutral<protonated<deprotonated<zwitterionic.

Similar common trends for both tautomers were observed for E_{HOMO}, E_{LUMO}, E_{LUMO} - E_{HOMO}: protonated<neutral<zwitterionic<deprotonated, with thione having almost always larger values. One notable exception was the neutral case for E_{LUMO}, which was much lower for the thione tautomer. Protonation decreased E_{HOMO}, and deprotonation increased it, which suggest in the case of deprotonation molecule-surface interactions would increase through molecule electron donation to the surface. Similar trends for E_{LUMO} meant bonding interactions in the opposite direction - charge transfer from the surface to the molecule was favored in the case of molecule protonation. A significantly lower E_{LUMO} value for the neutral thione tautomer also suggested that such electron donation from surface to molecule LUMO type of interactions would occur more easily for it. The molecular orbital gaps E_{LUMO} - E_{HOMO} were always larger for thiol, except for the protonated form. The electronegativity and electrophilicity index trends for protonated, neutral, and deprotonated forms of a molecule show opposite trends. As the

molecule transitions from deprotonated to neutral to protonated, its electronegativity increases due to the electron-deficient nature of the protonated form, which attracts electrons more strongly. Conversely, the electrophilicity index decreases in the same order, as deprotonation creates an electron-rich species less prone to accept electrons. The protonated form, being electron-deficient, is more electrophilic - the index decreases as the molecule becomes more likely to accept electrons.

Experimental work in previous sections pointed towards having thione tautomer as the stable configuration. This can be also estimated from the thermochemistry of the molecules. To compare the relative stabilities of the thiol and thione tautomers of the molecule, the ratio of their concentrations at equilibrium can be derived from the Gibbs free energy difference between thiol and thione tautomers:

$$\Delta G = \Delta H - T\Delta S \tag{4.10}$$

where ΔH is the difference in enthalpy, ΔS is the difference in entropy, and T is the temperature in Kelvin. The relationship between calculated Gibbs free energy difference ΔG of the tautomers and their equilibrium constant K can be utilised as:

$$K = \frac{N_{\text{thiol}}}{N_{\text{thione}}} = e^{-\frac{\Delta G}{RT}} \tag{4.11}$$

where N_{thiol} and N_{thione} represent the number of molecules of the thiol and thione tautomers, respectively, ΔG is the Gibbs free energy difference between the two tautomers, R is the universal gas constant, and T is the temperature in Kelvin. From these calculations, the ratio of the thiol to thione tautomer was found to be approximately 1 to 130, indicating that there is a strong thermodynamical preference towards the thione tautomer.

The conditions in solutions would not necessarily enforce the same conditions in the vicinity of the surface, yet this finding agrees with the results observed in previous sections towards stable bonding configuration involving the thione form. However most likely both forms take part in transient bonding configurations observed in previous experimental analysis.

Previously discussed quantum chemical trends can be summarised for thiol and thione tautomers, which seem to indicate different bonding configurations, specifically for the thione: i) a much bigger dipole moment in the opposite direction (partial negative charge pointing away from the sulfur in thione vs. from the amino group in thiol), ii) similar E_{HOMO} yet much lower E_{LUMO} value, iii) a higher electronegativity and lower electrophilicity index. From this it is likely that both tautomers take part in electron donation to the surface, but especially the thione tautomer also is likely to be involved in electron donation from surface to the molecule LUMO, and possibly retrodonation.

The specific potential bonding sites can be analyzed with the help of Figures 4.8 (b) to 4.8 (e). Mulliken charges in Figure 4.8 (b) show lowest values for the triazole ring nitrogen with double bonds and sulfur atoms. Thione has a much lower value for the sulfur, indicating a more electron-rich environment, whereas a similar case is observed for the double bonded triazole nitrogen for the thiol. Rest of the atoms have similar charges. This suggests a surface bonding through the triazole ring for the thiol, and a bidentate-like bonding that involves both nitrogen and sulfur for thione tautomer.

Figures 4.8 (c) and Figure 4.8 (d) show HOMO and LUMO orbitals. Thiol HOMO shows activity over the whole molecule, whereas thione surfaces are more prominent on sulfur and double bonded nitrogen. Thiol LUMO isosurfaces are concentrated around sulfur, in contrast thione LUMO is spread over the whole structure. We once again need to remark that HOMO-LUMO interactions is an indication rather than a rule, as it was shown that even a molecule as small as simple triazole can interact with surfaces with orbitals other than HOMO and LUMO [71].

Figure 4.8 (e) plot the Fukui analysis. The lime colour indicate areas with excess negative Fukui charge, corresponding to sites prone to electrophilic attack, while purple highlights regions with excess positive Fukui charge, suggesting nucleophilic reactivity. For the thiol tautomer, the potential electron donation sites were the amino group and neighbouring double bond nitrogen, and the electron-withdrawal site was the carbon bonded to sulfur. For the thione tautomer, the potential electron donation sites were the sulfur and nitrogen with proton closest to sulfur, and the electron-withdrawal sites were the carbon atoms.

4.3.7. Mechanistic hypothesis for a quasi-sustained corrosion inhibition

3-amino-1,2,4-triazole-5-thiol is a good corrosion inhibitor for NaCl containing environments, previously exhibiting corrosion inhibition behaviour for alloys of magnesium [268], copper [269–271], iron [272, 273], and aluminium [80, 109, 136, 274–276]. However what makes it unique among other good organic corrosion inhibitors is that it displays a remnant of its original corrosion inhibition activity even when it is no longer sustained in the environment.

The literature hints at a hypothesis for the reason behind this behaviour. Previously it was observed that the adsorption of the molecule on aluminium surfaces increased the stability of the aluminium oxide and assisted the formation of Al-O bonds by preventing aluminium chloride and oxychloride complexes [276]. For AA2024-T3 substrates, molecules covered the whole surface with a film, where adsorption on Al-matrix resulted in anodic inhibition, while a concurrent adsorption on Cu-rich intermetallics led to cathodic inhibition [275]. Similar molecule structures with a triazole-ligand also provided hydrophobicity to the adsorbed surfaces [277]. Periodic DFT and molecular dynamics calculations of 3-amino-1,2,4-triazole-5-thiol on copper surfaces showed that main adsorption took place through the 1,2,4-triazole ring nitrogen, where coordination bonds with copper d-orbitals are formed, resulting in a bonding configuration where molecules lie flat on the triazole ring [271]. One work on Al/Cu galvanic model systems inhibited by Ce/3-amino-1,2,4-triazole-5-thiol found that whereas the molecule interacted with both surfaces and prevented ingress of chloride ions to reduce galvanic coupling, during sputtering nitrogen XPS signal was more present for the Al surface [274].

Adsorption studies of 3-amino-1,2,4-triazole-5-thiol on Ag and Au surfaces revealed that in solutions exceeding 0.1 mM molecule concentrations, the molecule primarily adsorbs onto the surfaces by forming a metal–thiolate bond (with also a minor thione contribution on Ag). In this configuration, molecule adopts an approximately perpendicular orientation that is stabilised by π–π stacking between adjacent triazole rings, as well as hydrogen bonding involving either neighbouring amine groups or surrounding solvated species. In contrast, for conditions of limited adsorbate availability at the surface, the molecule can alternatively bind through nitrogen atoms of the triazole ring in a deprotonated form, resulting in a flat-lying orientation. These two distinct binding configurations can reversibly interconvert by adjusting solution concentration, pH, or applied electrode potential, which would have a profound implication on the electron transfer properties of the resulting surface [250]. This reversible adsorption behaviour was also observed for Ag ions reversibly adsorbing on a chelating polymer derived from 3-amino-1,2,4-triazole-5-thiol [278].

Building on these past findings and results presented in this work, we propose that the sustained corrosion inhibition arises from differences in the adsorption strength and configuration of the molecule on the matrix compared to the intermetallics. Our electrochemical results indicate that the dominant corrosion inhibition effect is through the inhibition of the anodic reactions, with a minor influence on localised corrosion. In the presence of the molecule in sufficient amount, an insulating film is created by molecule-surface complexes formed on both the matrix and the intermetallics. As visible from the SKPFM measurements, this film is on the order of tens of nanometers thick, and covers the complete sample surface. Intermetallics seem to be covered with a thicker layer, which was similar to the previously observed behaviour for when pure copper was acting as cathode in an Al-Cu galvanic couple [274]. Compared to pristine surfaces, molecule adsorption resulted in higher surface potentials of the matrix, and decreased surface potential differences between the matrix and the intermetallics, which is in line with surfaces covered with insulating multilayers. In the subsequent absence of the molecule, a thinner layer was still sustained on the matrix, but this layer became even thinner on the intermetallics. The thinning layer and matrix-intermetallic surface potential differences comparable to the pristine surfaces indicate a reversible sort of interaction on intermetallics as compared to the matrix.

It has been previously argued that low molecular weight amines tend to desorb, and fail to exhibit a protective thin film effect on Fe or Zn, and even initiate Cu corrosion when used as volatile corrosion inhibitors - thus their presence in the corrosive medium must be sustained for corrosion protection; whereas heteroalkylated amines exhibited stable adsorption [279, 280]. Our experiments seem to show a similar behaviour, where a bonding through sulfur and oxide seems to be the dominant mechanism, and responsible for the quasi-sustained corrosion inhibition. When the molecule is found in sufficient concentration in the environment, a thicker multilayer of the molecule in different forms are adsorbed on the whole surface, but when the molecule drops below the critical concentration most of this bonding is washed away and only a sort of sulfate bond remains on the surface. $Al(OH)_3$ was calculated to be the most stable Al product between pH 4-12 [274], and our results suggest that the S moiety of the molecule forms a bond with this matrix (hydr)oxide, resulting in a molecule-oxide sulfate-like structure. When the molecule is washed away from the environment, this chemisorbed bonding configuration remains on the matrix. For adsorption on the intermetallics, the situation seems to be similar to the behaviour previously observed for Ag and Au [250], where approximately perpendicular bonding converts into a flat configuration with the triazole ring parallel to the surface. This mechanism is depicted as a schematic in Figure 4.9.

A strong molecule–surface interaction involved in chemisorption should leave sig-

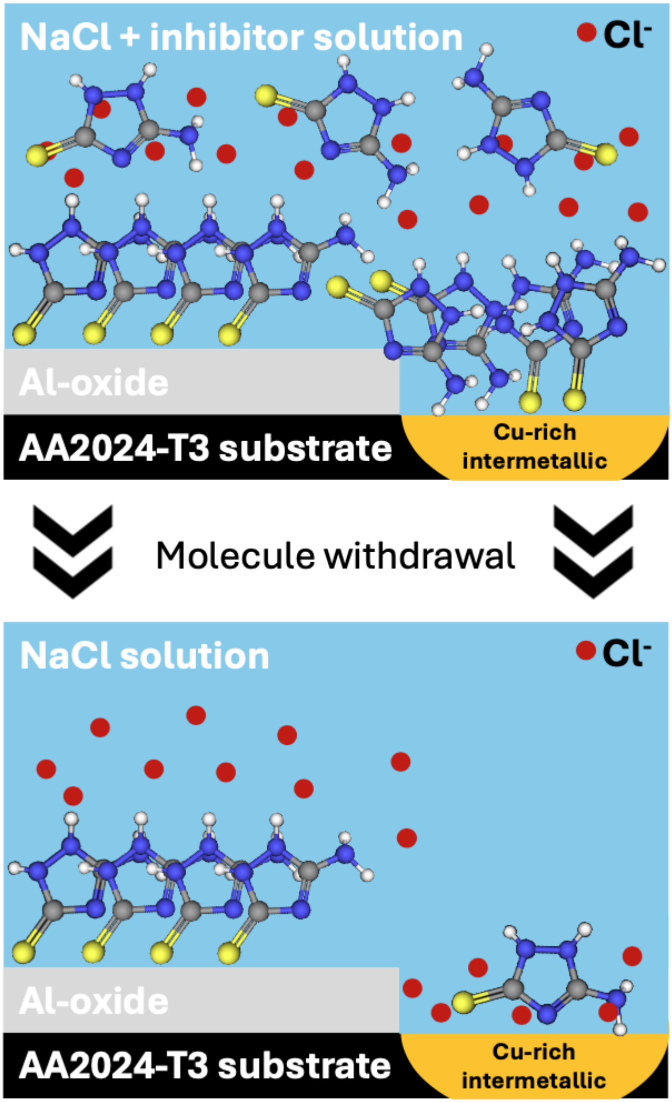


Figure 4.9: Schematic illustration of the quasi-stable inhibition behaviour offered by 3-amino-1,2,4-triazole-5-thiol.

natures detectable by spectroscopy [76]. Only the sulfate-like signals remained for various used spectroscopies, with the exception of ToF-SIMS. Assuming a majority thione tautomer of the molecule, the considerable reduction in surface potentials in the subsequent absence of the molecule suggests a perpendicular adsorption configuration, driven by significant dipole moment with sulfur oriented toward the surface, supported by the absence of parallel bonding evidence. XPS, Raman, and FTIR confirm bonding primarily through sulfur (as stable sulfate species post-washing) and transient nitrogen interactions. Cu surfaces favored nitrogen bonding, and Al/Al-OH surfaces showed chemisorbed S–O bonds. TOF-SIMS indicated bonding fragments by both S and N, which was higher for Cu surfaces and AA2024-T3 intermetallics - this suggests either coexistence of multiple major bonding configurations, or a dominant flat configuration where molecule is bonded through all active moieties of amino, triazole ring and thiocarbonyl/mercapto group.

In the presence of the molecule, S and π-bonds possibly work together in a bidentate sort of configuration where π-bonds most likely have weak interactions with Al or Al-OH, and S chemically bonds with O. Possibly amino group also assists this bonding, or does its own weak physisorption. Amino groups likely facilitate hydrogen bonding to the surface, as hinted by appearing symmetric and absent asymmetric -NH_2 FTIR stretches in the subsequent absence of the molecule, suggesting restricted amino motion. In contrast, modes such as the NH-NH stretch and S=C-NH stretch remain unchanged, implying these vibrations are less involved in surface interactions or that the molecule retains flexibility in these regions. The increase in C=N stretch suggests that the molecule regains this vibration mode which was possibly used for a transient bidentate configuration of S and N, where potentially S is chemisorbed to oxygen and N is physisorbed to Al. Tentative out of plane bending for N-H and time-dependent Al–N peak evolution on pure Al could indicate dynamic reorientation from perpendicular to parallel adsorption, consistent with Raman selectivity for perpendicular bonds [250], though absence of Raman nitride signals on AA2024-T3 points to a stable bonding relying on sulfatised Al-oxide rather than nitrogen based interactions. Upon inhibitor removal, S remains on the surface via sulfate bonds, while NH groups retain protonation, ruling out strong chemisorption via the triazole ring. Therefore the data points to triazole ring and amino group interactions on alloy matrix being transient, and sulphur being the central figure behind quasi-stable corrosion inhibition, which would be boosted by the low LUMO of the thione tautomer that would facilitate electron exchange.

Even when present on the surface after the molecule withdrawal from the environment, the thinner molecule-intermetallic bonding would not be enough to act as a barrier to chloride ingress in saline media. A recent DFT study on Al/Al-oxide surfaces has shown that although chloride penetration barrier increases with the thickness of the self-assembled monolayer, steric hindrance alone is insufficient to effectively prevent chloride penetration, as structural inhomogeneities within the monolayer exert a significantly greater influence [281]. With a parallel adsorption configuration, intermetallics would both have a lower penetration barrier, and more structural inhomogeneities for chloride penetration. This seems to be resulting in an adsorbed layer that is not effective in corrosion inhibition. The remaining bonding only from sulfur (as opposed to a potential bidentate configuration with π-π stacking, or a surface fully covered with a multilayer) reduces quality of the self-assembly layer as well, more easily allowing chloride-like corrosive species to the interface. This would cause a decrease in corrosion inhibition efficacy, but still be superior to a surface unexposed to the molecule.

4.4. Conclusions

This research was conducted to understand the quasi-sustained corrosion inhibition behaviour observed for AA2024-T3 exposed to saline media as a result of the presence of 3-amino-1,2,4-triazole-5-thiol. When the molecule was present in 0.1 M NaCl in 1 mM concentrations, it resulted in an inhibition efficiency of 91%. In its subsequent withdrawal from the environment, the molecule exhibited a quasi-sustained inhibition behaviour, and after three days of exposure to only 0.1 M NaCl, the molecule still provided 42% inhibition efficiency.

The spectroscopy measurements and quantum chemical calculations performed to unveil this phenomena suggest that when present in the environment in sufficient amounts, the molecule covers the surface completely. A sort of sulphate-like bonds to the Al-(hydr)oxide matrix, and intermetallic-molecule interactions with N and S moieties were observed, both of which most likely adsorbed approximately perpendicular to the surface. When the molecule was no longer supplied in the environment, most adsorbed molecules on the matrix and intermetallics desorbed. The remaining molecules on the intermetallics changed their orientation to a flat configuration, decreasing their corrosion inhibition likelihood. Meanwhile, a sulphatised Al-(hydr)oxide kept stabilising the oxide film through hindering the ingress of aggressive ions, thus sustaining the corrosion inhibition, albeit at a reduced efficacy.

4.5. SUPPLEMENTARY INFORMATION

4.5.1. X-RAY PHOTOELECTRON SPECTROSCOPY (XPS)

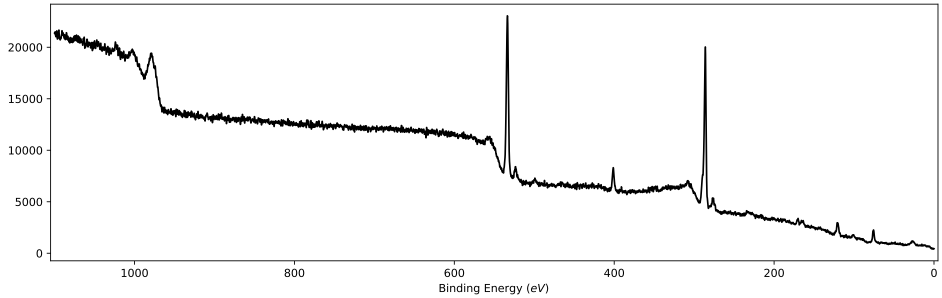


Figure 4.10: Full survey XPS spectra of AA2024-T3 sample exposed to 1 mM 3-amino-1,2,4-triazole-5-thiol dissolved in water for 24 hours.

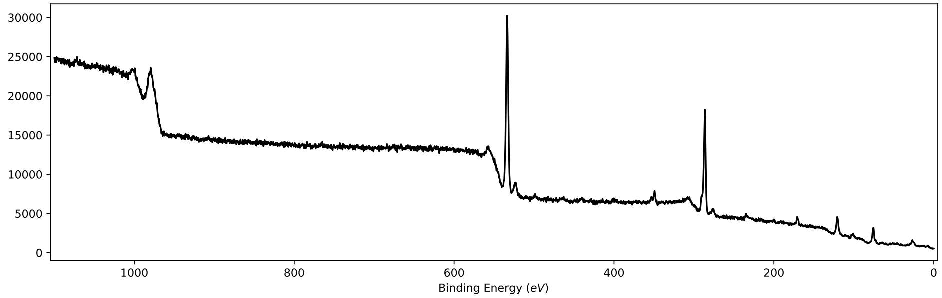


Figure 4.11: Full survey XPS spectra of AA2024-T3 sample exposed to 1 mM 3-amino-1,2,4-triazole-5-thiol dissolved in water for 24 hours, followed by exposure to only water for 2 hours.

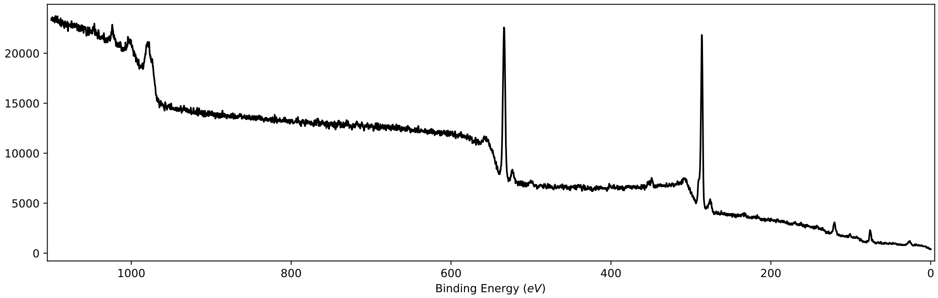


Figure 4.12: Full survey XPS spectra of AA2024-T3 sample exposed to only water for 24 hours, followed by exposure to only water for 2 hours.

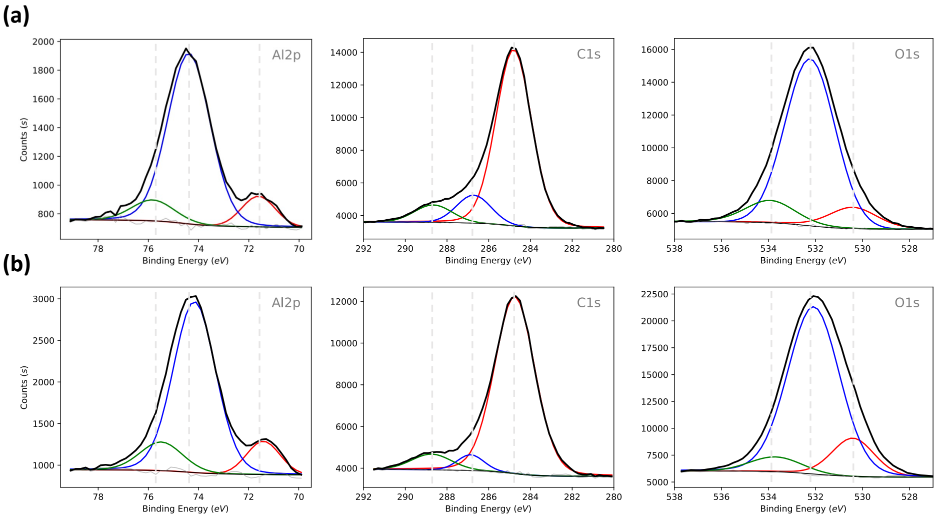


Figure 4.13: Remaining high-resolution XPS spectra for (a) inhibitor presence, (b) subsequent inhibitor absence.

4.5.2. SHELL-ISOLATED NANOPARTICLE-ENHANCED RAMAN SPECTROSCOPY (SHINERS)

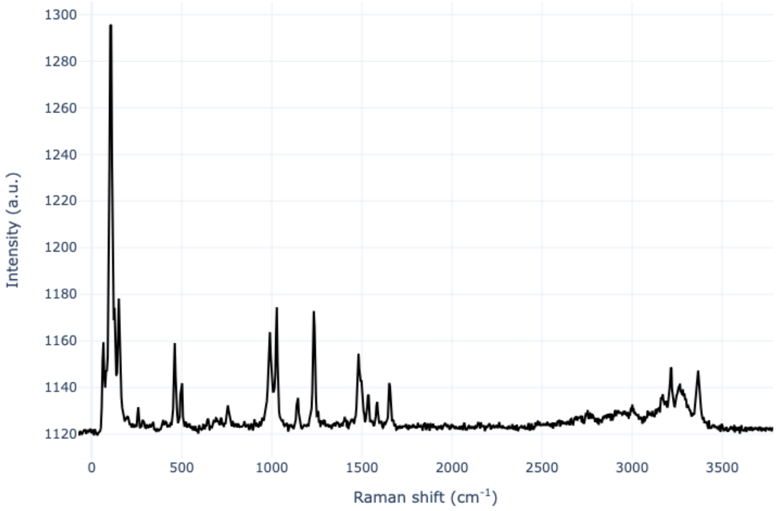


Figure 4.14: Raman spectra of 3-amino-1,2,4-triazole-5-thiol powder.

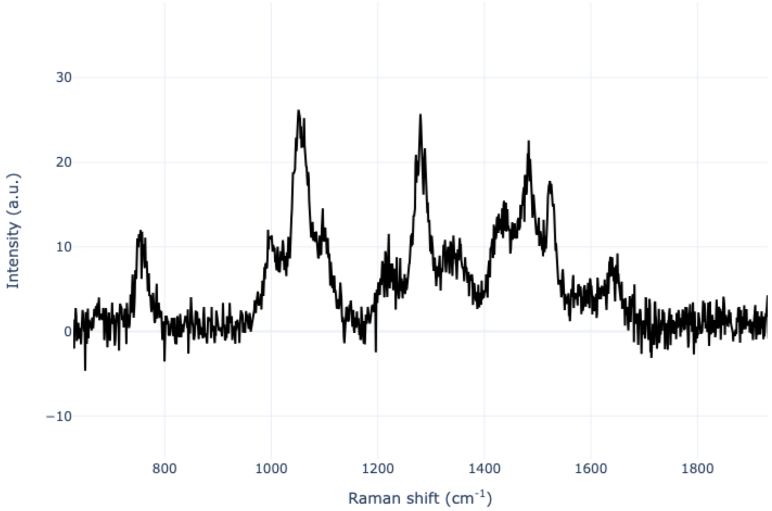


Figure 4.15: Raman spectra of 3-amino-1,2,4-triazole-5-thiol powder dissolved in water.

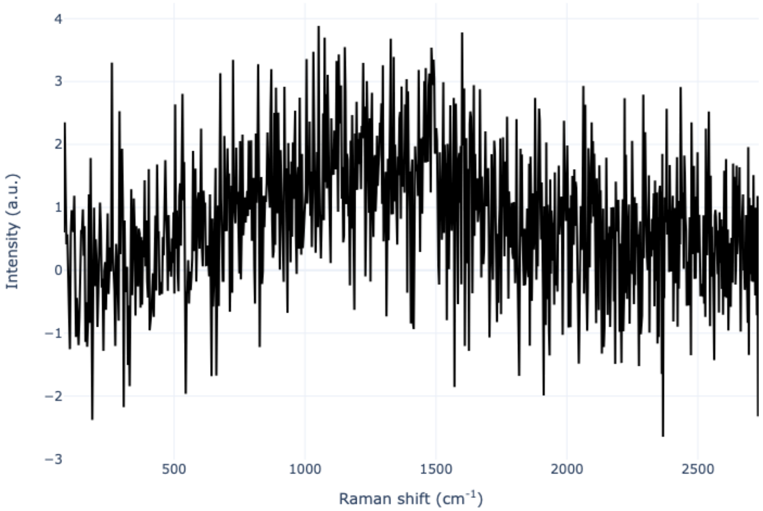


Figure 4.16: Raman spectra of AA2024-T3 surface without nanoparticles.

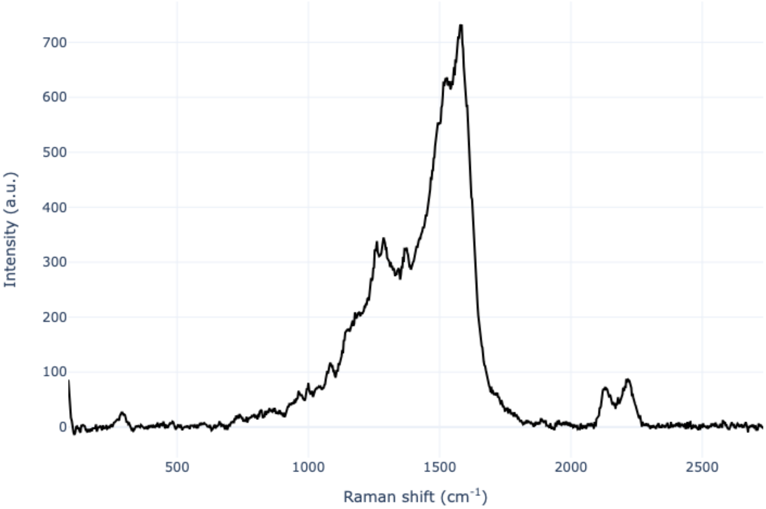


Figure 4.17: Raman spectra of AA2024-T3 surface with nanoparticles.

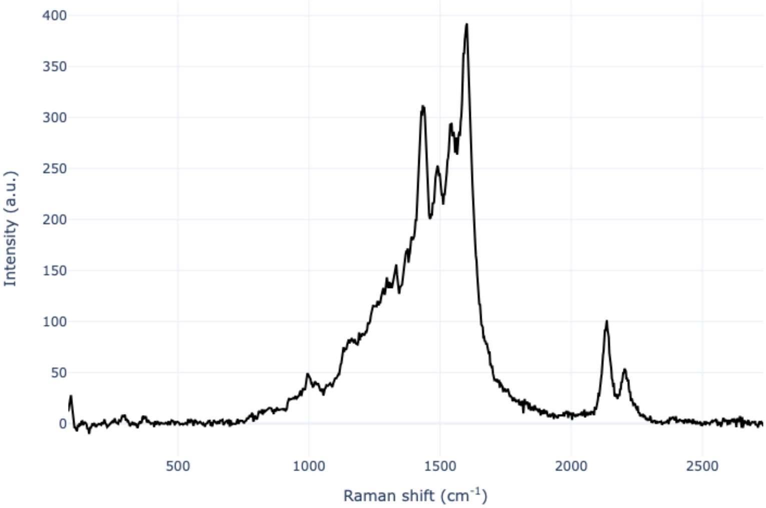


Figure 4.18: Raman spectra of AA2024-T3 with nanoparticles exposed to water.

4.5.3. TIME-OF-FLIGHT SECONDARY ION MASS SPECTROMETRY (TOF-SIMS)

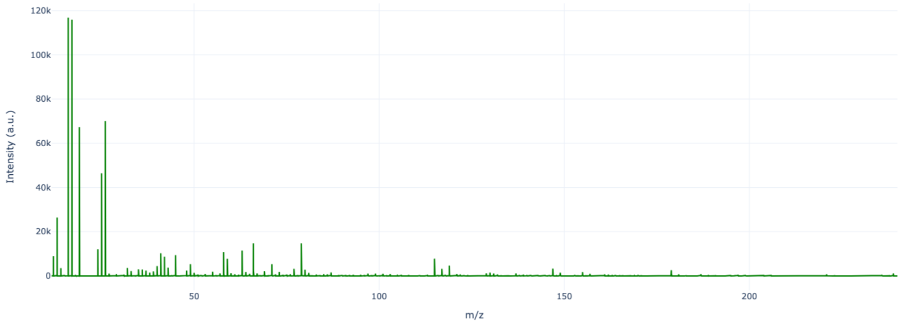


Figure 4.19: Negative ion spectra of AA2024-T3 exposed to 3-amino-1,2,4-triazole solution.

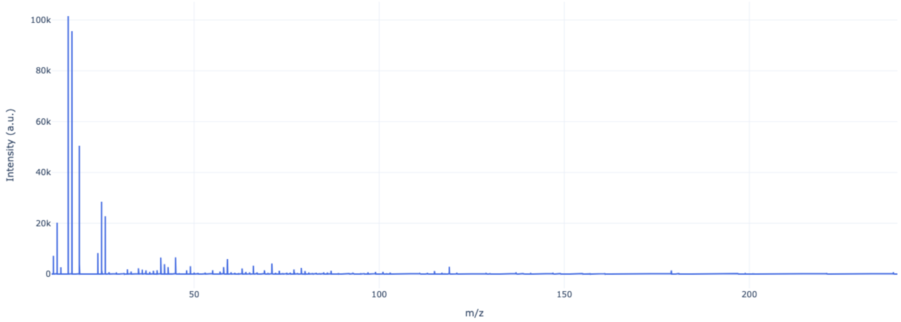


Figure 4.20: Negative ion spectra of Al exposed to 3-amino-1,2,4-triazole solution.

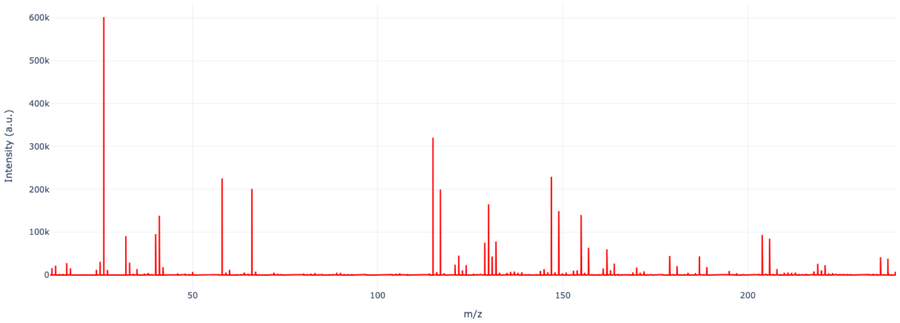


Figure 4.21: Negative ion spectra of Cu exposed to 3-amino-1,2,4-triazole solution.

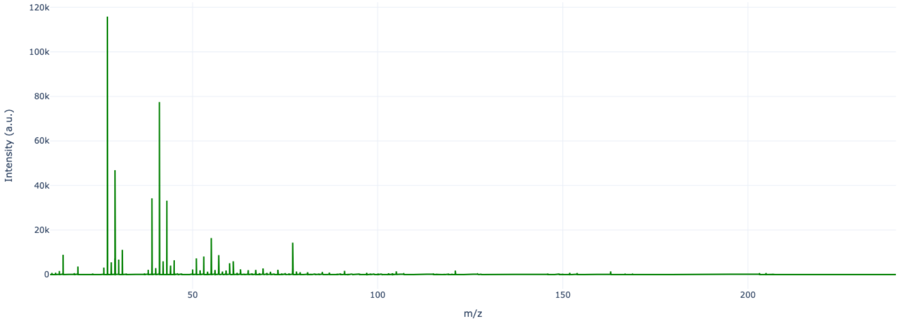


Figure 4.22: Positive ion spectra of AA2024-T3 exposed to 3-amino-1,2,4-triazole solution.

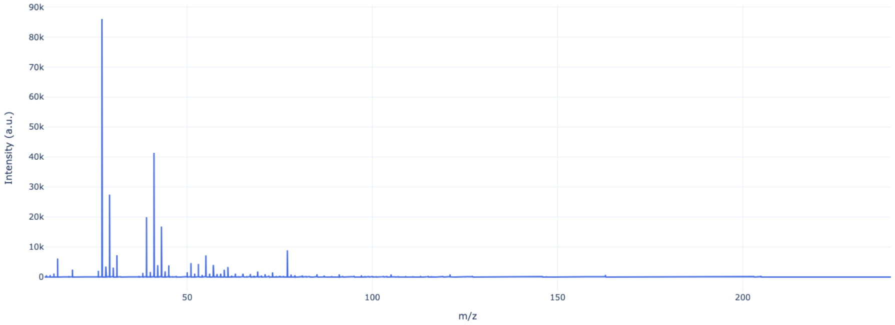


Figure 4.23: Positive ion spectra of Al exposed to 3-amino-1,2,4-triazole solution.

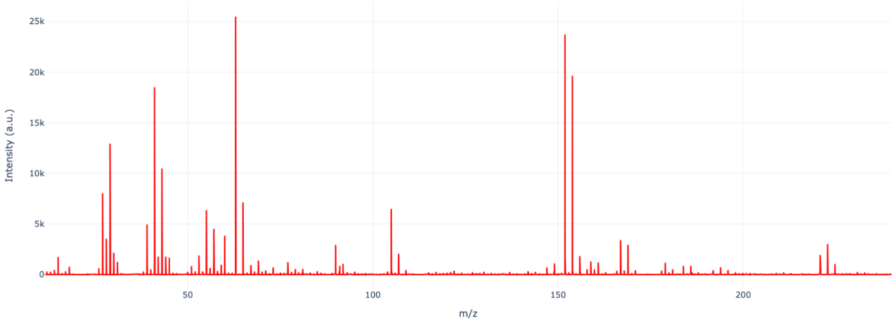


Figure 4.24: Positive ion spectra of Cu exposed to 3-amino-1,2,4-triazole solution.

4.5.4. DENSITY FUNCTIONAL THEORY (DFT) CALCULATIONS

An example ORCA 6.0 input file for the thione tautomer of 3-amino-1,2,4-triazole-5-thiol in its neutral form is presented below.

```
# thione-neutral

! B3LYP D3BJ OPT FREQ RIJCOSX def2-TZVPD TightSCF CPCM PAL8
%cpcm
    smd true
    SMDsolvent "WATER"
end
! LargePrint PrintBasis PrintMOs

* xyz 0 1
   S        -2.51783       1.87005       0.12780
   C        -1.37229       0.66709       0.11623
   N        -0.00169       0.78971       0.05579
   C         0.42091      -0.42487      -0.12458
   N        -0.54238      -1.39186      -0.21701
   N        -1.70947      -0.67617       0.10310
   N         1.70968      -0.79406      -0.24729
   H         2.35181      -0.03966      -0.02451
   H         1.98173      -1.67614       0.16238
   H        -0.38007      -2.20139       0.37446
   H        -2.58944      -0.94072      -0.32637
*
```

5

Describing Inhibition

All models are wrong, some are useful.

George Box

If you want to understand function,
study structure!

Francis H. Crick

Despite the remarkable success of machine learning in materials science, challenges persist in gaining mechanistic insights, especially in low-data regimes where dataset sizes limit the precise applicability of machine learning. The prevailing reliance on high-confidence predictions from the models often leaves the underlying decision-making mechanisms opaque, limiting scientific understanding. This study presents an alternative approach that emphasises understanding the model decision-making process over individual predictions, enabling the extraction of scientifically meaningful insights from small datasets. We reveal common trends by reverse engineering the best-performing models based on featurisation methods of physicochemical descriptors, hashed fingerprints, and structural keys, which we integrate with domain knowledge to create a molecular substructure template for candidate molecules. Using this template, we filter a toxicity database to identify non-toxic corrosion inhibitors, aiming to replace the de facto but hazardous corrosion inhibitor hexavalent chromium. The resulting candidate's efficacy is validated through electrochemical testing, illustrating the feasibility of achieving mechanistic insights from statistical models in data-scarce environments.

This chapter has been published with some adjustments as *Gaining scientific understanding with small data machine learning: explainable molecule representations and their consensus*, npj Materials Degradation (2025) [282].

5.1. Introduction

In Douglas Adams' Hitchhikers Guide to the Galaxy an alien race seeks the "Answer to the Ultimate Question of Life, the Universe, and Everything". A planet-sized computer, Deep Thought, is tasked with calculating the answer. After seven and a half million years, it finally reveals the answer: 42. This result baffles the programmers, as they realise they don't actually know what the "Ultimate Question" is. Deep Thought explains that without understanding the true nature of the question, the answer has no real meaning.

We face a similar challenge today in machine learning applications to materials science. On the one hand, machine learning models have been widely successful in materials sciences for autonomous experimentation, materials property interpolations, structure optimisation, and chemical space exploration [283–288]. Recent advances, such as creating continuous material property representations through generative autoencoder approaches [289, 290], and utilising the inherent graph structure of molecules through graph neural network architectures [291, 292], allowed researchers to investigate previously unexplored chemical spaces. On the other hand, training such methods is resource-intensive, requiring substantial computational power and large datasets — more than ten thousand, in some cases tens of millions of data points for predictive accuracy. Data scarcity and expensive target generation prevent cutting-edge architectures to be used with experimental datasets, and limits the use of such models to simplified DFT simulation predictions. Yet, even with high benchmark scores, a recent study showed that model predictions may not generalise well to new materials spaces [293]. Most importantly, generalised or not, the scientific insights within these models remain largely opaque, meaning that even when models make accurate predictions, the mechanisms behind these "answers" are not transferred to scientists [294, 295]. Like Deep Thought's answer, these predictions are limited in their impact without a deeper understanding of the "questions" they are addressing.

The materials science community has already made significant progress in building the foundation for explainable and interpretable models [296–300]. However, most of the published literature on explainability focuses on bigger models, despite most materials discovery problems happening in low- to no-data regime. For this reason, with this paper we intend to go in the opposite direction of the present mainstream deep-learning focus and ask: Is it possible to gain scientific insights from predictive models with lower prediction metrics based on small datasets? The answer, we believe, is yes. We see that representation is key; with the right representation and reverse engineering machine learning models based on different data representations we can identify common trends - which combined with domain expertise can allow the scientist to see new trends that were previously unattainable.

Here we present an unorthodox framework to transform statistical models into scientific insight. Instead of relying on low-confidence individual predictions of models trained with scarce data, we train and then reverse engineer multiple models at once to understand their decision-making mechanism. We combine the resulting insights with the scientific intuition of the domain expert. Our work is performed on small organic molecules and their corrosion inhibition properties, but it can principally be applied to any materials discovery task.

Although approaches to interpret machine learning models such as permutation feature importance [301], local interpretable model agnostic explanations [302], and Shapley additive explanations [303] have been widely used in other scientific fields, specifically in the domain of corrosion and inhibition prediction there have been limited works [1, 304]. Compared to neighbouring scientific fields more often than not two points plagued the corrosion inhibition researchers (i) the features most important in predictive performance were identified, but no scientific reasoning were built on top of the models to reason why the model predictions were improving when they were present in the prediction process due to the complex multi-scale nature of corrosion phenomena, and (ii) the limited size of the datasets (often around 20 samples or so) caused eager but early conclusions to be drawn, such as the debated correlations between quantum chemical descriptors and corrosion inhibition performance [1, 130]. We aimed to instil more trustworthy conclusion to be drawn from predictive models by coming up with a method of reasoning of how different model representations can be aligned to solve the same problem, hypothesize for potential influences of the features, and validate our hypothesis experimentally.

For gaining scientific insights from statistical investigations, achieving the best representation is important. Here we converge to the best representation for our dataset by systematically analyzing 29 widespread open-access methods that convert molecules into a set of features (hereinafter referred to as “featurisation"), and 9 different target representations of experimentally acquired electrochemical data on around 100 molecules (for details see our previous work [80]), to be used as model targets. After settling on the optimum description by looking at different feature-target combinations (which results in more than 12 thousand model configurations), we can use complementary descriptions with lowest root mean squared error (RMSE) together to capture common trends existing in all. We use such trends to come up with a searchable template molecule that can be used to filter existing larger databases. In our case we use a toxicity database, as our goal is to find a non-toxic molecule with the potential to replace the currently in-use domineering hexavalent chromium: a highly inhibitive but deadly corrosion inhibitor prohibited by the EU REACH (Registration, Evaluation, Authorisation, and Restriction of Chemicals) regulation [8]. We validate the insight gained from this approach by performing electrochemical measurements of the recommended molecule, showing that gaining mechanistic insight with statistical models is possible for small datasets. The result of this work is expected not only to assist in developing green and sustainable alternative inhibition approaches to corrosion which eats away 3.1% of the global GDP [206], but also expected to serve as a guiding framework for other data-scarce materials discovery problems.

5.2. METHODS

5.2.1. GENERATION OF ELECTROCHEMICAL TARGETS

TARGET generation experiments are discussed in a detailed manner in our previous publications [80]. This work is based on 107 organic molecules tested as corrosion inhibitors for the same substrate AA2024-T3, which are provided in Supplementary Information. The targets corresponded to three different electrochemical experiments: electrochemical impedance spectroscopy performed at 24th hour (_EIS24h), linear polarisation resistance experiments performed at 24th hour (_24h) and linear polarisation resistance values averaged through time (_avg) with trapezoidal integration [79]:

$$\langle R_p \rangle = \frac{1}{t_f - t_0} \int_{t_0}^{t_f} R_p(t)dt \tag{5.1}$$

$$\approx \frac{1}{t_f - t_0} \sum_{k=1}^{N} \frac{R_p(t_{k-1}) + R_p(t_k)}{2} (t_k - t_{k-1}) \tag{5.2}$$

where t_f is the final measurement time, t_0 is the initial measurement time, and k is the index for the performed discrete measurements. The results from these experiments are represented in three different forms: raw electrochemical polarisation resistance values R_p, R_p values converted into inhibition efficiencies (IE) with:

$$IE = \frac{R_p^{inh} - R_p^{blank}}{R_p^{inh}} \tag{5.3}$$

and into inhibition power (IP) with

$$IP = 10\log_{10} \frac{R_p^{inh}}{R_p^{blank}} \tag{5.4}$$

where superscripts "inh" and "blank" stand for samples exposed to organic molecules or only to NaCl, respectively. These three different experimental approaches with three different representations resulted in nine different potential target values.

5.2.2. GENERATION OF FINGERPRINTS AND PHYSICOCHEMICAL DESCRIPTORS

THE SMILES strings of 107 small organic molecules are first desalted (removal of ionic metal parts from the strings) for correct descriptor calculation, then converted into structural fingerprints and physicochemical descriptors with the open-source python cheminformatics packages `RDKit` (v. 2023.03.3) [141] and `molfeat` (v. 0.9.2) [305] to use as features of the machine learning models. Using these packages, 29 different methods were chosen for converting molecules into tabular numeric features. These represented the most popular cheminformatics tools for digitising molecules. Every feature dataset was supplemented with pH-based experimental features: pH_{before} (pH measurement before the experiments), pH_{after} (pH measurement after the experiments), $pH_{average}$ (average of before and after values), and $pH_{bef-aft}$ (difference between before and after values). The resulting datasets contained between 18 and 2052 features.

5.2.3. MACHINE LEARNING MODEL TRAINING AND COMPARISON

TARGETS

THE 107 samples are split into two sets: a training set with 95 samples, and a set-aside validation set with 12 samples. This was achieved through the `verstack` (v. 3.9.2) [306] package using a continuous stratified split so that the target data distributions in both sets are statistically similar.

FEATURES

The feature datasets for training samples are cleaned with the help of `scikit-learn` package (v. 1.5.0) [307]. If the training datasets had missing values for any samples, these are filled with the median of that feature. To eliminate redundancies found in the molecular representation, the model features with variances lower than 0.1, and features correlated to others with a Pearson correlation value of more than 0.8 are removed. Afterwards, features are scaled with three different `scikit-learn` scaler functions to assist the model learning process: `MinMaxScaler`, `StandardScaler`, `PowerTransformer`. Features with and without scaling are algorithmically selected with two different sparse feature selection methods: one based on the recursive feature elimination method of `scikit-learn` which uses impurity-based feature importance on RF estimators (RFE), and the other based on recursive feature elimination of the `Probatus` package (v. 3.0.0) [308] which uses SHAP-based feature importance (RFE_{SHAP}). RFE was repeated 1000 times with random seeds, RFE_{SHAP} used 5-fold randomised cross validation search. Feature selection was carried out to prevent flooding the model with irrelevant features, as high-dimensionality of the feature space would result in fitting the noise rather than the signal, commonly known as overfitting. Another reason was to capture only the features most relevant to the mechanism of corrosion inhibition, which is expected from highly predictive features. The top ten selected features from RFE, or the optimum number of selected features revealed from RFE_{SHAP} were used for the actual models.

MODELS

After scaling and feature selection, different featurisation schemes are combined with different target representations to be modeled with four different regression architectures, three implemented in `scikit-learn`: `random forest` [309], `support vector machine` [310], `k-nearest neighbours` [311], and `xgboost` implemented in the `xgboost` package (v. 1.7.6) [312]. The optimisation scoring function used was negative root mean squared error. Bayesian optimisation was employed using the `bayes-opt` package (v. 1.4.3) [313] to find the optimal hyperparameters based on a 10-fold cross-validation score, where the data is divided into 10 random subsets. In each iteration, 9 subsets are used for training and the remaining 1 subset is used for testing, allowing the model's generalisation performance to be evaluated across different train test splits. Optimised hyperparameters and their ranges were:

- Random forest: number of trees (10, 1000), maximum tree depth (1, 50), minimum number of samples required to split (2, 25), maximum ratio of used features: (0.1, 1)
- Support vector machine: regularisation parameter C (0.001, 1000), the margin of tolerance ε (0.001, 10), kernel coefficient γ (0.001, 100), radial basis function kernel
- K-nearest neighbours: number of neighbours (1, 10), weighing for the neighbours (uniform or distance-weighed), the distance metric to be used for calculating 'neighbourhood' (Euclidean, Manhattan or Minkowski)
- XGBoost: number of trees (100, 1000), maximum tree depth (2, 10), learning rate (0.01, 0.1), fraction of the training data to be randomly sampled for tree construction (0.1, 1.0), fraction of features randomly sampled for tree construction (0.1,

1.0), minimum loss reduction gamma required for further leaf node partition (0.1, 1.0), L1 regularisation (0.001, 100), L2 regularisation (0.001, 100)

Detailed explanation of hyperparameters can be found in the `scikit-learn` documentation and textbooks [314]. Learning curves and prediction plots are recorded for further analysis. Regression performance was quantified with R^2, RMSE and MAE. After quantification, models are retrained with all training set with optimised hyperparameters and saved as pickle files for further experiments.

5.2.4. Visualising fingerprints

Atompair-count, rdkit-fp and ECFP fingerprints are analyzed to identify which molecular fragments the features correspond to, and further visualised through the code provided.

5.2.5. Generating best pseudomolecules through Bayesian Optimisation

The retrained models are optimised with Bayesian optimisation. Now, the optimised parameters were not the model hyperparameters, but the model input values of algorithmically selected features and the predictions of the selected model. The acquisition function used was upper confidence bound with the default implemented hyperparameters. The bounds for optimisation for each feature were set based on the minimum and maximum values observed in the original database, ensuring interpolation rather than extrapolation. On top of all initial real molecule samples, 2000 random samples were used to initialise the optimisation, and 1000 iterations were performed for optimising the pseudomolecule. Pseudomolecule feature scaling is inverted for further use for similarity analysis. The resulting features represent the optimal artificial molecule parameters according to the model, leading to the best target property and creating an ideal pseudomolecule.

5.2.6. Curating the toxicity dataset for pseudomolecule similarity hits

To find the molecules most similar to the pseudomolecule, a query molecule database is necessary. A database with a large collection of SMILES strings and experimental toxicity values was chosen as the candidate database [315]. The choice of selecting a toxicity database was deliberate. Aside from predictions from our model, such a database would provide information on the toxicity of compounds. The training and evaluation sets from Supplementary Table 2 of the original study were combined, and the SMILES strings along with their experimentally determined U.S. Environmental Protection Agency (EPA) toxicity hazard classifications were extracted. The EPA classifications corresponded to: I highly toxic, II moderately toxic, III slightly toxic, IV practically non-toxic [316]. After desalting the molecules, generating descriptors, and cleaning the dataset, this process yielded over 10,000 candidate molecules. The similarity between molecules was calculated with cosine similarity $S_{\cos}$, where the similarity between two vectors is calculated as:

$$S_{\cos}(M,P) = \frac{M \cdot P}{||M||||P||} = \frac{\sum_{i=1}^{n} M_i P_i}{\sqrt{\sum_{i=1}^{n} M_i^2} \cdot \sqrt{\sum_{i=1}^{n} P_i^2}}, \tag{5.5}$$

where M_i and P_i are the i^{th} components of vectors M and P, respectively, corresponding to the vector of query molecule and optimised pseudomolecule, respectively. The resulting cosine similarity values span from 1 to -1, from 1 meaning the vectors are oriented in the same direction (complete similarity), to -1 meaning vectors are oriented in the opposite direction (complete dissimilarity), and 0 indicating orthogonal vectors (decorrelation). In-between values indicate intermediate similarity/dissimilarity.

5.2.7. SHAP (SHAPLEY ADDITIVE EXPLANATIONS) ANALYSIS

SHAP values are a concept from cooperative game theory used to fairly distribute the *payout* among players based on their contributions. In machine learning, each feature value of the instance is a player in a game where the prediction is the payout. SHAP values are applied to interpret complex models by attributing the contribution of each feature to the model's prediction for a specific instance. `SHAP` package (v. 0.42.1) [317] was used to create SHAP beeswarm plots for optimised models based on 3 different featurisation methods: atompair-count, PaDEL and MACCS.

5.2.8. VALIDATION EXPERIMENTS THROUGH ELECTROCHEMICAL MEASUREMENTS

AA2024-T3 sheets with a thickness of 2 mm (Salomon's Metalen B.V., the Netherlands) were cut into 20 x 20 mm specimens using an automatic shear cutter. The samples were then sequentially ground with 320, 800, 1200, 2000, and 4000 grit papers on a rotating plate sander under running water, followed by cleaning in isopropanol for 15 minutes and drying with compressed air. The resulting specimens were used for electrochemical measurements. The electrochemical investigations consisted of observing the open circuit potentials (OCP) for 24 hours, where a linear polarization resistance (LPR) measurement was performed every hour to observe the time-dependent behaviour. For LPR measurements the potentials were scanned from -10 to +10 mV vs. OCP at a rate of 0.5 mV/s. After concluding the 24-hour observation, electrochemical impedance spectroscopy (EIS) measurements were performed, where a 10 mV peak-to-peak amplitude sinusoidal AC perturbation was applied from 10 kHz to 10 mHz frequency range with 10 frequency point per logarithmic decade. Flat three-electrode electrochemical cells (Corrtest Instruments, China) were used to perform the experiments at room temperature. The sample was used as the working electrode, platinum mesh was used as the counter electrode, and Ag|AgCl (saturated KCl) was used as the reference electrode. The exposed surface area was 0.785 cm^2, exposed to a 250 ml 0.1 M NaCl 1 mM 2-thiobarbituric acid electrolyte. The pH of the electrolyte was adjusted to 7.0 with an adequate amount of NaOH, by analysing the solution pH with a Metrohm 913 pH meter. All chemicals were purchased from Sigma-Aldrich. The electrochemical measurements were controlled with Biologic VSP-300 multichannel potentiostats with the help of EC-

Lab software. The electrochemical experiments were repeated three times to confirm the reproducibility of the experiments.

5.3. Results and Discussion

5.3.1. Describing corrosion inhibition

The description of the problem is key for the machine learning models to extract all possible information from the signal present in the data. This involves not only selecting the appropriate set of features (the set of numbers used for prediction) but also choosing the correct form for both the features and targets (the set of numbers to predict). This is not only a factor to consider when improving the prediction performance of models. It becomes even more important when models are used for gaining mechanistic insights similar to various spectroscopies, as demonstrated in this paper.

To reveal the best description of corrosion inhibition, we have set up a large span of target and feature representations: 9 targets, 29 featurisation methods, 3 feature selection approaches (include all, recursive feature elimination (RFE), SHAP-based), combined with 4 different feature scaling methods (no-scaling, minmax, standard, power) and 4 different regression model architectures (random forest (RF), XGBoost (XGB), support vector machine (SVR), k-nearest neighbours (KNN)). This search space of the optimal description consisted of more than 12,000 configurations. The models based on these configurations were trained with 95 small organic molecules with 10-fold cross validation, and validated with a left-out validation set of 12 other molecules. Target values were obtained from time-resolved electrochemical experiments of electrochemical impedance spectroscopy and linear polarisation resistance, discussed in more detail in our previous publication [80]. After training, the best models are identified by comparing cross-validation root mean squared (CV-RMSE) values (all converted into the scale of inhibition power IP). Afterwards, the models' hyperparameters are optimised with Bayesian optimisation, and the left-out validation set was used to check the prediction performance of the models. The ranking of the predictive performances of the models can be found in the Supplementary Figure 1-2.

Figure 5.1 presents the ranking of different featurisation methods for the best 4 targets, and their mean. In this case “best" means the ranking of models with the lowest CV-RMSE error for a given representation. The inset shows the CV-RMSE performance distribution of models with different featurisations, feature selection methods, feature scaling, model architectures; pooled for different targets. For the best target (IE_EIS24h), a similar pooling for all other configurations but the featurisation methods resulted in the performance distribution of different featurisation schemes unfolded on the right, with featurisation schemes corresponding to same labels as featurisation ranking for different targets on the left. This results in featurisation methods ordered from best to worst for target IE_EIS24h. Lowercase labels correspond to hashed fingerprints or structural keys (also highlighted with _fp suffix), capitalised labels correspond to physicochemical descriptors.

Targets

The 9 targets are the combination of data coming from 3 different electrochemical experimental methods, denoted with respective suffixes (Bode modulus at 10^{-2} Hz measured

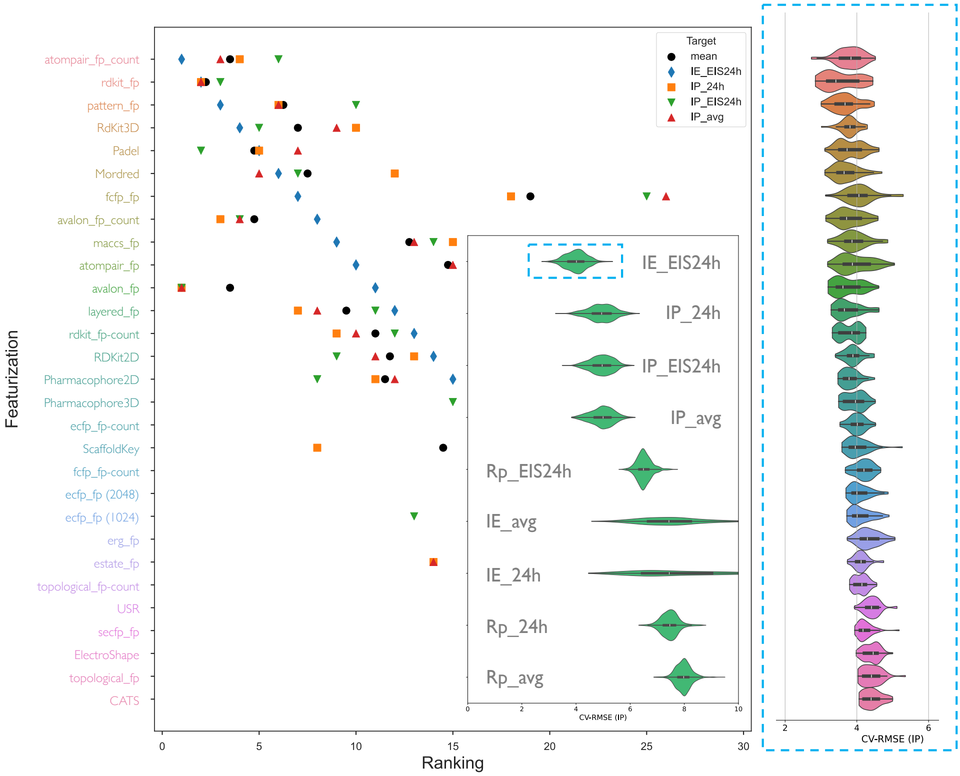


Figure 5.1: Ranking of featurisation methods based on prediction performance for the top four targets, and their mean. The y-axis lists featurisation methods, ordered from best to worst for the target IE_EIS24h, which resulted in models with the lowest cross-validation root mean square error (CV-RMSE). Physicochemical descriptors are capitalised, whereas structural keys/fingerprints are denoted with suffix _fp in case of absence/presence based encoding, and _fp_count for count based encoding. The inset displays CV-RMSE distributions for models across all targets, incorporating variations in featurisation methods, feature selection approaches, feature scaling, and model architectures. For the target IE_EIS24h, the distribution of prediction performance for each featurisation method is shown on the right, aligned with the corresponding rankings on the left.

at 24th hour through electrochemical impedance spectroscopy, _EIS24h; linear polarisation resistance measured at 24th hour, _24h; linear polarisation resistances averaged through the first 24 hours, _avg), represented in 3 different forms (raw electrochemical data, R_p; inhibition efficiency, IE; inhibition power, IP).

Compared to the rest of the factors, target representation by far had the most impact on the predictive performance of the models. Looking at the CV-RMSE distributions for the different targets, the best model for the best target IE_EIS24h resulted in a CV-RMSE of 2.73, whereas the best model for the worst target Rp_avg resulted in a CV-RMSE of 6.87, an increase of 152%.

Models based on time-averaged experiments (_avg) performed worse than others, indicating that prediction of time-dependent phenomena might be more difficult than prediction of a stabilised reaction after a given time, in this case after 24 hours. R_p models also showed poorer prediction performance compared to IE and IP models.

It was observed that although the best models were in the form of IE, the prediction performance from IP models was more consistent. IE_avg and IE_24h had very large distributions of predictive performance. It is hard to say with certainty whether consistent behaviour from IP models is dataset-independent – if that is the case this would mean that the logarithmic form of IP ($c\log_{10}(R_p)$), in comparison to the hyperbolic form of IE ($c\ R_p^{-1}$) stabilises the model prediction performance, which would make sense for finding edge cases such as good corrosion inhibitors. R_p models had lesser distribution, but their performances were consistently worse. Worst IP and IE_EIS24h models were almost always better than the best R_p models. This underlines the importance of target representation – it seems that normalisation offered by IE and IP transformations allows models to capture corrosion inhibition in a more accurate manner.

Features

The 29 different featurisation methods create a wide span of features in a different manner, consisting of 0D (bulk properties and physicochemical descriptors that contain no information about molecule geometry or atom connectivity, e.g., molecular weight, logP octanol-water partition coefficient, contained atom presence and counts), 1D (representations that include information on bonding or bonding fragments, e.g., presence/absence of molecular fragments, hydrogen bond donor or acceptor counts, number of rings, number of functional groups), 2D (molecule graph invariant properties, e.g., topological polar surface area, autocorrelation descriptors), and 3D (topographical molecule shape information, e.g., geometrical, three-dimensional distances and connectivities) descriptors. The featurisation methods can broadly be split into three different categories: physicochemical descriptors (e.g., `PaDEL` [318]), structural keys (e.g., `MACCS` [319]), and hashed fingerprints (e.g., `atompair-count` [320]).

Physicochemical calculators generate information about the physical and chemical properties of the whole molecule, such as the surface area occupied by polar atoms and their attached hydrogens, the number of electronegative atoms that can act as hydrogen bond donor/acceptor, the molecule Van der Waals radii surface area of its atoms, or even more obscure and derivative properties such as the topological Balaban index that measures the branching and connectivity of a molecular graph, among many others. The combination of these types of features holds promise in highlighting the properties most relevant for the target molecule behaviour we are interested in.

Structural keys encode the molecule structure into a binary bit value (0 or 1) where each bit corresponds to a *pre-defined* structural feature, such as the presence/absence of a benzene ring. If the molecule has the pre-defined feature, the bit position corresponding to this feature is set to 1, otherwise it is set to 0. It is important to realise that structural keys cannot encode structural features not pre-defined in their fragment library.

Hashed fingerprints solve this problem by not requiring a pre-defined fragment library, where all possible molecular fragments smaller than the specified size are converted into numeric values using various algorithms. A data of arbitrary size can then be converted into a fixed size vector using a hash function. The size of this vector is often chosen to be a power of two, default option used being 1024 or 2048. The values of such a vector correspond to the absence/presence of particular molecular fragments, which are denoted as "bits".

One such molecular fragment generation approach would be path-based fingerprints (e.g., Daylight fingerprint based `rdkit_fp` [321]), where branching paths in the molecular graph are analyzed for a given length and hashed in a fixed vector. A different path-based approach would be atom pairs [320], where pairs of atoms with the shortest path connecting them would form substructures to be hashed. Circular fingerprints offer another option, where the circular environments of each atom up to a given radius are used to construct molecular fragments (e.g., extended-connectivity `ECFP`, functional-class `FCFP` fingerprints [322]). Such fingerprints can be encoded in binary to indicate the presence/absence of a given molecular fragment, or can specify the number of occurrences of that molecular fragment by taking integer values for the count-based fingerprinting approach. The flexibility offered by hashing might also cause problems in interpretability, however, since a molecular database may contain a very large amount of molecular fragments and hashing them into a fixed range can result in "bit collisions", where different molecule fragments would be converted into the same hashed bit value.

For the best target IE_EIS24h, the top three featurisation methods were all based on hashed fingerprints: `atompair_fp-count`, `rdkit_fp` and `pattern_fp`. Looking at the mean ranking of the best four targets, `atompair_fp-count`, `rdkit_fp`, `pattern_fp`, `avalon_fp`, `avalon_fp-count`, `layered_fp` hashed fingerprints, and `PaDEL`, `RdKit3D`, `Mordred` physicochemical descriptors resulted in average rankings of less than 10, on average producing more predictive models than other featurisation approaches. Compared to alternatives, `ECFP` and `pharmacophore` featurisation methods commonly used in many drug discovery problems were inferior in describing corrosion inhibition.

The distribution of the ranking for different targets shows that the choice of the featurisation method is heavily dependent on the target. For all IP target representations, however, it was remarkable that avalon_fp consistently resulted in models with lowest CV-RMSE values. The trends of using fingerprints based on presence/absence compared to counts were also dependent on the featurisation method: `atompair` performed better for counts, but `rdkit`, `fcfp` and `avalon` fingerprints performed better in presence/absence binary form. Addition of 3D descriptors to the `RDKit2D` improved ranking consistently, highlighting the importance of 3D molecule effects for corrosion inhibition. Whereas for `pharmacophore` descriptors ranking of 2D seem better but both are quite similar in CV-RMSE values, indicating models do not take advantage of additional 3D descriptors offered by this featurisation method. Given that `pharmacophore` descriptors were created to work with molecule interactions with a specific biological target such as a protein or an enzyme, it is normal that the important 3D features do not directly transfer to other problem domains.

Looking at the best target (IE_EIS24h) CV-RMSE distributions for the featurisation methods, the best model for the best featurisation method atompair_fp-count resulted in a CV-RMSE of 2.73, whereas the best model for the worst featurisation method resulted in a CV-RMSE of 4.06, an increase of 49%. The distribution for every featurisation method was a result of scaling, feature selection, and model architecture.

There was no one best method for choosing any of these details for model configurations, as they all result in similar distributions for CV-RMSE (see Supplementary Figure 3-6). However, by examining the ranking of all models based on CV-RMSE (see Supplementary Excel file), we can qualitatively identify trends among the top-performing models.

Out of the ten models with lowest CV-RMSE, all had IE_EIS24h as target. For featurisation methods two were based on `atompair_fp-count`, six on `rdkit_fp`, one was `pattern_fp` and one on `RdKit3D` descriptors. It is interesting to note that only one model was based on physicochemical descriptors, and the rest were based on hashed fingerprints. Feature scaling showed a mix of methods, with the key takeaway being that any scaling is beneficial compared to none: only one model did not use scaling (which was based on RF architecture, which is a scale-independent model), while the others were evenly split with three models each using minmax, power, and standard scaling.

For feature selection nine out of ten models used RFE, suggesting it may be a better choice for standardised use despite the drawback of requiring manual selection of the number of features beforehand. The rest of the analysis in this paper used RFE-based feature selection to refine the feature set, ultimately selecting down to the top 10 features of every configuration.

For model architectures, seven models used SVR, two RF and one XGB. This indicates that SVR architecture's robustness to outliers and noisy data may be particularly valuable when working with real-world experimental data.

5.3.2. Gaining mechanistic insight through algorithmic feature selection

Having established an optimal model description, we now shift focus to our primary objective: leveraging this refined description to uncover novel mechanistic insights. The premise is that the features that make a model more predictive are also likely to be those most relevant to the underlying physicochemical mechanisms of corrosion inhibition. This makes feature selection methods in machine learning not only a routine for improving model prediction performance, but also a tool for extracting scientific insight hidden in the data statistics.

After identifying key features through algorithmic selection, it is essential to develop an intuition about their relevance to the system. Our goal was to identify whether and how the algorithmically chosen features can be used as a tool to gain mechanistic insight about corrosion inhibition. With that goal in mind we selected one featurisation method from each category for further experimentation: `PaDEL` for physicochemical descriptors, `maccs_fp` for predefined structural keys, and `atompair_fp-count` for hashed fingerprints. The selection was based on the highest predictive performance of each category.

The choice of `MACCS` (Molecular ACCess System) over a potentially more predictive featurisation method is based on two key reasons. First, given the narrow range of IE_EIS24h CV-RMSE distributions, explainability takes precedence over pure predictive performance. While accurately predicting the behaviour of individual molecules may be challenging for such small-scale models, the primary objective is to predict general mechanistic trends based on the selected features, a task that is comparatively more feasible. `MACCS` is ideal for this, as due to its predefined nature it is extremely clear what each feature corresponds to. Second, not every featurisation method allows explainability in the first place. In our preliminary experiments with visualising `rdkit_fp` features, we have observed considerable amount of bit collision — where different molecule substructures are mapped to the same feature column bit. This not only most likely reduces predictive performance, but also complicates explainability, as it becomes unclear which substructure is driving the prediction. Other fingerprints such as Avalon does not allow visualisation in the first place, preventing explainability.

The next sections first analyze the interpretation of hashed `atompair_fp-count` fingerprints through bit visualisation. We show that by making artificial changes in the prediction queries we can gauge the response of a given feature. Since visualisation into substructures is not possible for physicochemical descriptors and other featurisation methods, in the following section we show how to use Bayesian optimisation as a tool for understanding model decision-making process with `PaDEL` featurisation. Finally, we use SHAP analysis for deciphering feature influence for best models based on `atompair-count`, `PaDEL`, and `MACCS`, and we demonstrate how we can use models based on different featurisation methods to gain a united mechanistic insight.

5.3.3. Visualising algorithmically selected fingerprints for finding the corrosion inhibition structural building blocks

Figure 5.2 demonstrates the algorithmically selected features from the best `atompair-count` featurisation visualised as corresponding molecular substructures, here denoted as ’bits’. The substructures are recorded at the time the fingerprints are generated. Later, these substructures are used to map the features back to the corresponding parts of each molecule. As mentioned before, `atompair` featurisation describes the molecular substructures as two atoms and the number of atoms between them. For example the bit corresponding to feature 478 is a sulfur and nitrogen with two atoms in between, and can be written in a SMILES-like format as S-(2x)-n, where x corresponds to any atom, and uppercase/lowercase denotes aliphaticity/aromaticity. In the same manner, bit 479 would be S-(2X)-C, bit 576 S=C, and so on.

Based on the selected features we can see that the model "thinks" that substructures involving triplets of aromatic (bit 200, bit 311) and aliphatic carbons (bit 1295), nitrogen-nitrogen couples (bit 1016), sulfur directly attached to carbon (bit 576), sulfur attached to nitrogen with two atoms in between (bit 478), sulfur attached to carbon with two atoms in between (bit 479), and aliphatic carbon attached to aromatic carbon with three atoms in between (bit 453) are important in predicting corrosion inhibition. Feature selection identifies these substructures as corrosion inhibition-critical substructures.

These substructures align with the mainstream literature conclusions, which highlight that sulfur and nitrogen atoms often serve as anchoring points to the surface. Additionally, aromatic groups may contribute not only through steric effects that help repel detrimental chloride ions but also through their electron-donating or -withdrawing properties, which can influence direct interactions with the metallic surface or modulate electron density redistribution within attached functional groups. Meanwhile, long aliphatic chains further enhance steric hindrance, both factors being critical for effective corrosion inhibition [1]. Especially the fact that sulfur, critical for inhibition of copper intermetallics of AA2024-T3 [105, 106], was identified as important solely by the model, with no prior domain expertise or previous scientific insight, shows that models can cap-

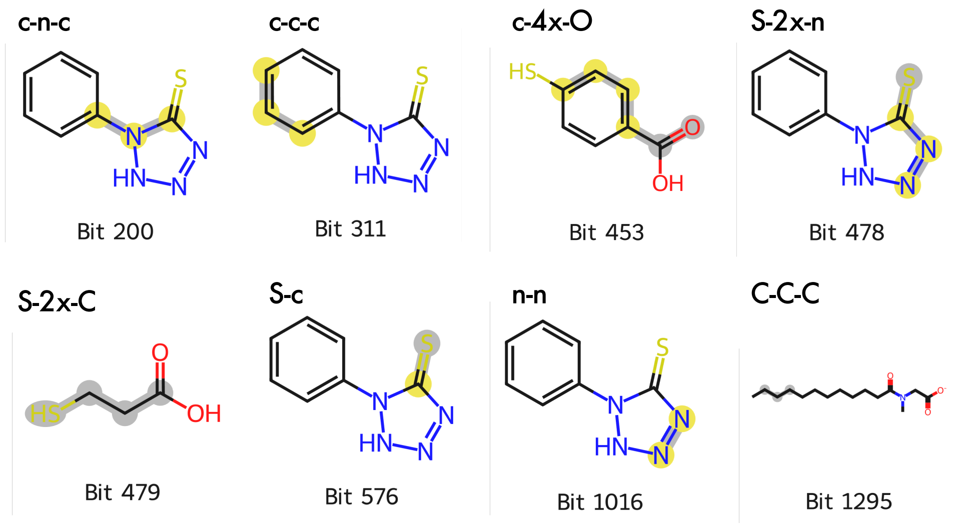


Figure 5.2: Example molecules that contain key features identified by the feature selection algorithm, visualised as 'bits'. The SMILES-like strings corresponding to the bit are presented in top left, and the corresponding structures are highlighted in yellow and gray on the molecules. Aliphatic atoms are highlighted in gray and in uppercase letters, aromatic atoms in yellow and in lowercase letters.

ture physicochemical insights. This shows promise in reverse engineering statistics into mechanisms, and for that understanding the model decision-making process is key.

The visualised molecule substructures already give mechanistic tips, but to understand how every feature contributes to the model in a detailed manner, and to explore what would've happened if only that particular feature had a different value, we have produced counterfactual predictions. To form counterfactuals, we have kept every other feature constant while changing only the analyzed feature to its maximum or minimum value found in the dataset, and then examined how predictions of the model changed.

Figure 5.3 presents examples from the counterfactual predictions for top-performing molecules for the selected features and their corresponding visualised bits. For a counterfactual prediction, the feature to be analyzed is modified to its maximum and minimum value found in the dataset, while keeping all the other features at their original values. In this way, the effect of every feature on a set of given molecules can be analyzed independently from other features.

Here we present the influence of feature 478 vs. 479, and feature 576 vs. 1016 as they capture interesting interpretable trends. Feature 576 is directly related to the surface bonding opportunity offered by the sulfur atom, where the predicted efficiency increases with an increase in number of thione bonds. The only cases where the maximum does not correspond to an increase are the two derivatives of 1,2,4 triazoles. This makes sense structurally, as the maximum sulfur amount found in the dataset is four, which in the case of the smaller five-ring structures might hinder bonding instead of supporting. Therefore, depending on the ring size, excess sulfur not contributing to bonding might not be beneficial for inhibition.

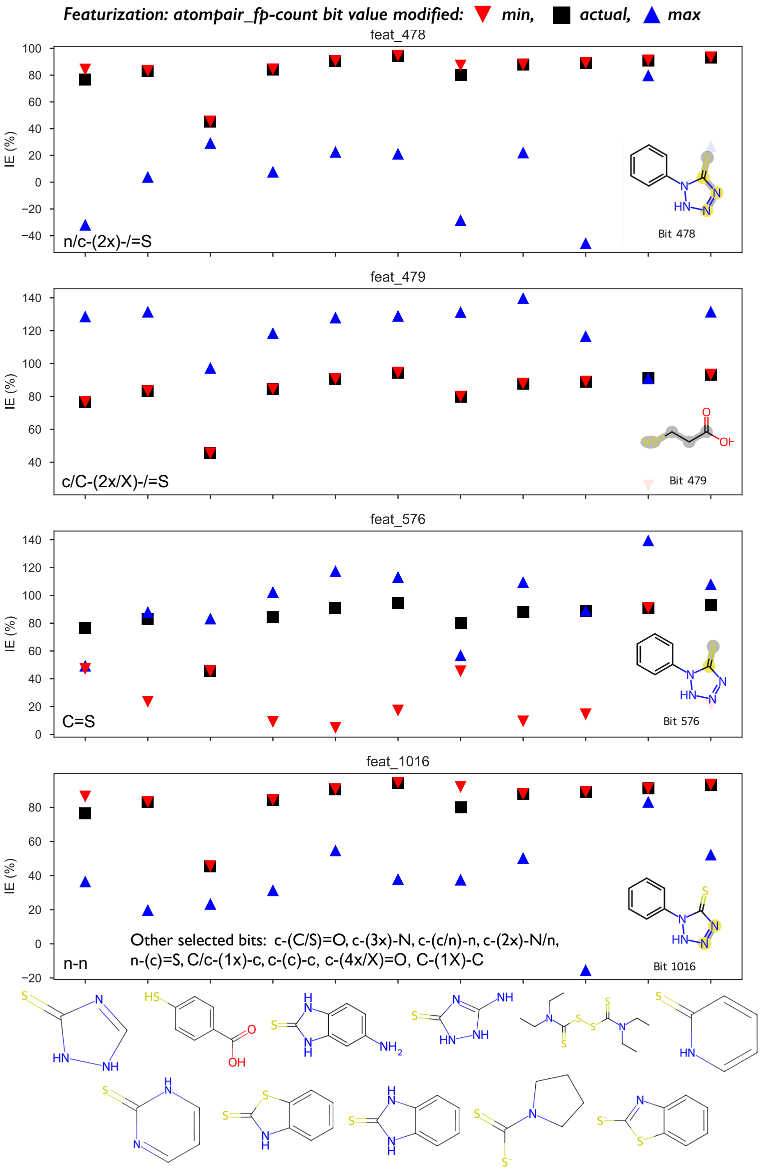


Figure 5.3: Creating counterfactual predictions for molecules with highest target values in the experimental dataset. The x-axis shows the molecules, and the y-axis shows the predictions. For actual predictions, the `atompair-count` model with lowest CV-RMSE is used for predictions, for min/max only the value for the corresponding feature is changed to the min/max found in the original featurisation dataset, and then the same model is used for predictions. Visualisation of the features as bits on example molecules is shown at the bottom right of the plots, and corresponding SMILES-like strings are shown on bottom left.

Feature 478 and 479 are also connected to the sulfur behaviour. Feature 478 and 479 correspond to very similar substructure bits with the only difference being the end atom: bit 478 is a sulfur atom connected to a nitrogen atom with two atoms in between (S-2x-n), bit 479 is a sulfur atom connected to a carbon atom with two atoms in between (S-2x-c). Despite their similar structures, an increase of feat 478 resulted in a decrease of predicted inhibition efficiency for all molecules, whereas an increase of feat 479 on the contrary increased predictions.

Clearly there is something important about this bond distance to be present in 20% of the features. In cyclic structures made up of five or six atoms, in either case where sulfur is in the ring or attached as a branching functional group, this sulfur-nitrogen distance would put sulfur and nitrogen on opposite sites of the ring. If the molecule benefits from sulfur and nitrogen being close to one another for corrosion inhibition – such as the formation of bidentate chelates – this position would prevent nitrogen from working together with sulfur in bonding, and highly electronegative nitrogen would draw excess electrons necessary for bonding away from the sulfur donation centers. In combination with trends of feature 1016, where maximum nitrogen-nitrogen pairs cause a decrease in predicted corrosion efficiency, it is clear that the position of nitrogen is very important for maximising corrosion inhibition performance.

This can also be used as a design principle: (i) as the atompair distance at a topological distance of 4 between C-S increases relative to N-S, and (ii) as the presence of neighbouring nitrogen atoms that do not contribute to surface binding decreases, the corrosion inhibition performance increases.

From these results, we argue that the potential of counterfactuals for gaining mechanistic insight is promising. Before diving deeper into mechanistic insights, we would like to also demonstrate a way of analyzing the physicochemical features, and afterwards combine both with a more complete analysis in the section on SHAP analysis.

5.3.4. Bayesian Optimisation as a tool for understanding model decision-making process

Bayesian optimisation is a statistical method for optimising any black-box objective function that lacks an analytical form and is expensive to evaluate. Instead of the true objective function, Bayesian optimisation uses a surrogate model that is an approximation of the objective function. This cheaper-to-analyze alternative is used to extrapolate the function with a measure of uncertainty. An acquisition function is used to select the next point to sample. This selection can be a combination of exploration (searching areas of the n-dimensional search space where the surrogate model is uncertain, active learning) and exploitation (searching areas where the surrogate model predicts high objective function values, Bayesian optimisation). In a sequential manner, the objective function is evaluated at the selected point, the surrogate is updated based on the gained information, and the acquisition function decides on the next point to be evaluated. This process is repeated iteratively, while with every step the global optimum of the surrogate model converges towards the global optimum of the objective function.

Bayesian optimisation can also assist in illuminating the black-box function of corrosion inhibition. The advantage of using physicochemical descriptors as model features is the clarity of the features – every feature is defined clearly, whether it is heteroatom content, ring number, electronegativity or any other interesting physicochemical quality. However, the disadvantage is often that the calculated descriptors are quite arcane, therefore it is difficult to have an intuitive understanding of what kind of molecule structure would result in the quantitative value of a descriptor. This can potentially make interpretation difficult. The reversed problem is even more difficult: given multiple such features, a molecular chemist or a materials scientist would have a hard time converting these quantitative parameters into an actual molecule. Bayesian optimisation offers a way out of this thorny reverse-design problem.

In our case, the black-box function to be optimised was the best model based on PaDEL featurisation. We were looking for the selected feature values that would result in the highest inhibition efficiency. After initialising the optimisation with samples in our dataset combined with 2000 samples with randomised features to be analyzed, Bayesian optimisation was run for 1000 iterations. The features that resulted in the optimised maximum were the optimal molecule parameters that the model predicts will lead to the best inhibition efficiency. We call this artificial creation an optimised "pseudomolecule".

This ideal 'pseudomolecule' can be used as a template for finding real molecules that are similar. We can compare the pseudomolecule with molecules from any given database, and find the molecules most similar to it. For this, we used a previously published toxicity database [315] that contains over 10,000 molecules. The choice of selecting a toxicity database was deliberate – aside from predictions from our model, the database would also provide information on the toxicity of compounds. We calculated similarity between our candidate pseudomolecule and the molecules in the toxicity database with the cosine similarity metric. The most similar 20 molecules are presented in Figure 5.4 in descending order in similarity. The EPA toxicity classifications of the molecules are also presented with the molecules, which correspond to: 1 highly toxic, 2 moderately toxic, 3 slightly toxic, and 4 practically non-toxic [316].

What we see from this figure is a complementary picture to the results from the previous model with atompair-count featurisation. The most similar real molecule in the dataset to our pseudomolecule is 2-mercaptobenzimidazole, which is a known, good corrosion inhibitor and already present in our training dataset (measured IE 94.6%). Other molecules are not present in our dataset. One common theme across all molecules was the presence of bulky benzene groups. Most benzene groups were without any heteroatoms, but some molecules had pyridine, pyrimidine, or dioxolane rings. Connected to the benzene groups, most of the molecules had hydrocarbon chains with secondary/tertiary amines, ketone or carboxylic acid functional groups. Some molecules had sulfur, always in close presence to nitrogen. Similar to the `atompair-count` case, S-N at four-atom distance was not present in the molecules.

This type of analysis is particularly useful for gaining a mechanistic understanding with greater confidence when working with limited data. One key advantage is its ability to serve as a tool for interpolation rather than extrapolation. Since optimisation focuses on feature value boundaries already present in the dataset, the model operates within familiar territory, enabling more reliable interpolation within those boundaries. The addition of toxicity values also gives the choice of selecting non-toxic molecules for further experimentation.

In the next section, we show how the combination of this sort of reverse engineering

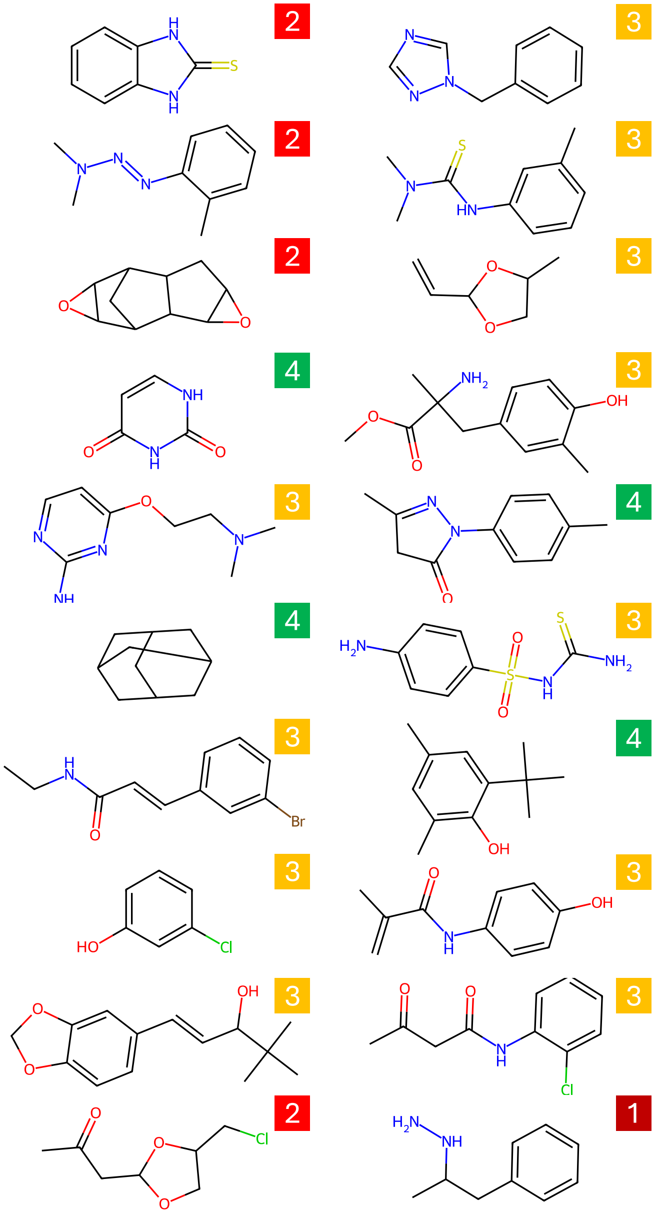


Figure 5.4: Molecules most similar to the pseudomolecule for the similarity metric cosine similarity. EPA toxicity classification shown next to the molecules: 1 highly toxic, 2 moderately toxic, 3 slightly toxic, and 4 practically non-toxic. Molecular featurisation was done with PaDEL, and the dataset used was a toxicity database curated from previously published work [315].

and SHAP analysis helps to understand what the 'perfect' molecule for corrosion inhibition would be according to statistical models.

5.3.5. SHAP analysis for deciphering feature influence

SHAP (SHapley Additive exPlanations) analysis is a method originating from game theory [303]. The SHAP value in the context of a machine learning model is the expected individual contribution of a feature to the model prediction. The SHAP value for any given feature i is calculated as:

$$\varphi_i(v) = \sum_{C \subseteq N-i} \frac{|C|!\,(n-|C|-1)!}{n!} \{v(C \cup \{i\}) - v(C)\}, \tag{5.6}$$

where v is a characteristic function that maps every coalition of n features to a prediction. Here, v is the machine learning model, and C, is such a coalition – a group of features working together. |C| is the number of features in coalition C. |C|! is the number of ways coalition C can form. $(n-|C|-1)!$ is the number of ways the rest of the features can join to the coalition after feature i joins. n! gives the number of ways to form a coalition from n features. The resulting term $\frac{|C|!\,(n-|C|-1)!}{n!}$ is the weight for marginal contribution, or the probability of feature i making a contribution to coalition C. The term $v(C \cup \{i\}) - v(C)$ is the marginal contribution of feature i to the coalition C. All the marginal contributions of a feature with their probability of making those contributions are weighed with respect to the weights for marginal contribution, then summed over all coalitions that feature can make a marginal contribution to. This gives the expected marginal contribution, in other terms, the SHAP value. In this way, all the possible coalitions that a feature can contribute to are considered, and a feature's individual contribution as well as the interactions between features are evaluated.

Figure 5.5 presents SHAP beeswarm plots for (a) atompair-count, (b) PaDEL and (c) MACCS featurisation methods. The beeswarm plot represents the distribution of feature impact of SHAP values across a dataset. Each point in the plot represents the SHAP value of a feature for a specific instance, with colour indicating feature value. Beeswarm plots can identify which features influence the model's predictions the most, through the direction (positive or negative) and magnitude of these influences across the dataset. Positive SHAP value contributions mean that the value of that feature is expected to increase the corrosion inhibition efficiency of a given instance, and vice versa. This helps in interpreting the model, uncovering feature importance, and detecting patterns in the predictions.

The importance of pH has been identified in our previous work [80], and it is also present as a feature in all of the presented model featurisations. A detailed analysis will not be repeated here, however, we would like to highlight that pH was always chosen as one of the most important features of the models, despite having no linear correlation with targets. Based on SHAP dependency plots in our previous works and the beeswarm plots here, we observe that models have very negative SHAP values for very large and very small feature values. The reason becomes clear when one observes the Pourbaix diagram of Al, where aluminium oxide is stable and protective in the pH range of 4 to 9, but starts to disintegrate below and above this range. The models capture that behaviour

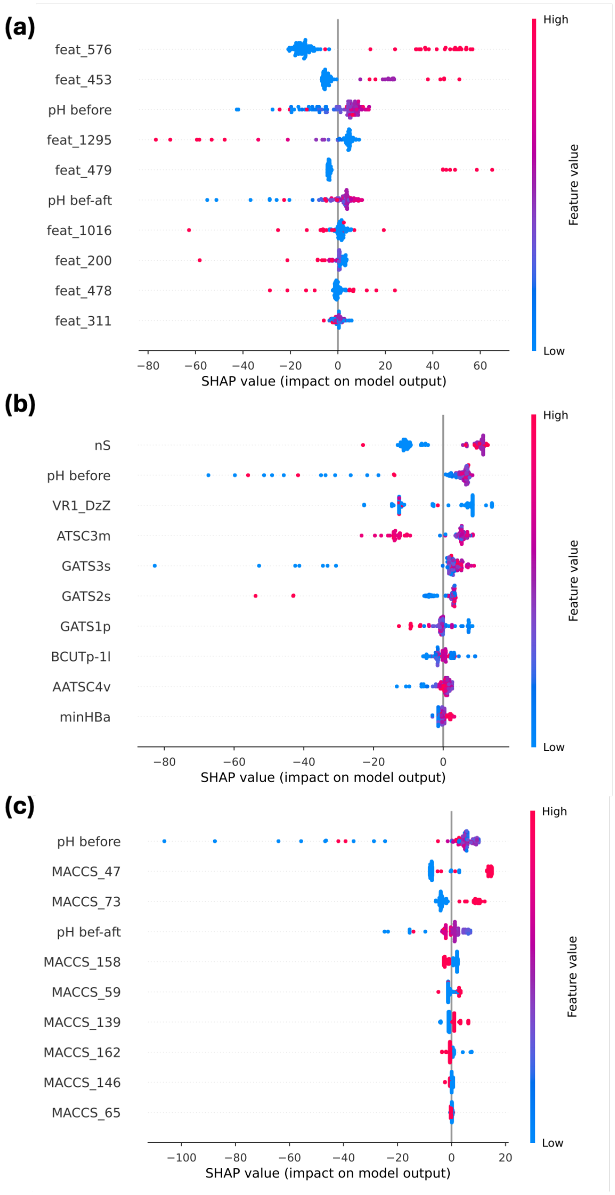


Figure 5.5: SHAP beeswarm plots displaying how features in a dataset impact model output for featurisation (a) atom pair-count, (b) PaDEL, (c) MACCS. Each dot represents an individual model instance (molecule), which pile up along each feature row to show density. Each row corresponds to one feature, which are sorted by the mean of absolute SHAP values. Colour is used to display the original value of a feature, whereas the SHAP value is the impact of a given feature value on the model output. Large values correspond to larger expected model impact.

quite well, proving that given the right features, mechanistic behaviour resulting from the environment-substrate interactions can be captured.

ATOMPAIR-COUNT FINGERPRINTS

Figure 5.5 (a) shows SHAP beeswarm plots for `atompair-count` fingerprints. The visualisation of the molecule substructures as bits as seen in the previous section allows us to explain the feature SHAP behaviour. The notation used in this section is used as before: uppercase for aliphatic, lowercase for aromatic atoms, / for denoting structures corresponding to multiple atoms or bonds and nx for the number n of any atoms in between.

Feature 576 corresponds to substructure bit c/C=S, a carbon-sulfur double bond. The presence of sulfur is expected to increase the inhibition efficiency predictions, and it has the biggest impact on model predictions. This is in line with trends seen from literature which mention the high tendency of sulfur to bond with copper [1, 323], which would also allow the organic molecules to bond with Cu-based intermetallics of the AA2024-T3 substrate, which are root cause for localised corrosion [39, 42, 55]. If the intermetallics are protected, it would greatly decrease the microgalvanic driving forces that cause localised corrosion of the alloy. For this reason, it is not surprising that sulfur presence is the most important feature, but it is nonetheless remarkable that the model has learned the importance of bonding with such a clear tendency. Analysis of other model SHAP values shows that this is not a coincidence.

Feature 453 corresponds to substructure bit c/C-4x=O. High values have a positive impact on the model, whereas low values expect to have a minor negative one. Analysis of molecules that contain this bit reveal that the majority had a ccccC=O substructure. That corresponds to an aromatic ring with alcohol, ketone or carboxylic acid functional groups. Assuming that they are not near the substrate anchoring sulfur/nitrogen groups, such structures would indeed push away the corrosive Cl^- ions through steric hindrance.

Feature 1295 corresponds to substructure bit C-X-C. Higher feature values result in a sharp decrease of SHAP values, low values of it result in minor positive values. The majority of molecules containing this feature include CCC and CNC substructures, which are characteristic of aliphatic hydrocarbon chains. These chains exhibit low reactivity, limiting their interaction with both the surface and the surrounding environment. The prevalence of these groups is a characteristic quality of surfactants, where a long aliphatic tail is typically attached to a carboxylic or amino group. Such surfactants are recognised as effective inhibitors under specific conditions for substrates like carbon steels [324, 325]. In these cases, the presence of long tails likely contributes to corrosion inhibition by providing steric hindrance. However this was not the case for this alloy system. An excessive presence of such chains can lead to a bulky structure with decreased molecule solubility without contributing to surface bonding, which must have been the dominant negative effect.

Feature 479 corresponds to substructure bit c/C-2x-S, sulfur bonded to any two atoms bonded to a carbon atom. Whereas feature 576 contained information on double-bonded sulfur, this one contains information on single-bonded sulfur. Analogous to feature 576, higher values correspond to significantly increased SHAP values. It seems that sulfur with this topological distance to carbon is predicted to contribute significantly to corrosion inhibition. As previously discussed through counterfactual analysis, this sulfur-carbon distance is notably peculiar. For cyclic structures composed of five- or six-membered rings, where sulfur is incorporated within the ring or attached as a functional group, this distance positions sulfur and carbon farther apart, on the opposite sides of the ring. This means that this feature value can be maximised through dithiocarbamate-like structures, or S attached to ring structures. This observation suggests a structural configuration in which sulfur is either bonded as a functional group to one of the ring's vertices or directly integrated into the ring structure.

Trends for the rest of the features are less straightforward to analyze. *Feature 1016* corresponds to n-n, aromatic nitrogen connected together. It seems that a high n-n presence results in more activity, expected to push the predictions more to higher and lower values. *Feature 200* corresponds to c-x-c/C, where higher values decrease the SHAP values, which might be related to aromaticity degree. *Feature 478* corresponds to n-2x-/=S, where the presence of it is making a molecule take more extreme SHAP values. *Feature 311* corresponds to ccc, which again is related to the aromaticity degree and influence on the model is low.

Based on these observations it seems that the correct combination of c/C=S, c/C-4x=O, c/C-2x-S and n-2x-/=S might result in ideal model predictions.

PADEL DESCRIPTORS

Figure 5.5 (b) shows SHAP beeswarm plots for PaDEL descriptors. The descriptions of the computationally generated descriptors are quite often not adequately documented, which requires double-checking multiple sources. Analysis of the descriptors below are primarily based on the book *Molecular descriptors for chemoinformatics* [326], and the documentation pages of numerous descriptor calculator packages.

nS represents the number of sulfur atoms in the molecule. An increase in number of sulfur is expected to increase model predictions. It is the descriptor with the highest impact, and presence or absence of sulfur is predicted to be critical in the inhibition property of the molecule. It is directly related to feature 576 and 479 of the atompaircount fingerprints.

ATSC stands for Centered Autocorrelation of a Topological Structure (also known as Moreau-Broto autocorrelation). Autocorrelation descriptors calculate the correlation between a specific atomic property, such as atomic mass, at a defined topological distance within the molecule. They capture how a property is distributed across the molecular structure. The property values are "centered" by subtracting the mean property value across the molecule.

ATSC3m reflects how the atomic mass is distributed and correlated across atoms that are three bonds apart in a molecule. *3* refers to the "lag", which indicates the topological distance between atoms being considered in the molecule, in this case, three bonds. *m* denotes that the descriptor is weighted by atomic mass. A higher ATSC3m value suggests significant variation in atomic masses at this specific distance, indicating that heavier and lighter atoms are more differently positioned in relation to each other. A lower value indicates that there is little variation in the atomic masses of atoms that are three bonds apart. Since high ATSC3m values are expected to result in a significant drop in the majority of prediction values, neighbours that are two atoms apart and similar to one another in atomic mass could be more suitable for inhibitor molecule structures.

AATSC4v quantifies the autocorrelation of atomic van der Waals volumes within a molecule at a topological distance of four bonds, further normalised with respect to molecule size before calculating the autocorrelation. Higher values did not markedly improve prediction performance, but lower values certainly hindered it. This can be observed for straight-chain structures larger than butane. However, for higher values, individual or fused ring systems exhibit greater topological distances and hence greater potential. Examples include carbon atoms in aromatic rings such as benzene, benzimidazole, benzotriazole, or cyclopentanes. Additionally, the presence of two neighbouring heteroatoms, such as nitrogen in five-membered rings like imidazoles, can also contribute to these increased distances.

The behaviour of *VR1_DzZ* was difficult to analyze. Dz are a modification of ATSC descriptors that use topological distance in conjunction with properties of atoms (see Dz^K pg.33 [326]). Official PaDEL documentation describes VR1_DzZ as a "Randic-like indices eigenvector-based index from Barysz matrix / weighted by atomic number" (see Randic-like pg.164 , VRA1 pg.717 [326], calculation of Barysz distance matrix [327]), which is defined by coefficients of the eigenvector associated with the largest negative eigenvalue. It is related to local vertex invariants able to provide discrimination among graph vertices. However, its non-linear complicated effect is difficult to analyze in isolation, where high values seem to hinder the inhibition efficiency, therefore it is not further discussed as tying it to the molecular structure is not accessible.

GATS stands for Geary Autocorrelation of Topological Structure. Like ATSC, GATS descriptors are used to quantify the autocorrelation of a specific atomic property over a defined topological distance in a molecule. GATS differs from other types of autocorrelation by including a normalisation factor, which adjusts for the number of atoms and bonds considered, providing a scale-independent measure. A strong positive correlation produces low GATS values between 0 and 1, negative autocorrelation produces values larger than 1, whereas no correlation corresponds to a value of 1 (pg.32 [326]).

GATS3s provides a measure of how the Sanderson electronegativity varies across the molecule at a topological distance of three bonds. *3* again refers to the lag. *s* denotes that the descriptor is weighted by atomic Sanderson electronegativity, which is a specific measure of electronegativity that describes the ability of an atom to attract electrons in a chemical bond. High GATS3s values indicate a significant variation in electronegativity values among atoms that are three bonds apart. This might occur in molecules with a mix of atoms that have widely differing electronegativities, as for heteroatoms (e.g., sulfur, nitrogen, oxygen) in an organic molecule. Low GATS3s values suggest uniformity in electronegativity at this distance. Similar electronegativities would indicate less variation in the ability to attract electrons across the molecule. The SHAP values seem to increase with increasing GATS3s values, suggesting that for ideal inhibitors atoms at 3-bond distance should have a higher electronegativity difference.

GATS2s is similar to GATS3s with the only difference being the topological distance, which is 2 in this case. This suggests molecules with atoms at 2-bond distance with differing electronegativities would result in higher target values. Unlike GATS3s, high GATS2s values decrease the model performance for two outliers.

GATS1p is also similar to GATS3, but in this case the topological distance is 1, so it considers neighbouring atoms. *p* indicates that the descriptor is weighted by atomic polarisability. GATS1p therefore is a measure of how the property of atomic polarisability varies over the structure of the molecule for neighbouring atoms. Polarisability is a measure of how easily the electron cloud around an atom can be distorted by an electric field, which is related to the size of the atom and its electron density distribution. Lower GATS1p values seem to increase the model predictions. A low GATS1p value suggests that the atomic polarisability of adjacent atoms are quite similar. This would occur in molecules where atoms have similar sizes and electronic environments, leading to little variation in how easily their electron clouds can be distorted.

The *BCUTp-1l* descriptor reflects the distribution of polarisable atoms in a molecule. *BCUT* stands for Burden - CAS - University of Texas eigenvalues. It refers to a set of molecular descriptors derived from the Burden matrix, a matrix which captures a desired property correlation between every atom in a molecule. *p* indicates that the descriptor is weighted by atomic polarisability. *1l* signifies the lowest eigenvalue obtained from the Burden matrix. A low eigenvalue typically indicates that the molecule's polarisability is relatively evenly distributed or that there are no extreme variations in polarisability across the molecule. Conversely, a higher eigenvalue suggests more significant variations, possibly indicating regions of the molecule with high and low polarisability. In the case for this model its effect was not straightforward to analyze, but it was observed that higher values corresponded to more limited absolute impact, suggesting less active molecules, which may not be desirable for inhibitor molecule design.

minHBa refers to the calculated minimum hydrogen bond acceptor strength in a molecule. A hydrogen bond is the electrostatic attraction between a hydrogen atom covalently bonded to a more electronegative atom or group, the "donor", and another electronegative atom that has a lone pair of electrons, the "acceptor". Main hydrogen bond donors and acceptors are electronegative atoms like N and O, which have lone pairs of electrons that can attract the hydrogen atom. The minHBa descriptor specifically focuses on identifying the weakest hydrogen bond acceptor within the molecule – a high minHBa value would mean that even the least effective hydrogen bond acceptor in the molecule has a relatively high hydrogen bonding potential. A high minHBa seemed to increase the prediction values. This suggests that C atoms with higher hydrogen bond acceptor values would assist in improving predictions. This might be related to the aromaticity: as it was found that aromatic rings act as hydrogen bond acceptors [328], therefore compared to aliphatic C chains the presence of aromatic rings might increase the minHBa values. A high hydrogen bonding capacity would help in the self-assembly process by creating more intact monolayers as the organic molecules adsorb to the surface with one part, and attach with one another through hydrogen bonding. [329, 330] A tighter bonding between adsorbed molecules would hinder chlorides from penetrating in between. However, the influence of the descriptor on the model is weaker than the rest of the features.

Summarising the strongest interpretable influences that would result in a higher predicted inhibition efficiency for AA2024-T3 alloy:

1. High number of sulfur atoms.

   The molecule likely contains multiple sulfur atoms. Sulfur is relatively electronegative (though less so than oxygen and nitrogen) and can participate in various chemical environments, such as thiols (-SH), thioethers (R-S-R'), or disulfides (R-S-S-R').

2. High GATS3s and GATS2s: high variation in electronegativity at a three- and two-bond distance.

   This suggests that at a distance of three- and two-bonds, there is a significant difference in the Sanderson electronegativity values. This could mean that there are alternating patterns of atoms with high and low electronegativities. The presence of highly electronegative heteroatoms such as sulfur, nitrogen and oxygen, combined with less electronegative atoms such as carbon at three- and two-bonds distance, would result in higher descriptor values.

3. Low ATSC3m: low variation in atomic mass at a three-bond distance.

   The low ATSC3m value indicates minimal variation in atomic mass at a distance of three-bonds. This implies that the atoms in the molecule, despite the different types, have similar masses. Since sulfur has a relatively high atomic mass compared to carbon, oxygen, nitrogen, or hydrogen, this would result from structures with only a small amount of sulfur at the periphery of the structure, leading to a more uniform mass distribution.

4. Low GATS1p: low variation in polarisability at a one-bond distance.

   The low GATS1p value indicates uniformity in atomic polarisability for neighbouring atoms. This suggests that the atoms connected directly to each other do not vary much in their polarisability, which could be the case if they are similar types of atoms or atoms with similar electronic environments.

Molecular structures corresponding to such trends would include several sulfur atoms, either in a linear arrangement or part of cyclic structures. Alternating electronegativities of N/O with carbon at two/three-bonds distance (structures of -S/N/O-X-C- / -S/N/O-X-X-C-) would result in high GATS2s/GATS3s. Sanderson electronegativities increase in order of C<S<N<O (2.75<2.96<3.19<3.65, in Pauling units), therefore N/O coupled with C would contribute more to the increase in GATS2s and GATS3s. This also coincides with feature 478 and 479 bits c/C-2x-S and n-2x-/=S from atompair-count fingerprints. Despite the presence of heavy sulfur, the molecule's structure would need to consist mainly of atoms of similar mass, meaning structures with long carbon chains or multiple cyclic structures are necessary to give rise to low ATSC3m. In addition, adjacent atoms should have similar polarisability values, indicating a lack of highly polarisable atoms directly bonded to less polarisable ones, again pointing towards long carbon chains or cyclic structures. Derivatives of larger thiols (R-SH) and thioethers (R_1-S-R_2) as well as cyclic thiophene and benzothiophene-like molecules would satisfy such criteria.

MACCS keys

Figure 5.5 (c) shows SHAP beeswarm plots for MACCS keys. Despite lower prediction performance MACCS was added to the previously studied featurisation methods because MACCS features are completely predetermined and very interpretable. The

MACCS keys were interpreted based on the Mayachemtools MACCS keys documentation [331].

The *MACCS 47* key corresponds to the S-/=x/X-n/N substructure. The presence of such a substructure is expected to increase the inhibition efficiency predictions for almost all molecules. This matches with the GATS2s requirements from PaDEL featurisation, as the S-X-N structure would have a higher electronegativity difference at two-bond distance.

The *MACCS 73* key corresponds to the S=X substructure, which matches the bit 576 from atompair-count featurisation. The presence of a double bond with sulfur, along with the S-X-N substructure, has the largest impact on higher inhibition value predictions. The rest of the features influence the model significantly less. As discussed before, sulfur presence is critical for inhibition, and multiple models consistently using related features underline this.

The *MACCS 158* key corresponds to the C-N substructure. Nitrogen presence is expected to decrease the model predictions. This was counterintuitive, as S-X-N presence was expected to increase the predictions. One explanation might be that in the presence of sulfur the model overshoots the predictions, and this feature decreases it to the expected values. This was actually observed for counterfactual predictions (Figure 5.3), where when feature 576 (corresponding to C=S) is artificially replaced with the maximum values found in the dataset, the predictions went above the theoretical maximum of 100% for the majority of well-performing molecules. Meanwhile, if there's no sulfur present, the molecules are often just not expected to work as corrosion inhibitors for the selected AA2024-T3 substrate.

The *MACCS 59* key corresponds to the S-x-x substructure (sulfur bonded to any atom with a non-aromatic bond, whereas that atom is bonded to another with an aromatic bond). For structures where sulfur is bonded to an aromatic ring structure, its presence often can be correlated with an increase in the model predictions, although the underlying relationship seems to be complex.

The *MACCS 139* key corresponds to the OH substructure. This would be present in carboxylic acids and alcohols, and its presence is expected to increase the model predictions. MACCS 139 shows similarity to feature 453 from atompair-count featurisation.

The *MACCS 162* key corresponds to presence of an aromatic substructure. Its presence is expected to slightly decrease the model predictions. This could be working with MACCS 59, where the aromaticity effect in combination with sulfur presence determines the complete effect of the ring structures.

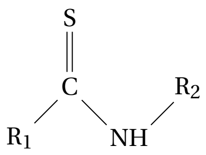
The *MACCS 146* key corresponds to condition where O < 2. This would act as a carboxylic acid detector, as one acid group would need at least two oxygen atoms. When this condition is true, and there are no acid groups on the molecule, the predictions of the model are expected to decrease. In combination with MACCS 139 this would determine the influence of single carboxylic acid functional groups.

The *MACCS 65* key corresponds to the c-n substructure. Its influence on the model is very weak and mixed.

Taken together, the combined information of all keys suggest that molecules containing S=C/C–N substructures, coupled with carboxylic acid groups and limited C–N bonding, may exhibit strong corrosion inhibition potential.

The common trends

Combining the insights from all three featurisations, we deduce that presence of S=C would improve the model predictions. S=C can form in sulfur analogues of carbonyl and carboxyl group thiocarbonyl and dithiocarboxyl groups, and would act as the anchor binding the molecule to the substrate. This substructure would ideally have N as its neighbour to C, which seems to have a positive influence on the inhibition. This might potentially be a result of N assisting with the bonding through the S, or through its electronegativity stabilise the hydrogen bonding formed between the molecules during the self-assembly process. This gives us a molecular structure template for an ideal corrosion inhibitor:



where R_1 can be S for dithiocarboxylic acids, or a longer chain that starts with S for dithiocarbamate structures. R_1 and R_2 can contain and/or be merged together into single or fused ring structures. This in combination with carboxyl presence would fulfill the criteria from different featurisations.

These patterns are found in structure of the commonly used corrosion inhibitors such as 2-mercaptopyrimidine, ammonium pyrrolidinedithiocarbamate, and 3-amino-1,2-4-triazole-5-thiol [80]. Literature suggests that S and N heteroatom containing organic molecules can stabilize AA2024-T3 aluminium oxide by covering the surface through sulfatization, or can adsorb on the copper-rich intermetallics, suppressing the cathodic reactions which often is the driving force of corrosion in the surrounding area [205]. Notably, even without any expert-guided feature selection, through obserbing the trends hidden in the dataset statistics alone, the results of this methodology corroborate the previous spectrostroscopy results that aimed to uncover mechanisms responsible for structures responsible for corrosion inhibition of various organic molecules [1].

The C(=S)N (and also c(=S)N, C(=S)n, c(=S)n for aromaticity variants) SMILES string can be converted into a SMARTS pattern, with which molecule databases can searched for this substructure. For the toxicity database we have used in this study (which contains more than 10,000 molecules), this search ends up in 123 hits of this database, which can be used as lead candidates for exploring potential, yet untested, corrosion inhibitors. These resulting lead molecules can be further constrained to include the trend coming from atompair-count featurisation of feature 453 c/C-4x=O, where its presence is expected to increase model predictions. The presence of c/C-4x=O molecular fragment further decreases the lead molecules to 10.

Among these, the molecules with EPA classification 3 and 4 are displayed in figure 5.6: 5-ethyl-5-(1-(ethylthio)ethyl)-2-thiobarbituric acid, sulfocarbathione,2-thiobarbituric acid and 5-methyl-2-thiohydantoin.

Out of the displayed four molecules, only 2-thiobarbituric acid (figure 5.6 lower-left) was available to purchase off the shelf. To show the validity of our gained insight, we have

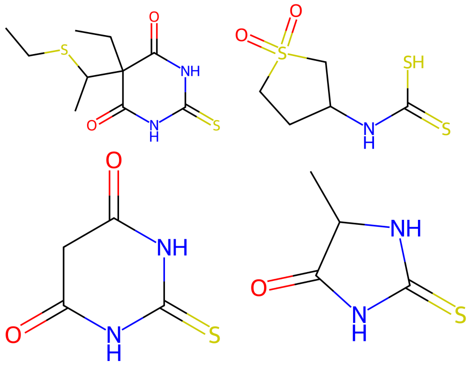


Figure 5.6: Non-toxic molecules from the toxicity database that fit the trends observed from different featurisation methods.

conducted electrochemical experiments using the same methodology used to acquire previous targets to curate the original training dataset [80]. We have tested 1 mM 2-thiobarbituric acid and adjusted its pH to 7.0, as the original solution had a pH of 2.3, much lower than the thermodynamic stability window of Al_2O_3, which is between 4 to 8.5 [145].

Figure 5.7 displays the results of the electrochemical measurements. Electrochemical impedance measurements performed after 24 hours of electrolyte exposure show that 2-thiobarbituric acid is indeed a promising molecule for corrosion inhibition. A comparison of the diameters of the suppressed semicircles shown in the Nyquist plot of figure 5.7 (a) show that the addition of thiobarbituric acid enlarges the diameter significantly, which is related to an increase in the polarisation resistance and overall corrosion inhibition of the surface. Bode plots of figure 5.7 (b) demonstrate that the addition of thiobarbituric acid increased the impedance modulus values measured at 10^{-2} Hz, which represents the corrosion resistance of the inhibitor-surface interface [143]. Impedance modulus values were raised to 64.2±14.5 kOhm cm^2 in the presence of thiobarbituric acid, which corresponded to an inhibition efficiency of 84.1± 3.5%. An interesting observation consistent throughout samples was that while the open circuit potential values were constant around -520 mV vs. Ag||AgCl throughout the first 24 hours (similar to initial uninhibited values), the linear polarisation resistance values kept increasing throughout time, without showing any signs of slowing down. These observations indicate that thiobarbituric acid can work as a strong corrosion inhibitor, and other corrosion inhibitor candidate molecules presented at Figure 5.6 should also be tested for their potential.

5.4. Conclusions

This paper demonstrates that mechanistic insights can be derived from machine learning models to design novel functional molecules. Rather than focusing on predicting individual molecule performance from small datasets—a task that is inherently limited by dataset size—we can reverse-engineer statistical models to understand

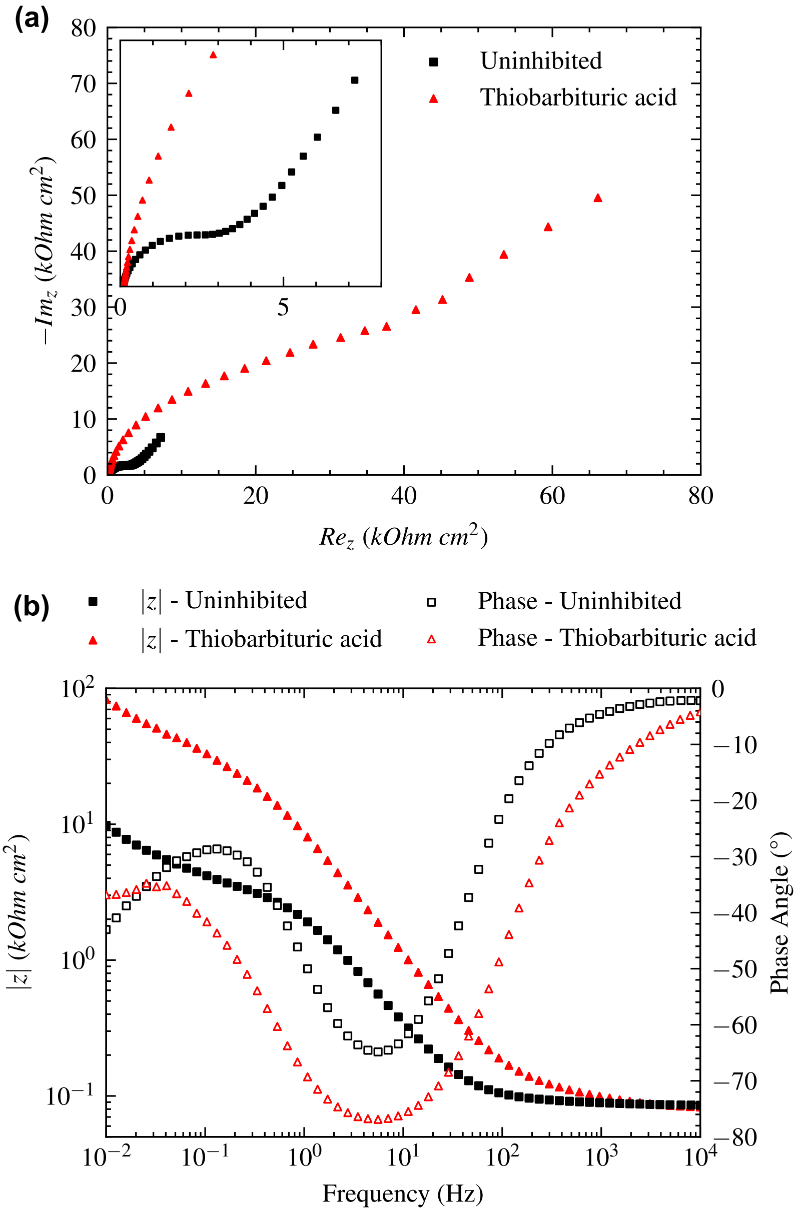


Figure 5.7: Electrochemical impedance spectroscopy of AA2024-T3 alloy exposed to 0.1M NaCl electrolytes with or without 2-thiobarbituric acid for 24 hours - (a) Nyquist, (b) Bode modulus and phase angle plots. Inset shows a zoom in of lower resistance values. Filled markers in the presence, empty markers in the subsequent absence of inhibitor.

their decision-making processes. By representing molecules using various featurisation methods and applying feature elimination techniques to identify the most important features, we gain insight into which feature combinations represent the problem best. These insights can then be integrated with the domain knowledge of scientists to utilise machine learning models beyond their typical "black-box" functionality.

However, it is crucial to remain aware of the limitations of different molecular representations. For example, while hashing-based methods are more generalisable, they may lead to bit collisions, making models volatile and less interpretable. Fingerprint techniques, though useful, may overlook subtle molecular changes if these do not alter the structural fragments being represented, whereas physicochemical descriptors may be more suitable for capturing such nuances. Nevertheless, fingerprints can effectively capture broader trends, as they are closely tied to molecular structure, and their interpretation is often more straightforward since they represent visualisable substructures—provided there are no bit collisions.

The combination of diverse molecular representations holds significant, largely untapped potential for scientific discovery via statistical models. Agreement across models with different featurisation methods can allow feature selection to be used as a tool akin to using various spectroscopic tools in materials science. Additionally, SHAP (SHapley Additive exPlanations) analysis offers promise in isolating the effects of complex trends, and it is highly effective in creating controlled variables within a materials research framework. Insights gained from different representations can complement one another, forming the basis for testable hypotheses, as illustrated here in the discovery of a novel corrosion inhibitor 2-thiobarbituric acid for AA2024-T3.

Next to what has been studied in this work, further insights can be gained by manipulating molecular structures—such as adding or removing fragments—at no additional cost after the model has been trained, allowing for the testing of trends in the material properties of interest. If these insights can then be integrated into generative chemical foundation models, it would enable the rapid design of new molecules at a fraction of the original cost.

5.5. SUPPLEMENTARY INFORMATION

5.5.1. PREDICTIVE PERFORMANCE OF THE MACHINE LEARNING MODELS

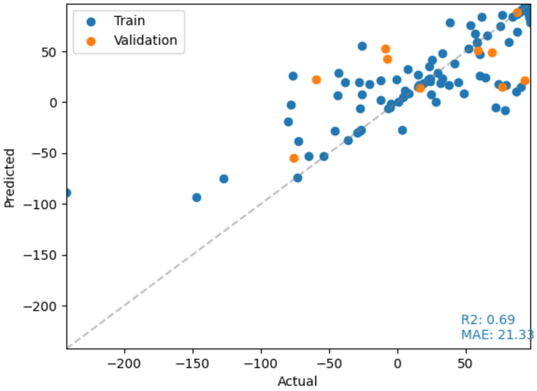


Figure 5.8: Prediction plots of training with cross-validation (in-blue) and set-aside validation split (in-orange) for the highest ranked model with IE_EIS24h target and atompair-count featurisation.

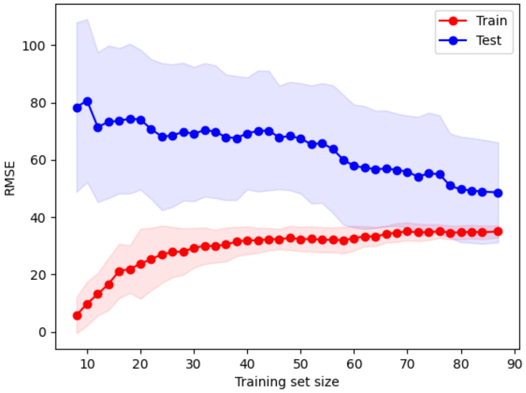


Figure 5.9: Learning curves for the highest ranked model with IE_EIS24h target and atompair-count featurisation.

5.5.2. DISTRIBUTION OF PREDICTIVE PERFORMANCE FOR DIFFERENT MODEL CONFIGURATIONS

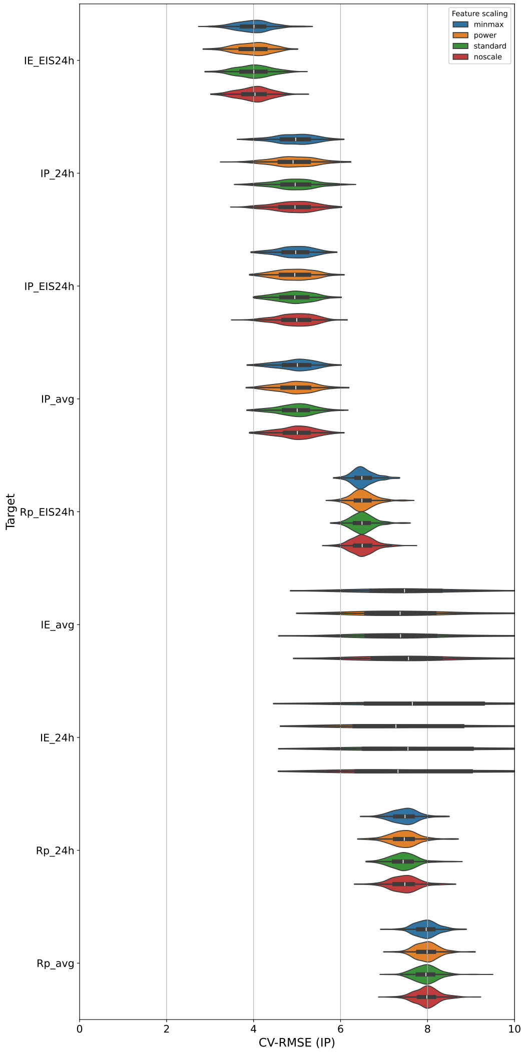


Figure 5.10: Influence of feature scaling on the CV-RMSE distributions.

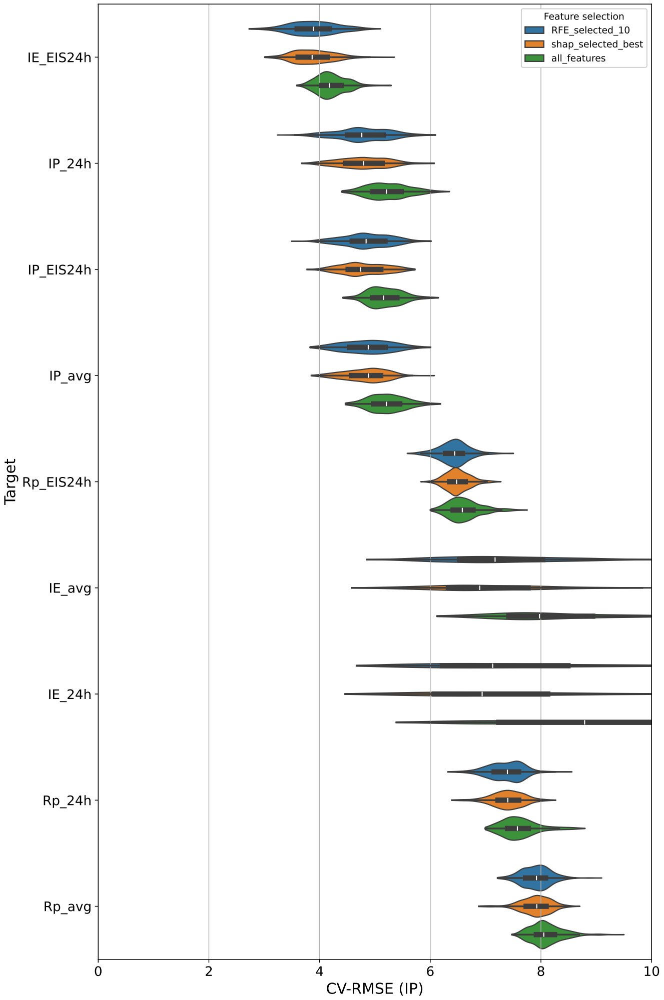


Figure 5.11: Influence of feature selection method on the CV-RMSE distributions.

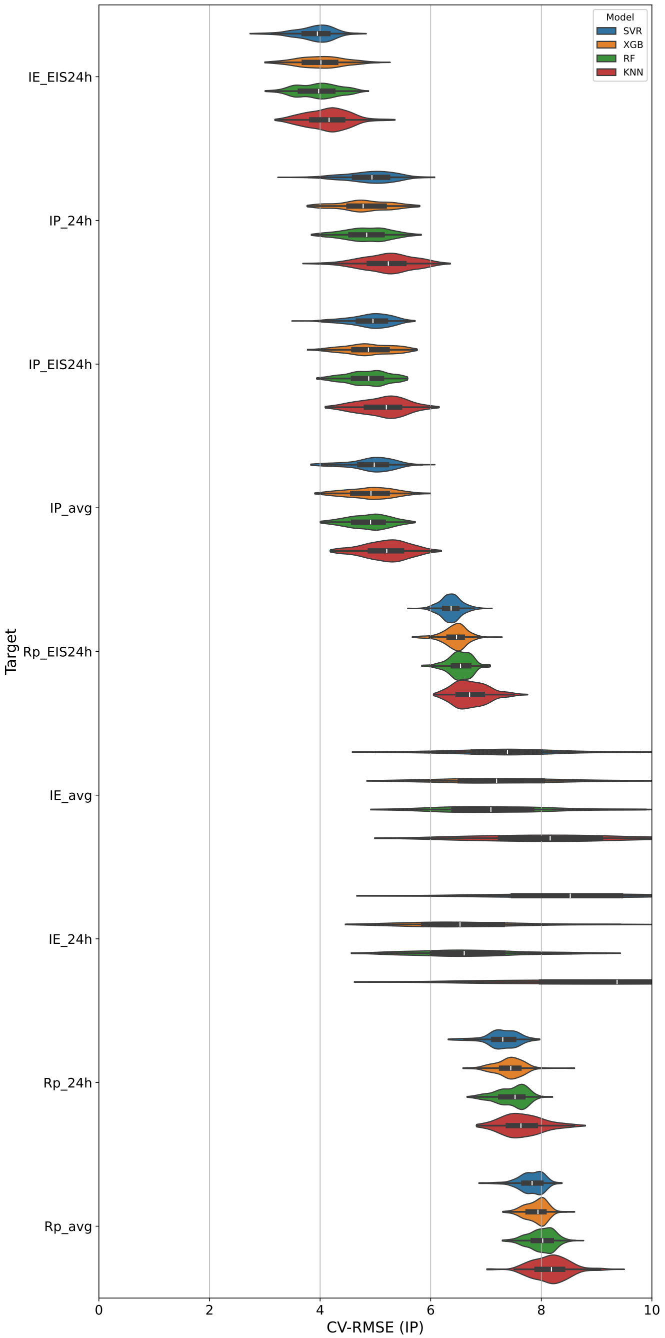


Figure 5.12: Influence of model architecture on the CV-RMSE distributions.

6

Conclusions and Outlook

The numbers have no way of speaking for themselves.
We speak for them. We imbue them with meaning.

Nate Silver

6.1. Key Scientific Contributions

The ultimate aspiration of this dissertation was to explore how the structure of organic molecules influences the electrochemical behaviour related to the corrosion of aerospace alloys, and if possible find new ways to capitalise on gained knowledge to develop better self-healing corrosion inhibition systems. Following the zeitgeist of the fourth paradigm of science, a data-driven approach laid the groundwork for the identification of trends arising from diverse molecule-surface interactions, aiming to evaluate the potential of organic molecules as environmentally friendly alternatives to hexavalent chromium. Rather than relying on simplified model alloys, the study focused directly on the AA2024-T3 with its complex microstructure to allow the results of this research to be directly transferable to real-world applications.

As the first step, the correct way of collecting corrosion inhibition data was needed to be determined. The aim was to simulate the conditions relevant to aerospace alloys - the corrosion inhibition behaviour arising from low organic molecule concentrations achievable through leaching from organic coatings, without dealing with the complexities of coating - molecule interactions. As the reported electrochemical corrosion inhibition performance of organic molecules varied inconsistently across the literature [107, 136], a time-resolved data generation process that combined linear polarisation resistance, electrochemical impedance spectroscopy, and potentiodynamic polarisation was designed. This methodology was used to create the first outcome of this dissertation: an electrochemical dataset of AA2024-T3 surfaces exposed to 0.1M NaCl in the presence of more than 100 small organic molecules in 1 mM concentrations.

This methodology allowed complete control over the experimental details: knowing exactly where the compounds came from, their concentrations, how surfaces-to-be-studied were prepared, whether any treatments were applied to the electrolytes, access to solubility and pH information before and after the electrochemical experiments, and even the influence of the electrochemical cell and potentiostat setup. Many of these critical details are often missing from published methodologies so by keeping all these free variables constant, the true analysis goal can be brought into focus: how molecular structure affects corrosion inhibition for the given substrate.

Prior to decisively connecting the molecular structure to electrochemical behaviour, some key critical factors influencing the corrosion inhibition behaviour was identified. Our findings show that it is not possible to draw definitive conclusions about the long-term corrosion inhibition performance of organic molecules without considering key variables such as exposure time, molecule concentration, environmental conditions, the physicochemical/electrochemical stability of molecules in dynamic systems, and potential synergies between multiple molecules interacting with each other and the surface.

While some organic inhibitors initially provided corrosion protection comparable to sodium dichromate at 1 mM concentrations within the first six hours, their effectiveness tended to stabilise with time, while chromate-based inhibition continued to improve with longer exposure. Increasing the concentration of organic inhibitors generally enhanced their performance, but only up to a critical concentration (typically 2–5 mM for the studied molecules), beyond which the inhibition either plateaued or even declined. Environmental conditions also played a crucial role. At extreme pH values (below 4.0 or above 8.5) some molecules that otherwise showed promise instead accelerated corrosion. This effect was linked to the electrolyte pH shifting the aluminium surface outside the thermodynamically stable Al_2O_3 window, making it more vulnerable to degradation.

It was observed that for most of the tested organic molecules, supplied corrosion inhibition was reversible. This limited their long-term effectiveness, particularly in dynamic environments where a constant inhibitor reservoir was not present. The best-performing inhibitors were not necessarily the most physically stable or irreversible - they did not always maintain higher inhibition after being removed from the electrolyte. However, when these organic molecules remained in the environment, they provided corrosion protection across a broad potential range. A small subset of molecules demonstrated quasi-stable corrosion inhibition, meaning they maintained corrosion inhibition properties even after their subsequent absence in the electrolyte, albeit at a reduced level. This behaviour suggests that long-term corrosion protection is possible with the right small organic molecules, even in the absence of a continuous inhibitor supply.

The observed quasi-sustained corrosion inhibition effect was further investigated and attributed primarily to dominant anodic protection. The prevailing hypothesis suggests that in the presence of the 3-amino-1,2,4-triazole-5-thiol, the molecule completely covers the surface in multilayers: it interacts with the Al-(hydr)oxide matrix to form a sulfate-like structure, while also adsorbing to intermetallic sites to suppress their susceptibility to corrosion initiation. In the subsequent absence of the molecule weakly bonded molecules desorb, corrosion inhibition on intermetallics decreases due to rearrangement of preferred bonding configuration which no longer is sufficient as a barrier to corrosive ions such as chloride; whereas sulfate-like structure keeps its presence over the Al-(hydr)oxide matrix further stabilising it, resulting in quasi-sustained corrosion inhibition. This understanding serves as a model for designing future systems capable of achieving fully corrosion-inhibited surfaces.

Inhibitor synergy offered by presence of multiple molecules proved to be powerful. It was shown that two organic molecules with individually unremarkable electrochemical performance could work together to produce significantly stronger and more stable corrosion inhibition. In some cases, these synergistic effects led to electrochemical performance surpassing that of state-of-the-art chromate inhibitors, even in the subsequent absence of the molecules in the environment.

The structural analysis of the molecules and their inhibition power distribution showed that simplistic correlations such as the presence of N/O/S heteroatoms, without considering where and how they are structured is not enough to gain insight on corrosion inhibition phenomena. This is also made clear in Figures A.1-A.3, which include example molecules with pyridine, pyrimidine, and azole-derivative structures visualised in a 2D projection of their chemical space, where molecules close to one another are chemically similar. Small changes such as the addition of a functional group in different positions, or the same addition for different previous skeleton structures can result in wildly different corrosion inhibition behaviour, implying a chemical space abundant in “activity cliffs”. These molecular structural relationships showed the need for advanced statistical models based on machine learning to capture the trends found in the experimental data.

A systematic analysis of the representation of the molecule and the target to be predicted revealed that the combination of diverse molecular representations holds significant, largely untapped potential for scientific discovery via statistical models. The inclusion of the mechanistic insights in machine learning models proved crucial, with experimentally obtained parameters like pH providing invaluable contextual information about the system. A methodology that focuses on reverse-engineering statistical models to understand their decision-making processes was created to integrate insights coming from the models with the domain-expert. The combinations that best represent the problem from complementary fingerprint- and physicochemical descriptor-based featurisation methods allowed the identification of the most critical fragments for corrosion inhibition. The fragments related to C(=S)N and C-4x=O were used to identify promising compounds from a toxicity database, which allowed 2-thiobarbituric acid to be discovered as a novel corrosion inhibitor.

In summary, this work has made several key contributions to the scientific corpus:

- Developed a methodology for creating a robust corrosion inhibitor database through electrochemical experiments. This resulted in a database with the single largest number of molecules measured for corrosion inhibition for a given state.
- Identified the most critical factors governing corrosion inhibition in the presence (and subsequent absence) of small organic molecules. Demonstrated that failing to account for parameters such as pH, exposure time, molecule concentration, and the continuous versus interrupted presence of the molecule can lead to incorrect conclusions about a compound's corrosion inhibition performance.
- Investigated the mechanisms underlying the rarely observed phenomenon of quasi-sustained corrosion inhibition. Built a hypothesis for the mechanisms that would sustain the inhibition, which will assist researchers in creating molecules that sustain their corrosion protection in the changing environmental conditions found in industry applications.
- Established a machine learning framework for extracting meaningful scientific insights from small datasets, enabling the discovery of novel functional molecules. Through the framework, a novel inhibitor molecule never before used as a corrosion inhibitor for aluminium alloys was found, and further validated through electrochemical experiments, showing the potential of such a method.
- Invented a patent-pending corrosion protection system based on molecule synergy. This system shows a never-before-seen corrosion inhibition performance superior even to the chromate both in the presence, and the subsequent absence of molecules in the environment.

6.2. OUTLOOK

As with any project, this research was also limited in scope by finite manpower, lack of infinite time, known, and unknown unknowns. Based on the experience gained while conducting this body of work, a few suggestions that can be realised in the short and long term are presented below.

The next step in corrosion inhibition research lies in an integrated approach that combines: (i) a clear mechanistic understanding of corrosion inhibition, (ii) a strong experimental foundation to reveal underlying processes, (iii) high-throughput methodologies to accelerate material screening, (iv) statistical models that guide both experiments and simulations through active learning, and (v) robust in-silico screening powered by machine learning to navigate the vast chemical spaces.

For this end, universities should capitalise more on learning from adjacent fields, and not fear cannibalising the already-available methods developed in industries such as pharmaceuticals and semiconductors. One highly relevant advancement is the high-throughput screening and automated experimentation setups based on self-driving labs. Such systems can be easily adapted to electrochemical sciences working on electrolyte-surface interactions, which make them ideal in kickstarting the materials discovery revolution for corrosion inhibitors. These automated platforms can reinvent materials discovery for corrosion inhibitors, rendering currently intractable problems, such as testing inhibitor systems with ternary or more complex compositions, or evaluating coatings with spatial gradients, both feasible and efficiently solvable [332, 333]. Such setups based on collecting quantitative electrochemical data can generate transdisciplinary datasets that can assist in various electrode-electrolyte based problems. By facilitating the generation of high-quality large-scale datasets, such systems are poised to enable a fully closed materials discovery loop encompassing materials characterisation, prediction, and synthesis. The resultant acceleration in knowledge acquisition could unlock vast potentials within molecular chemistry.

A critical next step is to leverage machine learning not merely as a black-box predictive tool, but as an integrated and interpretive component of scientific discovery. As shown in this work, with the right methodology, machine learning can serve as a characterisation tool that parallels established spectroscopy techniques. Future development along this research line should aim at enhancing the interpretability of models, which would extend their utility beyond prediction and toward facilitating nuanced insights into molecular behaviours and interactions.

Another route for future work is the standardisation of material data and metadata generation. A machine-readable experimental measurement and material representation are needed to amplify the scientist with the algorithmic insight. Given that the optimal representation of a material is inherently problem-specific, one pressing need is to develop novel descriptors that accurately capture surface–molecule interactions. These descriptors are not only critical for understanding corrosion inhibition but also relevant to a wide range of surface-related phenomena, including electrochemical sensors, carbon conversion, battery performance, and heterogeneous catalysis, among others. Establishing universal surface descriptors could enable the integration of diverse datasets that combine both organic and inorganic chemistries into combined foundational models. Such models would bridge domain-specific gaps and foster a more unified view of catalytic and inhibitory processes linked by consistent physicochemical principles.

The ultimate advantage of robust, working models lies in their capacity to rapidly and cost-effectively traverse the chemical space. Once properly trained, these models would become virtual laboratories where molecular structures can be systematically manipulated by adding or removing molecular fragments, and in this way allow analysing trends in material properties without incurring additional experimental costs. This approach would not only broaden understanding of material behaviour, but also would lay the groundwork for generative chemical models capable of designing entirely new molecules. Such models could revolutionise the rate and cost-effectiveness of chemical innovation, fundamentally altering the materials design norms.

Of course all these developments depend on whether the measurements measure, and data is generated for the right phenomena. Unfortunately, a recent study highlighted that accelerated tests often apply extreme conditions that may trigger degradation mechanisms not dominant in service, leading to inaccurate material lifetime predictions [334]. The lack of correlation between accelerated testing methods and material degradation under real service conditions is a clarion call for bridging the lab to real-world applications, which requires a deep involvement from both sides of the academia-industry complex. Integrating in-situ and operando techniques with the prevalent methods can prove key in revealing service-relevant mechanisms, while data-driven approaches can link lab and field data more reliably. Bridging this gap will prove foundational in converting science into societal developments.

Finally, the societal implications of these advancements should not be underestimated. The principles and methodologies developed here could extend well beyond corrosion inhibition of aerospace materials, or corrosion inhibition altogether. For example, the generated electrochemical dataset and explored sustained inhibition mechanism could directly be applied to the development of aluminium-based battery anodes, whereas developed machine learning methodologies can be directly applied to materials discovery problems related to heterogenous catalysis or carbon capture, contributing directly to the rapidly growing field of sustainable energy. By fostering interdisciplinary collaboration and technological innovation, the integration of automated experimentation, machine learning, and advanced data standardisation has the potential to make significant contributions to both scientific progress and societal well-being.

A

Visualising chemical space for azole and pyridine/pyrimidine derivative molecules

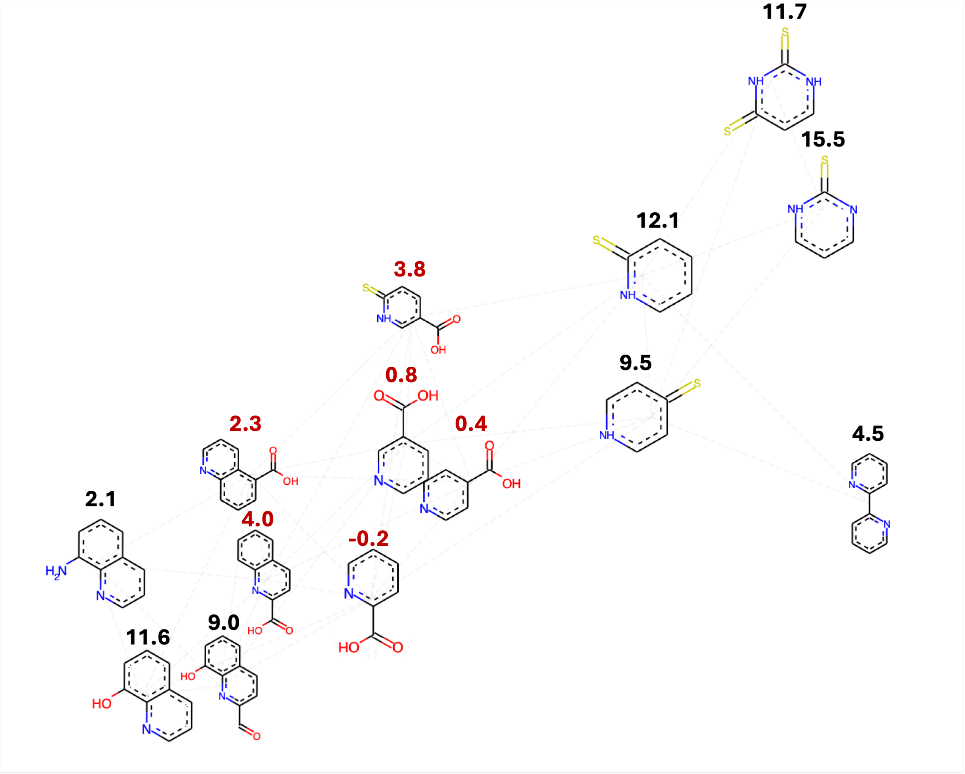


Figure A.1: The distribution of some of the studied nitrogen heteroatom containing 6-ringed molecules in the chemical space with their inhibition power (measured by EIS after 24 exposure to 1 mM concentration molecule and 0.1M NaCl). Distances between molecules correspond to their Tanimoto similarities. Molecules that are closer to one another are more chemically similar. Inhibition power is marked above each molecule, red marked font indicates pH values below or around 4.0.

The effect of addition or removal of functional groups can be visualised through chemical space networks. In the Figures A.1 - A.3, chemical space of azole and pyridine/pyrimidine derivative molecules are visualised as a case study, based on the methodology explored previously [335]. Linear relationships are often difficult to discern, as identical functional group additions can yield opposite outcomes. This complexity arises from several interrelated factors, including changes in the molecule's influence on the surface work potential, variations in adsorption properties, differences in steric and repulsive effects on corrosive ions (such as chloride), and molecule-induced pH shifts. Together, these factors underscore the intricate nature of even the simplest small molecules and their consequent corrosion inhibition efficacy.

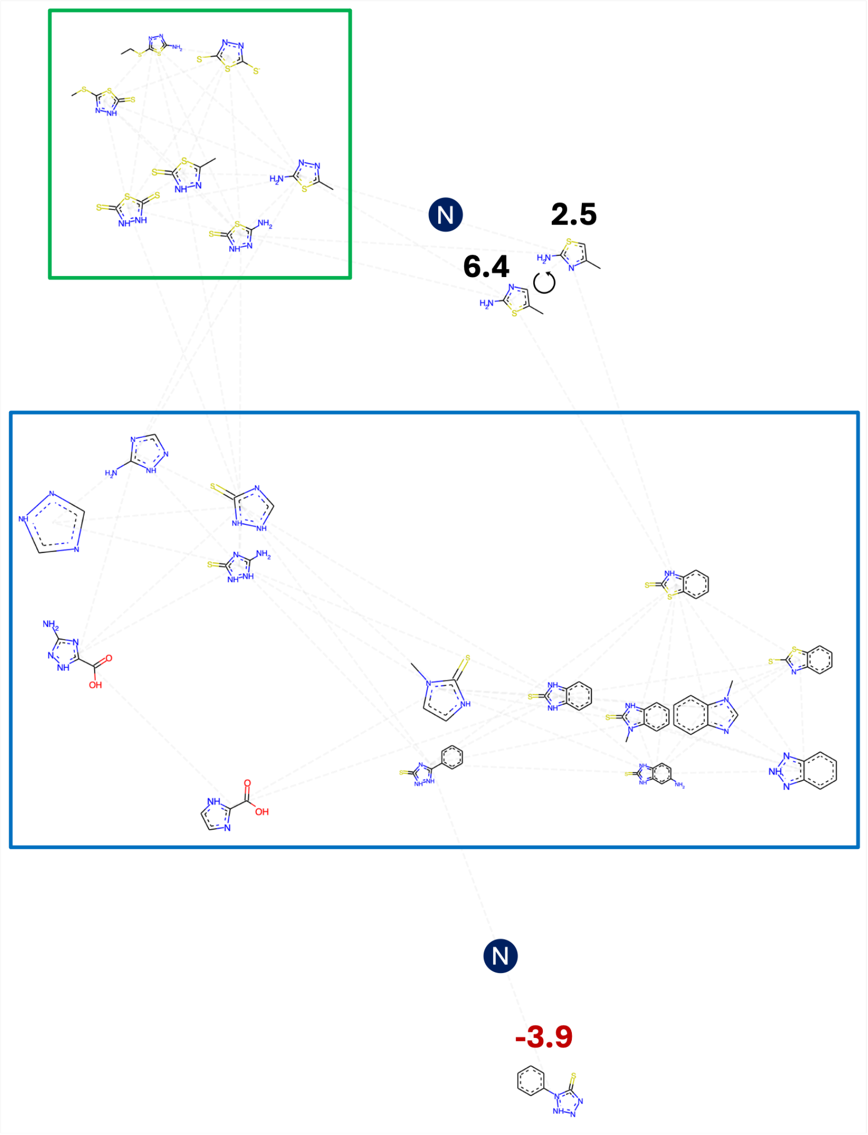


Figure A.2: The distribution of some of the studied nitrogen heteroatom containing 5-ringed molecules (azoles) in the chemical space with their inhibition power (measured by EIS after 24 exposure to 1 mM concentration molecule and 0.1M NaCl). Distances between molecules correspond to their Tanimoto similarities. Molecules that are closer to one another are more chemically similar. Inhibition power is marked above each molecule, red marked font indicates pH values below or around 4.0, N and arrow corresponding to addition of ring nitrogen and moving of a methyl bond to neighbouring position, respectively. Marked regions are zoomed-in in Figure A.3.

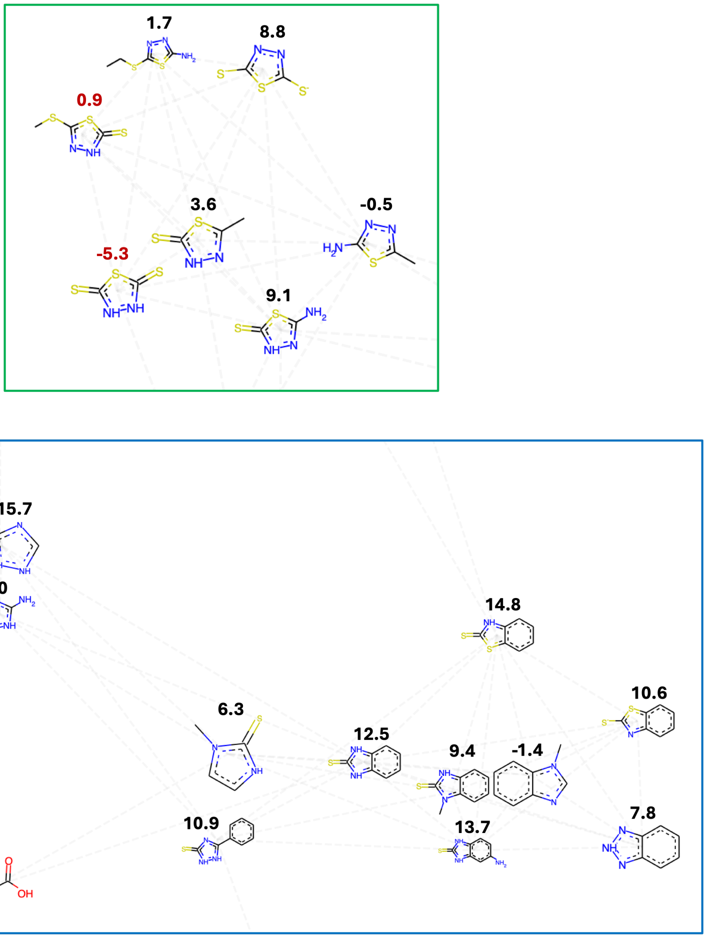


Figure A.3: Zoom in of chemical distributions of the molecules in marked regions found in Figure A.2.

B

Mol-dex

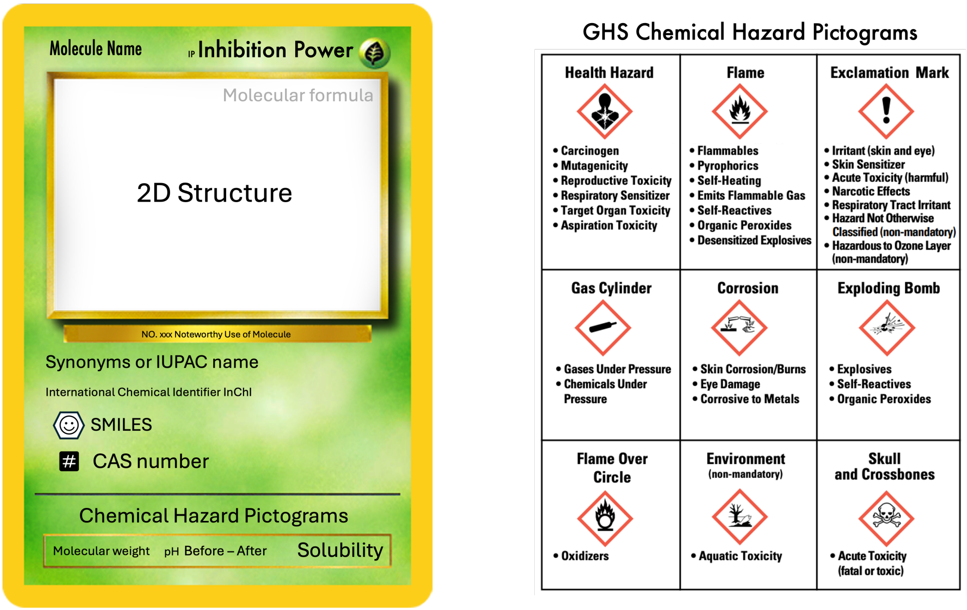


Figure B.1: Card template and GHS chemical hazard pictogram explanations.

This section collects all the studied molecules as cards designed as a visual aid to assist identification of trends, which were collected together in the form of a molecule index Mol-Dex.

The categorisation of molecules are performed based on the structure of the molecule. Each card collects information on the structure of the molecule, its corrosion inhibition performance, its common and IUPAC names and other chemical identifiers, experimental details such as the solubility of the molecule at 1 mM concentrations and resulting pH of the solution.

Each card contains information on:

- Molecule Name: purchase name of the molecule.
- IP Inhibition Power: Inhibition power calculated from the low frequency (10^{-1} Hz) impedance measurements performed with electrochemical impedance spectroscopy after 24 hour exposure to the 1 mM molecule in 0.1M NaCl.
- Molecular formula: closed formula of the molecule.
- 2D structure: depiction of the molecule as a 2D figure.
- Noteworth use of the molecule.
- Synonyms or IUPAC name.
- International Chemical Identifier InChI key: a textual identifier for chemical substances, created to standardise the encoding of molecular information and simplify the process of searching for this information in databases.

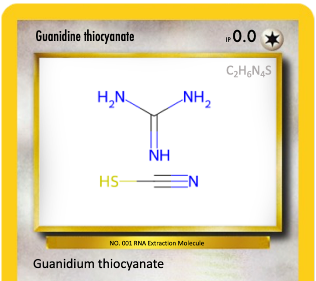
- SMILES (Simplified Molecular Input Line Entry System): a notation system that represents the structure of chemical molecules using short, ASCII text strings.
- CAS number: a unique identification number, assigned by the Chemical Abstracts Service (CAS) to index every chemical substance described in the open scientific literature.
- Chemical Hazard Pictograms: graphical symbols used to identify and communicate specific hazards associated with chemical substances and mixtures, as defined by the Globally Harmonised System of Classification and Labelling of Chemicals (GHS).
- Molecular weight.
- pH Before - After: pH of the solutions prepared at 1 mM molecule concentrations in 0.1M NaCl measured before and after the electrochemical screening experiments.
- Solubility: whether the molecule was fully soluble at 1 mM concentration.

Cards were assigned colors according to their ring structures, the presence of heteroatoms in the rings, and functional groups. The assignment was done in the following descending order:

- Green (leaf): fused benzene and 5-membered ring structures that contain nitrogen or sulfur.
- Blue (droplet): thiazole structures that contain both nitrogen and sulfur in a 5-membered ring.
- Yellow (thunder): triazole structures with 3 nitrogens in a 5-membered ring.
- Red (flame): imidazole or pyrole structures that contain at least 2 nitrogen in a 5-membered ring.
- Purple (eye): 6-membered heterocycles that contain at least 1 nitrogen or sulfur.
- Brown (fist): organic acids that contain carboxyl group.
- White (star): linear hydrocarbons with or without nitrogen or sulfur.
- Black (inverted triangle): rest of the molecules.

Guanidine thiocyanate IP 0.0

$C_2H_6N_4S$



NO. 001 RNA Extraction Molecule

Guanidium thiocyanate

InChI=1S/CH5N3.CHNS/c2-1(3)4;2-1-3/h(H5,2,3,4);3H

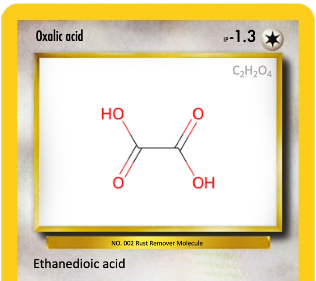
C(#N)S.C(=N)(N)N

593-84-0

118 g/mol pH 5.8 – 5.9 Soluble

Oxalic acid IP -1.3

$C_2H_2O_4$



NO. 002 Rust Remover Molecule

Ethanedioic acid

InChI=1S/C2H2O4/c3-1(4)2(5)6/h(H,3,4)(H,5,6)

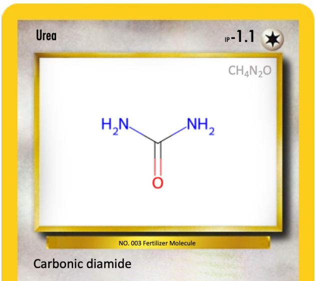
C(=O)(C(=O)O)O

144-62-7

90 g/mol pH 2.9 – 2.9 Soluble

Urea IP -1.1

CH_4N_2O



NO. 003 Fertilizer Molecule

Carbonic diamide

InChI=1S/CH4N2O/c2-1(3)4/h(H4,2,3,4)

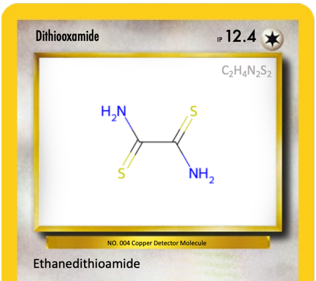
C(=O)(N)N

57-13-6

60 g/mol pH 5.6 – 5.7 Soluble

Dithiooxamide IP 12.4

$C_2H_4N_2S_2$



NO. 004 Copper Detector Molecule

Ethanedithioamide

InChI=1S/C2H4N2S2/c3-1(5)2(4)6/h(H2,3,5)(H2,4,6)

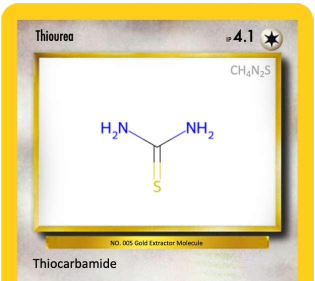
C(=S)(C(=S)N)N

79-40-3

120 g/mol pH 5.6 – 5.6 Soluble

Thiourea IP 4.1

CH_4N_2S



NO. 005 Gold Extractor Molecule

Thiocarbamide

InChI=1S/CH4N2S/c2-1(3)4/h(H4,2,3,4)

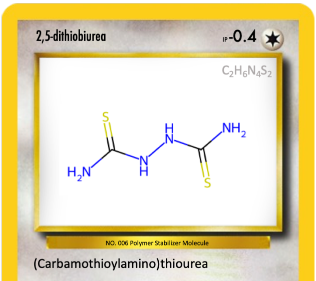
C(=S)(N)N

62-56-6

76 g/mol pH 5.9 – 6.0 Soluble

2,5-dithiobiurea IP -0.4

$C_2H_6N_4S_2$



NO. 006 Polymer Stabilizer Molecule

(Carbamothioylamino)thiourea

InChI=1S/C2H6N4S2/c3-1(7)5-6-2(4)8/h(H3,3,5,7)(H3,4,6,8)

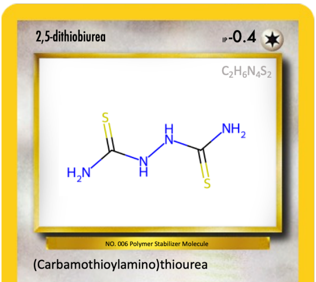
C(=S)(N)NNC(=S)N

142-46-1

150 g/mol pH 3.0 – 3.1 Soluble

2,5-dithiobiurea IP -0.4

$C_2H_6N_4S_2$



NO. 006 Polymer Stabilizer Molecule

(Carbamothioylamino)thiourea

InChI=1S/C2H6N4S2/c3-1(7)5-6-2(4)8/h(H3,3,5,7)(H3,4,6,8)

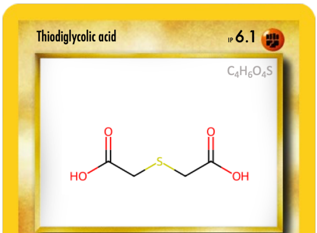
C(=S)(N)NNC(=S)N

142-46-1

150 g/mol pH 3.0 – 3.1 Soluble

Thiodiglycolic acid IP 6.1

$C_4H_6O_4S$



NO. 008 Cosmetics Reducing Agent Molecule

2-(carboxymethylsulfanyl)acetic acid

InChI=1S/C4H6O4S/c5-3(6)1-9-2-4(7)8/h1-2H2,(H,5,6)(H,7,8)

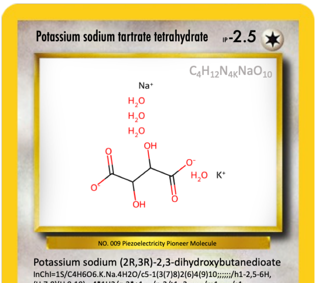
C(C(=O)O)SCC(=O)O

123-93-3

150 g/mol pH 3.5 – 3.5 Soluble

Potassium sodium tartrate tetrahydrate IP -2.5

$C_4H_{12}N_{4K}NaO_{10}$



NO. 009 Piezoelectricity Pioneer Molecule

Potassium sodium (2R,3R)-2,3-dihydroxybutanedioate

InChI=1S/C4H6O6.K.Na.4H2O/c5-1(3(7)8)2(6)4(9)10;;;;;;/h1-2,5-6H,(H,7,8)(H,9,10);;;4*1H2/q;2*+1;;;;/p-2/t1-,2-;;;;;;/m1....../s1

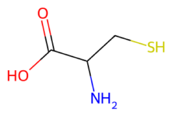
C(C(C(=O)[O-])O)(C(=O)[O-])O.O.O.O.O.[Na+].[K+]

6381-59-5

282 g/mol pH 6.1 – 6.5 Soluble

L-cysteine IP 15.2

$C_3H_7NO_2S$



NO. 010 Hydrophobic Amino Acid Molecule

(2R)-2-Amino-3-sulfanylpropanoic acid

InChI=1S/C3H7NO2S/c4-2(1-7)3(5)6/h2,7H,1,4H2,(H,5,6)/t2-/m0/s1

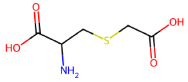
C(C(C(=O)O)N)S

52-90-4

121 g/mol pH 6.1 – 6.0 Soluble

Carbocysteine IP 1.3

$C_3H_4N_4O_2$



NO. 011 Mucolytic Molecule

(2R)-2-amino-3-(carboxymethylsulfanyl)propanoic acid
InChI=1S/C5H9NO4S/c6-3(5(9)10)1-11-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/t3-/m0/s1

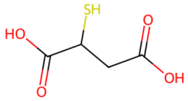
C(C(C(=O)O)N)SCC(=O)O

364638-23-3

179 g/mol pH 3.3 – 3.5 Soluble

Mercaptosuccinic acid IP 6.9

$C_4H_6O_4S$



NO. 012 Heavy Metal Poisoning Theraphy Molecule

2-Sulfanylbutanedioic acid

InChI=1S/C4H6O4S/c5-3(6)1-2(9)4(7)8/h2,9H,1H2,(H,5,6)(H,7,8)

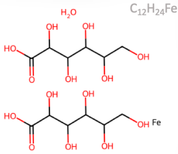
C(C(C(=O)O)S)C(=O)O

70-94-5

150 g/mol pH 3.0 – 3.4 Soluble

Iron gluconate monohydrate IP -2.0

$C_{12}H_{24}FeO_{15}$



NO. 013 Iron Modulator Molecule

Iron (2R,3S,4R,5R)-2,3,4,5,6-pentahydroxyhexanoate
InChI=1S/2C6H12O7.Fe/c2*7-1-2(8)3(9)4(10)5(11)6(12)13;/h2*2-5,7-11H,1H2,(H,12,13);/q;;+2/p-2/t2*2-,3-,4+,5-;/m11./s1

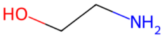
C(C(C(C(C(C(=O)O)O)O)O)O)O.C(C(C(C(C(C(=O)O)O)O)O)O)O.O.[Fe]

699014-53-4

464 g/mol pH 4.6 – 4.3 Soluble

Ethanolamine IP -3.6

C_2H_7NO



NO. 014 Cellular Membrane Maker Molecule

2-aminoethanol

InChI=1S/C2H7NO/c3-1-2-4/h4H,1-3H2

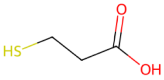
C(CO)N

141-43-5

61 g/mol pH 10.2 – 9.0 Soluble

3-mercaptopropionic acid IP 3.8

$C_3H_6O_2S$



NO. 015 Gold Nanoparticle Maker Molecule

3-Sulfanylpropanoic acid

InChI=1S/C3H4N4O2/c4-3-5-1(2(8)9)6-7-3/h(H,8,9)(H3,4,5,6,7)

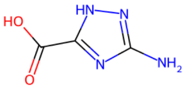
C(CS)C(=O)O

107-96-0

106 g/mol pH 3.4 – 3.5 Soluble

3-amino-1,2,4-triazole-5-carboxylic acid IP 0.7

$C_3H_4N_4O_2$



NO. 016 Medicine Builder Molecule

3-amino-1H-1,2,4-triazole-5-carboxylic acid

InChI=1S/C3H4N4O2/c4-3-5-1(2(8)9)6-7-3/h(H,8,9)(H3,4,5,6,7)

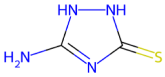
C1(=NC(=NN1)N)C(=O)O

3641-13-2

128 g/mol pH 3.7 – 3.8 Soluble

3-amino-5-mercapto-1,2,4-triazole IP 14.0

$C_2H_4N_4S$



NO. 017 Stable Corrosion Inhibition Molecule

5-amino-1,2-dihydro-1,2,4-triazole-3-thione

InChI=1S/C2H4N4S/c3-1-4-2(7)6-5-1/h(H4,3,4,5,6,7)

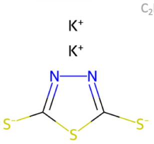
C1(=NC(=S)NN1)N

16691-43-3

116 g/mol pH 5.4 – 5.5 Soluble

1,3,4-thiadiazole-2,5-dithiol dipotassium IP 8.8

$C_2K_2N_2S_3$



NO. 018 Metal Sensor Molecule

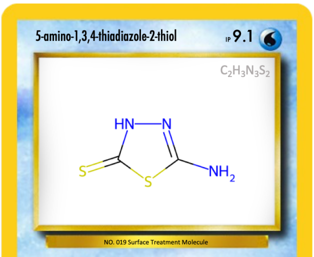
Dipotassium 1,3,4-thiadiazole-2,5-dithiolate
InChI=1S/C2H2N2S3.2K/c5-1-3-4-2(6)7-1;;/h(H,3,5)(H,4,6);;/q;2*+1/p-2

C1(=NN=C(S1)[S-])[S-].[K+].[K+]

4628-94-8

226 g/mol pH 8.4 – 7.4 Soluble

5-amino-1,3,4-thiadiazole-2-thiol IP 9.1



$C_2H_3N_3S_2$

NO. 019 Surface Treatment Molecule

5-Amino-3H-1,3,4-thiadiazole-2-thione

InChI=1S/C2H3N3S2/c3-1-4-5-2(6)7-1/h(H2,3,4)(H,5,6)

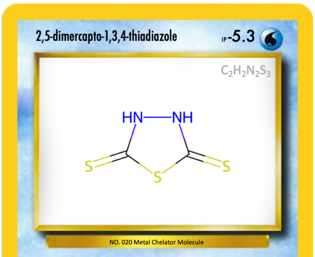
☺ C1(=NNC(=S)S1)N

2349-67-9

⚠

133 g/mol pH 4.5 – 4.7 Soluble

2,5-dimercapto-1,3,4-thiadiazole IP -5.3



$C_2H_2N_2S_3$

NO. 020 Metal Chelator Molecule

1,3,4-thiadiazolidine-2,5-dithione

InChI=1S/C2H2N2S3/c5-1-3-4-2(6)7-1/h(H,3,5)(H,4,6)

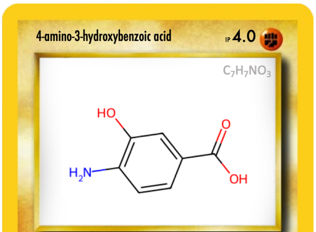
☺ C1(=S)NNC(=S)S1

1072-71-5

⚠

150 g/mol pH 3.0 – 3.2 Soluble

4-amino-3-hydroxybenzoic acid IP 4.0



$C_7H_7NO_3$

NO. 021 Dye Maker Molecule

4-Amino-3-hydroxybenzoic acid

InChI=1S/C7H7NO3/c8-5-2-1-4(7(10)11)3-6(5)9/h1-3,9H,8H2,(H,10,11)

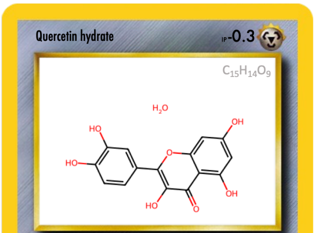
☺ C1=CC(=C(C=C1C(=O)O)O)N

2374-03-0

⚠

153 g/mol pH 3.7 – 4.1 Soluble

Quercetin hydrate IP -0.3



$C_{15}H_{14}O_9$

NO. 022 Dietary Flavonoid Molecule

2-(3,4-dihydroxyphenyl)-3,5,7-trihydroxychromen-4-one

InChI=1S/C15H10O7.H2O/c16-7-4-10(19)12-11(5-7)22-15(14(21)13(12)20)6-1-2-8(17)9(18)3-6;/h1-5,16-19,21H;1H2

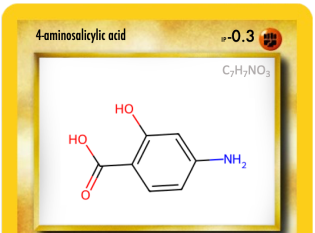
☺ C1=CC(=C(C=C1C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O)O)O.O

849061-97-8

☠

320 g/mol pH 5.6 – 6.0 Insoluble

4-aminosalicylic acid IP -0.3



$C_7H_7NO_3$

NO. 023 Tuberculosis Antibiotic Molecule

4-Amino-2-hydroxybenzoic acid

InChI=1S/C7H7NO3/c8-4-1-2-5(7(10)11)6(9)3-4/h1-3,9H,8H2,(H,10,11)

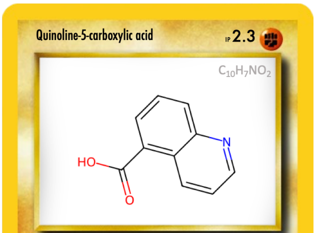
☺ C1=CC(=C(C=C1N)O)C(=O)O

65-49-6

⚠

153 g/mol pH 5.5 – 5.8 Soluble

Quinoline-5-carboxylic acid IP 2.3



$C_{10}H_7NO_2$

NO. 024 Medicine Builder Molecule

Quinoline-5-carboxylic acid

InChI=1S/C10H7NO2/c12-10(13)8-3-1-5-9-7(8)4-2-6-11-9/h1-6H,(H,12,13)

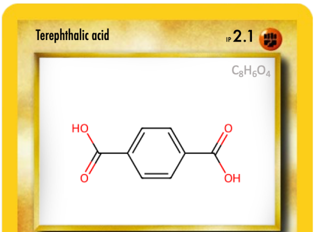
☺ C1=CC(=C2C=CC=NC2=C1)C(=O)O

7250-53-5

⚠

173 g/mol pH 4.2 – 4.3 Soluble

Terephthalic acid IP 2.1



$C_8H_6O_4$

NO. 025 PET Precursor Molecule

Terephtalic acid

InChI=1S/C8H6O4/c9-7(10)5-1-2-6(4-3-5)8(11)12/h1-4H,(H,9,10)(H,11,12)

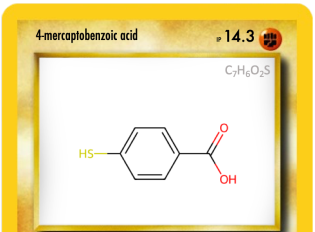
☺ C1=CC(=CC=C1C(=O)O)C(=O)O

100-21-0

⚠

166 g/mol pH 4.3 – 4.1 Soluble

4-mercaptobenzoic acid IP 14.3



$C_7H_6O_2S$

NO. 026 Cellular pH Indicator Molecule

4-sulfanylbenzoic acid

InChI=1S/C7H6O2S/c8-7(9)5-1-3-6(10)4-2-5/h1-4,10H,(H,8,9)

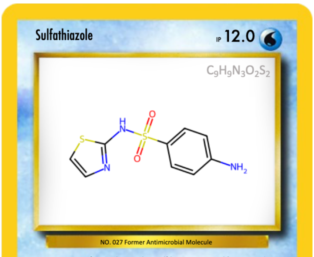
☺ C1=CC(=CC=C1C(=O)O)S

1074-36-8

⚠

154 g/mol pH 3.7 – 3.8 Insoluble

Sulfathiazole IP 12.0



$C_9H_9N_3O_2S_2$

NO. 027 Former Antimicrobial Molecule

4-amino-N-(1,3-thiazol-2-yl)benzenesulfonamide

InChI=1S/C9H9N3O2S2/c10-7-1-3-8(4-2-7)16(13,14)12-9-11-5-6-15-9/h1-6H,10H2,(H,11,12)

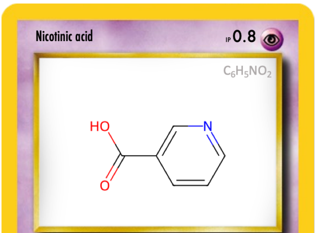
☺ C1=CC(=CC=C1N)S(=O)(=O)NC2=NC=CS2

72-14-0

⚠

255 g/mol pH 5.1 – 5.5 Soluble

Nicotinic acid IP 0.8



NO. 028 Cholesterol Reducing Molecule

Pyridine-3-carboxylic acid

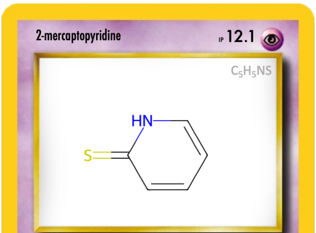
InChI=1S/C6H5NO2/c8-6(9)5-2-1-3-7-4-5/h1-4H,(H,8,9)

C1=CC(=CN=C1)C(=O)O

# 50-67-6

123 g/mol pH 3.8 – 4.0 Soluble

2-mercaptopyridine IP 12.1



NO. 029 Acylating Agent Molecule

1H-pyridine-2-thione

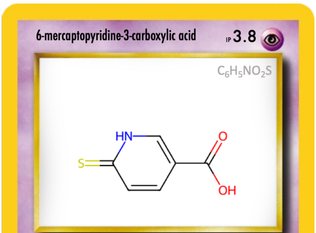
InChI=1S/C5H5NS/c7-5-3-1-2-4-6-5/h1-4H,(H,6,7)

C1=CC(=S)NC=C1

# 2637-34-5

111 g/mol pH 5.8 – 5.8 Soluble

6-mercaptopyridine-3-carboxylic acid IP 3.8



NO. 030 Metal Coordinating Molecule

6-sulfanylidene-1H-pyridine-3-carboxylic acid

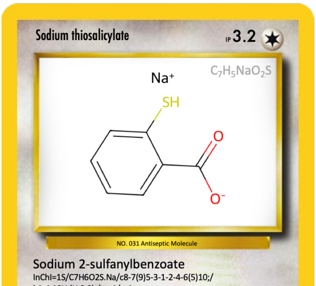
InChI=1S/C6H5NO2S/c8-6(9)4-1-2-5(10)7-3-4/h1-3H,(H,7,10)(H,8,9)

C1=CC(=S)NC=C1C(=O)O

# 17624-07-6

155 g/mol pH 3.2 – 3.3 Soluble

Sodium thiosalicylate IP 3.2



NO. 031 Antiseptic Molecule

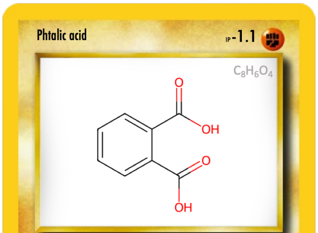
Sodium 2-sulfanylbenzoate
InChI=1S/C7H6O2S.Na/c8-7(9)5-3-1-2-4-6(5)10;/h1-4,10H,(H,8,9);/q;+1/p-1

C1=CC=C(C(=C1)C(=O)[O-])S.[Na+]

# 134-23-6

176 g/mol pH 5.3 – 5.0 Soluble

Phtalic acid IP -1.1



NO. 032 Plastic Flexer Molecule

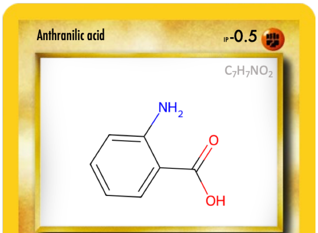
Benzene-1,2-dicarboxylic acid
InChI=1S/C8H6O4/c9-7(10)5-3-1-2-4-6(5)8(11)12/h1-4H,(H,9,10)(H,11,12)

C1=CC=C(C(=C1)C(=O)O)C(=O)O

# 88-99-3

166 g/mol pH 3.0 – 3.1 Soluble

Anthranilic acid IP -0.5



NO. 033 UV Absorber Molecule

2-aminobenzoic acid

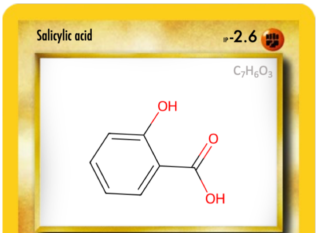
InChI=1S/C7H7NO2/c8-6-4-2-1-3-5(6)7(9)10/h1-4H,8H2,(H,9,10)

C1=CC=C(C(=C1)C(=O)O)N

# 118-92-3

137 g/mol pH 3.9 – 4.0 Soluble

Salicylic acid IP -2.6



NO. 034 Aspirin Precursor Molecule

2-hydroxybenzoic acid

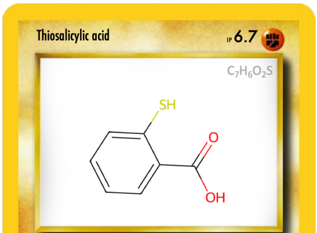
InChI=1S/C7H6O3/c8-6-4-2-1-3-5(6)7(9)10/h1-4,8H,(H,9,10)

C1=CC=C(C(=C1)C(=O)O)O

# 69-72-7

138 g/mol pH 3.1 – 3.2 Soluble

Thiosalicylic acid IP 6.7



NO. 035 Anti-inflammatory Molecule

2-sulfanylbenzoic acid

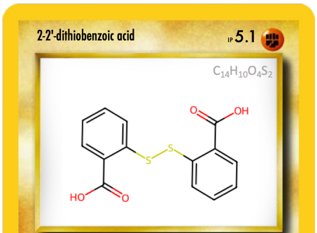
InChI=1S/C7H6O2S/c8-7(9)5-3-1-2-4-6(5)10/h1-4,10H,(H,8,9)

C1=CC=C(C(=C1)C(=O)O)S

# 147-93-3

154 g/mol pH 3.4 – 3.9 Insoluble

2-2'-dithiobenzoic acid IP 5.1



NO. 036 Cross-linker Agent Molecule

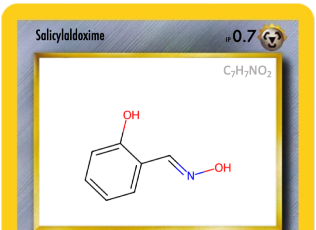
2-[(2-carboxyphenyl)disulfanyl]benzoic acid
InChI=1S/C14H10O4S2/c15-13(16)9-5-1-3-7-11(9)19-20-12-8-4-2-6-10(12)14(17)18/h1-8H,(H,15,16)(H,17,18)

C1=CC=C(C(=C1)C(=O)O)SSC2=CC=CC=C2C(=O)O

# 119-80-2

306 g/mol pH 3.7 – 3.8 Insoluble

Salicylaldoxime IP 0.7



$C_7H_7NO_2$

NO. 037 Recyclable Metal Extractor Molecule

2-[(E)-hydroxyiminomethyl]phenol

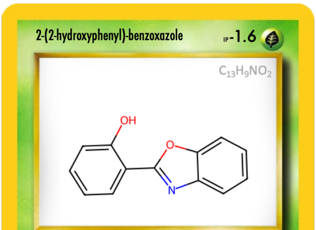
InChI=1S/C7H7NO2/c9-7-4-2-1-3-6(7)5-8-10/h1-5,9-10H/b8-5+

C1=CC=C(C(=C1)C=NO)O

94-67-7

137 g/mol pH 5.6 – 5.8 Soluble

2-(2-hydroxyphenyl)-benzoxazole IP -1.6



$C_{13}H_9NO_2$

NO. 038 Fluorescent Whitener Molecule

2-(1,3-benzoxazol-2-yl)phenol

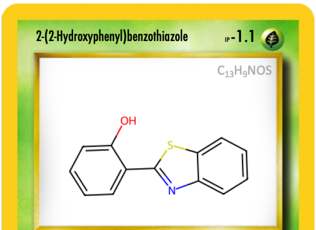
InChI=1S/C13H9NO2/c15-11-7-3-1-5-9(11)13-14-10-6-2-4-8-12(10)16-13/h1-8,15H

C1=CC=C(C(=C1)C2=NC3=CC=CC=C3O2)O

853-64-3

211 g/mol pH 6.7 – 6.0 Insoluble

2-(2-Hydroxyphenyl)benzothiazole IP -1.1



$C_{13}H_9NOS$

NO. 039 Fluorescent Sensor Molecule

2-(1,3-benzothiazol-2-yl)phenol

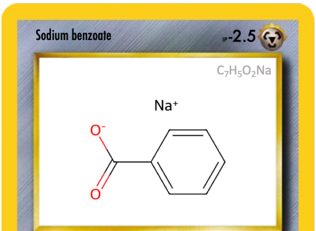
InChI=1S/C13H9NOS/c15-11-7-3-1-5-9(11)13-14-10-6-2-4-8-12(10)16-13/h1-8,15H

C1=CC=C(C(=C1)C2=NC3=CC=CC=C3S2)O

3411-95-8

227 g/mol pH 6.8 – 6.1 Soluble

Sodium benzoate IP -2.5



$C_7H_5O_2Na$

NO. 040 Preservative Molecule

Sodium benzoate

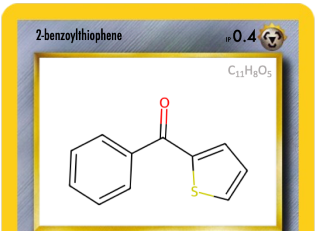
InChI=1S/C7H6O2.Na/c8-7(9)6-4-2-1-3-5-6;/h1-5H,(H,8,9);/q;+1/p-1

C1=CC=C(C=C1)C(=O)[O-].[Na+]

533-31-1

144 g/mol pH 6.3 – 6.2 Soluble

2-benzoylthiophene IP 0.4



$C_{11}H_8O_5$

NO. 041 Pharma Bridge Molecule

Phenyl(thiophen-2-yl)methanone

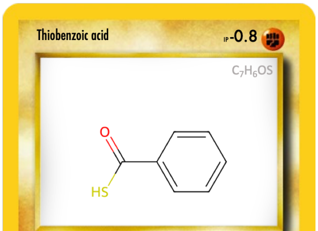
InChI=1S/C11H8OS/c12-11(10-7-4-8-13-10)9-5-2-1-3-6-9/h1-8H

C1=CC=C(C=C1)C(=O)C2=CC=CS2

135-00-2

188 g/mol pH 6.3 – 5.8 Insoluble

Thiobenzoic acid IP -0.8



C_7H_6OS

NO. 042 Preservative Molecule

Benzenecarbothioic S-acid

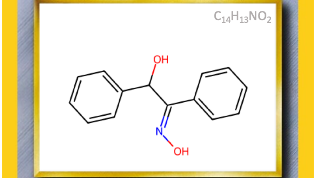
InChI=1S/C7H6OS/c8-7(9)6-4-2-1-3-5-6/h1-5H,(H,8,9)

C1=CC=C(C=C1)C(=O)S

98-91-9

138 g/mol pH 3.1 – 3.2 Soluble

Alpha-benzoin oxime IP 1.2



$C_{14}H_{13}NO_2$

NO. 043 Metal Ion Detector Molecule

(2Z)-2-hydroxyimino-1,2-diphenylethanol

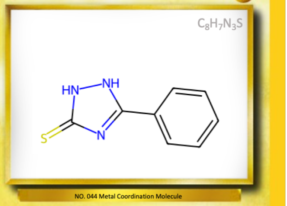
InChI=1S/C14H13NO2/c16-14(12-9-5-2-6-10-12)13(15-17)11-7-3-1-4-8-11/h1-10,14,16-17H/b15-13-

C1=CC=C(C=C1)C(C(=NO)C2=CC=CC=C2)O

441-38-3

227 g/mol pH 5.5 – 5.8 Insoluble

5-phenyl-1H-1,2,4-triazole-3-thiol IP 10.9



$C_8H_7N_3S$

NO. 044 Metal Coordination Molecule

5-phenyl-1,2-dihydro-1,2,4-triazole-3-thione

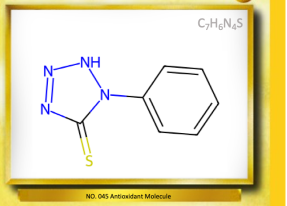
InChI=1S/C8H7N3S/c12-8-9-7(10-11-8)6-4-2-1-3-5-6/h1-5H,(H2,9,10,11,12)

C1=CC=C(C=C1)C2=NC(=S)NN2

3414-94-6

177 g/mol pH 4.8 – 5.1 Soluble

1-phenyltetrazole-5-thiol IP -3.9



$C_7H_6N_4S$

NO. 045 Antioxidant Molecule

1-phenyl-2H-tetrazole-5-thione

InChI=1S/C7H6N4S/c12-7-8-9-10-11(7)6-4-2-1-3-5-6/h1-5H,(H,8,10,12)

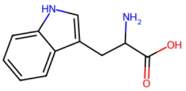
C1=CC=C(C=C1)N2C(=S)N=NN2

86-93-1

178 g/mol pH 3.1 – 3.0 Soluble

L-tryptophan IP -1.0

$C_{11}H_{12}N_2O_2$



NO. 046 Essential Amino Acid Molecule

(2S)-2-amino-3-(1H-indol-3-yl)propanoic acid

InChI=1S/C11H12N2O2/c12-9(11(14)15)5-7-6-13-10-4-2-1-3-8(7)10/h1-4,6,9,13H,5,12H2,(H,14,15)/t9-/m0/s1

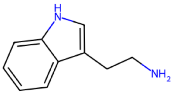
☺ C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)N

73-22-3

204 g/mol pH 5.4 – 5.8 Soluble

Tryptamine IP -2.5

$C_{10}H_{12}N_2$



NO. 047 Neuromodulator Molecule

2-(1H-Indol-3-yl)ethan-1-amine

InChI=1S/C10H12N2/c11-6-5-8-7-12-10-4-2-1-3-9(8)10/h1-4,7,12H,5-6,11H2

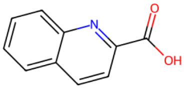
☺ C1=CC=C2C(=C1)C(=CN2)CCN

61-54-1

160 g/mol pH 10.5 – 9.3 Soluble

Quinaldic acid IP 4.0

$C_{10}H_7NO_2$



NO. 048 Pharmaceutical Chelating Molecule

Quinoline-2-carboxylic acid

InChI=1S/C10H7NO2/c12-10(13)9-6-5-7-3-1-2-4-8(7)11-9/h1-6H,(H,12,13)

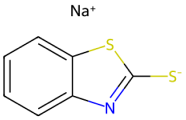
☺ C1=CC=C2C(=C1)C=CC(=N2)C(=O)O

93-10-7

173 g/mol pH 3.8 – 3.9 Soluble

Sodium 2-mercaptobenzothiazole IP 10.6

$C_7H_4NS_2Na$



NO. 049 Adhesive Biocide Molecule

Sodium 1,3-benzothiazole-2-thiolate

InChI=1S/C7H5NS2.Na/c9-7-8-5-3-1-2-4-6(5)10-7;/h1-4H,(H,8,9);/q;+1/p-1

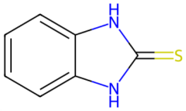
☺ C1=CC=C2C(=C1)N=C(S2)[S-].[Na+]

2492-26-4

189 g/mol pH 8.9 – 8.4 Soluble

2-mercaptobenzimidazole IP 12.5

$C_7H_6N_2S$



NO. 050 Metal Adsorbing Molecule

1,3-dihydrobenzimidazole-2-thione

InChI=1S/C7H6N2S/c10-7-8-5-3-1-2-4-6(5)9-7/h1-4H,(H2,8,9,10)

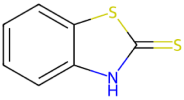
☺ C1=CC=C2C(=C1)NC(=S)N2

583-39-1

150 g/mol pH 5.6 – 5.8 Soluble

2-mercaptobenzothiazole IP 14.8

$C_7H_5NS_2$



NO. 051 Vulcanization Optimizer Molecule

3H-1,3-benzothiazole-2-thione

InChI=1S/C7H5NS2/c9-7-8-5-3-1-2-4-6(5)10-7/h1-4H,(H,8,9)

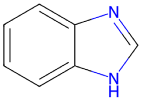
☺ C1=CC=C2C(=C1)NC(=S)S2

149-30-4

167 g/mol pH 4.9 – 4.9 Insoluble

Benzimidazole IP 1.5

$C_7H_6N_2$



NO. 052 Small Molecule Drug Foundation Molecule

1H-1,3-benzimidazole

InChI=1S/C7H6N2/c1-2-4-7-6(3-1)8-5-9-7/h1-5H,(H,8,9)

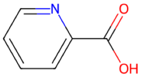
☺ C1=CC=C2C(=C1)NC=N2

51-17-2

118 g/mol pH 7.6 – 6.3 Soluble

2-picolinic acid IP -0.2

$C_6H_5NO_2$



NO. 053 Metal Ion Metabolism Molecule

Pyridine-2-carboxylic acid

InChI=1S/C6H5NO2/c8-6(9)5-3-1-2-4-7-5/h1-4H,(H,8,9)

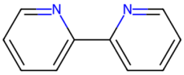
☺ C1=CC=NC(=C1)C(=O)O

98-98-6

123 g/mol pH 4.1 – 4.2 Soluble

2,2′-Bipyridine IP 4.5

$C_{10}H_8N_2$



NO. 054 Most Widely Used Ligand Molecule

2-pyridin-2-ylpyridine

InChI=1S/C10H8N2/c1-3-7-11-9(5-1)10-6-2-4-8-12-10/h1-8H

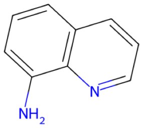
☺ C1=CC=NC(=C1)C2=CC=CC=N2

366-18-7

156 g/mol pH 6.6 – 6.4 Soluble

8-aminoquinoline IP 2.1

$C_9H_8N_2$



NO. 055 Anti Malaria Precursor Molecule

Quinolin-8-amine

InChI=1S/C9H8N2/c10-8-5-1-3-7-4-2-6-11-9(7)8/h1-6H,10H2

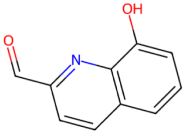
C1=CC2=C(C(=C1)N)N=CC=C2

578-66-5

144 g/mol pH 6.2 – 6.1 Soluble

8-hydroxy-2-quinolinecarboxaldehyde IP 9.0

$C_{10}H_7NO_2$



NO. 056 Fluorescent Probe Molecule

8-hydroxyquinoline-2-carbaldehyde

InChI=1S/C10H7NO2/c12-6-8-5-4-7-2-1-3-9(13)10(7)11-8/h1-6,13H

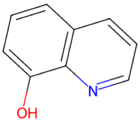
C1=CC2=C(C(=C1)O)N=C(C=C2)C=O

14510-06-6

173 g/mol pH 7.1 – 6.7 Soluble

8-hydroxyquinoline IP 11.6

C_9H_7NO



NO. 057 Trace Metal Detection Test Molecule

Quinolin-8-ol

InChI=1S/C9H7NO/c11-8-5-1-3-7-4-2-6-10-9(7)8/h1-6,11H

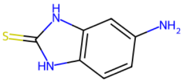
C1=CC2=C(C(=C1)O)N=CC=C2

148-24-3

145 g/mol pH 6.2 – 6.3 Soluble

5-amino-2-mercaptobenzimidazole IP 13.7

$C_7H_7N_3S$



NO. 058 Agrochemical Intermediate Molecule

5-amino-1,3-dihydrobenzimidazole-2-thione

InChI=1S/C7H7N3S/c8-4-1-2-5-6(3-4)10-7(11)9-5/h1-3H,8H2,(H2,9,10,11)

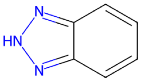
C1=CC2=C(C=C1N)NC(=S)N2

2618-66-8

165 g/mol pH 6.0 – 6.2 Soluble

Benzotriazole IP 7.8

$C_6H_5N_3$



NO. 059 Classical Corrosion Inhibitor Molecule

1H-1,2,3-Benzotriazole

InChI=1S/C6H5N3/c1-2-4-6-5(3-1)7-9-8-6/h1-4H,(H,7,8,9)

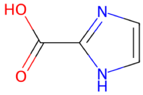
C1=CC2=NNN=C2C=C1

95-14-7

119 g/mol pH 5.7 – 5.8 Soluble

1-H-imidazole-2-carboxylic acid IP 8.7

$C_4H_4N_2O_2$



NO. 060 Enzyme Inhibitor Molecule

1-H-imidazole-2-carboxylic acid

InChI=1S/C4H4N2O2/c7-4(8)3-5-1-2-6-3/h1-2H,(H,5,6)(H,7,8)

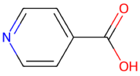
C1=CN=C(N1)C(=O)O

16042-25-4

112 g/mol pH 4.6 – 4.9 Soluble

Isonicotinic acid IP 0.4

$C_6H_5NO_2$



NO. 061 Tuberculosis Treatment Precursor Molecule

Pyridine-4-carboxylic acid

InChI=1S/C6H5NO2/c8-6(9)5-1-3-7-4-2-5/h1-4H,(H,8,9)

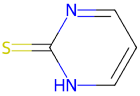
C1=CN=CC=C1C(=O)O

55-22-1

123 g/mol pH 3.8 – 4.0 Soluble

2-mercaptopyrimidine IP 15.5

$C_4H_4N_2S$



NO. 062 Allergen Molecule

1H-pyrimidine-2-thione

InChI=1S/C4H4N2S/c7-4-5-2-1-3-6-4/h1-3H,(H,5,6,7)

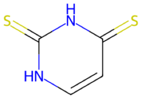
C1=CNC(=S)N=C1

1450-85-7

112 g/mol pH 5.0 – 5.4 Soluble

Dithiouracil IP 11.7

$C_4H_4N_2S_2$



NO. 063 Former Heart Medication Molecule

1H-pyrimidine-2,4-dithione

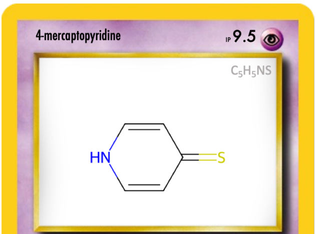
InChI=1S/C4H4N2S2/c7-3-1-2-5-4(8)6-3/h1-2H,(H2,5,6,7,8)

C1=CNC(=S)NC1=S

2001-93-6

144 g/mol pH 4.9 – 5.1 Insoluble

4-mercaptopyridine IP 9.5



NO. 064 Preservative Molecule

1H-pyridine-4-thione

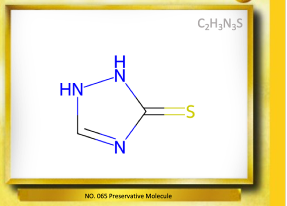
InChI=1S/C5H5NS/c7-5-1-3-6-4-2-5/h1-4H,(H,6,7)

C1=CNC=CC1=S

4556-23-4

111 g/mol pH 6.5 – 5.4 Soluble

1H-1,2,4-triazole-3-thiol IP 15.7



NO. 065 Preservative Molecule

1,2-dihydro-1,2,4-triazole-3-thione

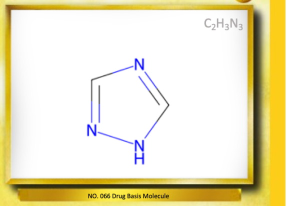
InChI=1S/C2H3N3S/c6-2-3-1-4-5-2/h1H,(H2,3,4,5,6)

C1=NC(=S)NN1

3179-31-5

101 g/mol pH 5.4 – 5.8 Soluble

1,2,4-triazole IP 1.7



NO. 066 Drug Basis Molecule

1H-1,2,4-triazole

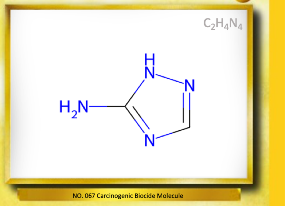
InChI=1S/C2H3N3/c1-3-2-5-4-1/h1-2H,(H,3,4,5)

C1=NC=NN1

288-88-0

69 g/mol pH 5.9 – 6.0 Soluble

3-amino-1,2,4-triazole IP 3.3



NO. 067 Carcinogenic Biocide Molecule

1H-1,2,4-triazol-5-amine

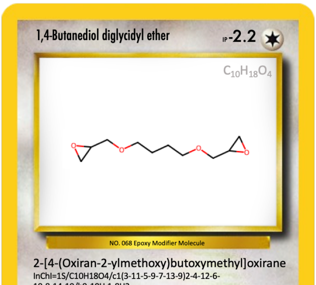
InChI=1S/C2H4N4/c3-2-4-1-5-6-2/h1H,(H3,3,4,5,6)

C1=NNC(=N1)N

61-82-5

84 g/mol pH 6.0 – 6.7 Soluble

1,4-Butanediol diglycidyl ether IP -2.2



NO. 068 Epoxy Modifier Molecule

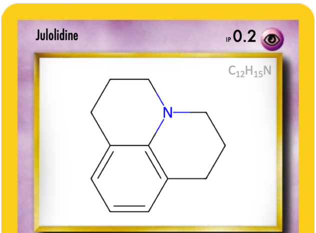
2-[4-(Oxiran-2-ylmethoxy)butoxymethyl]oxirane
InChI=1S/C10H18O4/c1(3-11-5-9-7-13-9)2-4-12-6-10-8-14-10/h9-10H,1-8H2

C1C(O1)COCCCCOCC2CO2

2425-79-8

202 g/mol pH 5.7 – 7.3 Soluble

Julolidine IP 0.2



NO. 069 Photochemistry Molecule

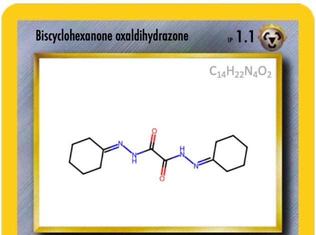
2,3,6,7-Tetrahydro-1H,5H-pyrido[3,2,1-ij]quinoline
InChI=1S/C12H15N/c1-4-10-6-2-8-13-9-3-7-11(5-1)12(10)13/h1,4-5H,2-3,6-9H2

C1CC2=C3C(=CC=C2)CCCN3C1

479-59-4

173 g/mol pH 7.5 – 7.5 Insoluble

Biscyclohexanone oxaldihydrazone IP 1.1



NO. 070 Demyelinating Copper Extracting Molecule

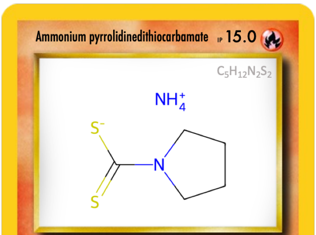
N,N'-bis(cyclohexylideneamino)oxamide
InChI=1S/C14H22N4O2/c19-13(17-15-11-7-3-1-4-8-11)14(20)18-16-12-9-5-2-6-10-12/h1-10H2,(H,17,19)(H,18,20)

C1CCC(=NNC(=O)C(=O)NN=C2CCCCC2)CC1

370-81-0

278 g/mol pH 5.7 – 5.7 Soluble

Ammonium pyrrolidinedithiocarbamate IP 15.0



NO. 071 Transition Metal Complexation Molecule

Azanium pyrrolidine-1-carbodithioate

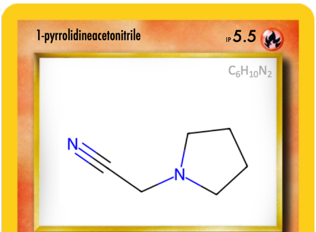
InChI=1S/C5H9NS2.H3N/c7-5(8)6-3-1-2-4-6;/h1-4H2,(H,7,8);1H3

C1CCN(C1)C(=S)[S-].[NH4+]

5108-96-3

164 g/mol pH 6.4 – 6.5 Soluble

1-pyrrolidineacetonitrile IP 5.5



NO. 072 Heterocycle Builder Molecule

2-pyrrolidin-1-ylacetonitrile

InChI=1S/C6H10N2/c7-3-6-8-4-1-2-5-8/h1-2,4-6H2

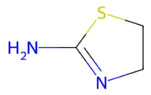
C1CCN(C1)CC#N

29134-29-0

110 g/mol pH 9.4 – 8.5 Soluble

2-amino-2-thiazoline IP 0.8

$C_3H_6N_2S$



NO. 073 Preservative Molecule

4,5-dihydro-1,3-thiazol-2-amine

InChI=1S/C3H6N2S/c4-3-5-1-2-6-3/h1-2H2,(H2,4,5)

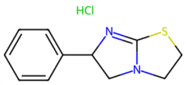
C1CSC(=N1)N

1779-81-3

102 g/mol pH 9.6 – 8.7 Soluble

Tetramisole hydrochloride IP 1.8

$C_{11}H_{13}N_2SCl$



NO. 074 Parasitic Worm Drug Molecule

6-phenyl-2,3,5,6-tetrahydroimidazo[2,1-b][1,3]thiazole

InChI=1S/C11H12N2S.ClH/c1-2-4-9(5-3-1)10-8-13-6-7-14-11(13)12-10;/h1-5,10H,6-8H2;1H

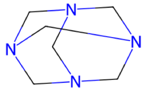
C1CSC2=NC(CN21)C3=CC=CC=C3.Cl

5086-74-8

241 g/mol pH 6.0 – 6.2 Soluble

Hexamethylenetetramine IP -1.9

$C_6H_{12}N_4$



NO. 075 Versatile Reagent Molecule

1,3,5,7-Tetraazaadamantane

InChI=1S/C6H12N4/c1-7-2-9-4-8(1)5-10(3-7)6-9/h1-6H2

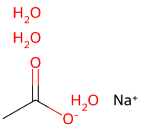
C1N2CN3CN1CN(C2)C3

100-97-0

140 g/mol pH 6.4 – 6.6 Soluble

Sodium acetate hydrate IP 1.2

$C_2H_5O_3Na$



NO. 076 Bacteria Culture Molecule

Sodium acetate hydrate

InChI=1S/C2H4O2.Na.H2O/c1-2(3)4;;/h1H3,(H,3,4);;1H2/q;+1;/p-1

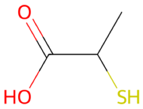
CC(=O)[O-].O.O.O.[Na+]

6131-90-4

144 g/mol pH 6.3 – 6.3 Soluble

Thiolactic acid IP 2.9

$C_3H_6O_2S$



NO. 077 Depilation Molecule

2-sulfanylpropanoic acid

InChI=1S/C3H6O2S/c1-2(6)3(4)5/h2,6H,1H3,(H,4,5)

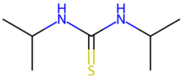
CC(C(=O)O)S

79-42-5

106 g/mol pH 3.1 – 3.2 Soluble

1,3-diisopropyl-2-thiourea IP 4.6

$C_7H_{16}N_2S$



NO. 078 Metal Halide Complex Synthesis Molecule

1,3-di(propan-2-yl)thiourea

InChI=1S/C7H16N2S/c1-5(2)8-7(10)9-6(3)4/h5-6H,1-4H3,(H2,8,9,10)

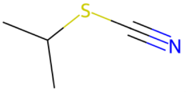
CC(C)NC(=S)NC(C)C

2986-17-6

160 g/mol pH 5.5 – 4.6 Soluble

Isopropyl thiocyanate IP 5.9

C_4H_7NS



NO. 079 Mercapto Building Block Molecule

Propan-2-yl thiocyanate

InChI=1S/C4H7NS/c1-4(2)6-3-5/h4H,1-2H3

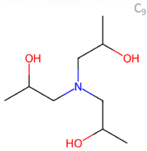
CC(C)SC#N

625-59-2

101 g/mol pH 7.3 – 7.3 Soluble

Triisopropanolamine IP -2.4

$C_9H_{21}NO_3$



NO. 080 Emulsifier Molecule

1,1′,1′′-Nitrilotri(propan-2-ol)

InChI=1S/C9H21NO3/c1-7(11)4-10(5-8(2)12)6-9(3)13/h7-9,11-13H,4-6H2,1-3H3

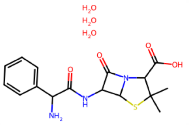
CC(CN(CC(C)O)CC(C)O)O

122-20-3

191 g/mol pH 9.5 – 8.5 Soluble

Ampicillin trihydrate IP 8.2

$C_{16}H_{25}N_3O_7S$



NO. 081 Aminopenicillin Antibiotic Molecule

Aminobenzylpenicillin trihydrate

InChI=1S/C16H19N3O4S.3H2O/c1-16(2)11(15(22)23)19-13(21)10(14(19)24-16)18-12(20)9(17)8-6-4-3-5-7-8;;;/h3-7,9-11,14H,17H2,1-2H3,(H,18,20)(H,22,23);3*1H2/t9-,10-,11+,14-;;;/m1.../s1

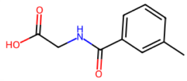
CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)C.O.O.O

7177-48-2

403 g/mol pH 4.9 – 5.0 Soluble

3-methylhippuric acid IP -1.5

$C_{10}H_{11}NO_3$



NO. 082 Xylene Exposure Biomarker Molecule

2-[(3-methylbenzoyl)amino]acetic acid

InChI=1S/C10H11NO3/c1-7-3-2-4-8(5-7)10(14)11-6-9(12)13/h2-5H,6H2,1H3,(H,11,14)(H,12,13)

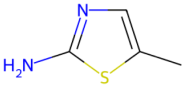
CC1=CC(=CC=C1)C(=O)NCC(=O)O

# 27115-49-7

193 g/mol pH 3.6 – 3.8 Soluble

2-amino-5-methylthiazole IP 6.4

$C_4H_6N_2S$



NO. 083 Antimicrobial Precursor Molecule

5-methyl-1,3-thiazol-2-amine

InChI=1S/C4H6N2S/c1-3-2-6-4(5)7-3/h2H,1H3,(H2,5,6)

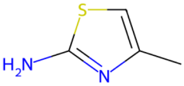
CC1=CN=C(S1)N

# 7305-71-7

114 g/mol pH 6.9 – 6.8 Soluble

2-amino-4-methylthiazole IP 2.6

$C_4H_6N_2S$



NO. 084 Vitamin B1 Precursor Molecule

4-methyl-1,3-thiazol-2-amine

InChI=1S/C4H6N2S/c1-3-2-7-4(5)6-3/h2H,1H3,(H2,5,6)

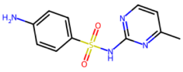
CC1=CSC(=N1)N

# 1603-91-4

114 g/mol pH 7.5 – 7.0 Soluble

Sulfamerazine IP 10.1

$C_{11}H_{12}N_4O_2S$



NO. 085 Sulfonamide Antibacterial Molecule

4-amino-N-(4-methylpyrimidin-2-yl)benzenesulfonamide

InChI=1S/C11H12N4O2S/c1-8-6-7-13-11(14-8)15-18(16,17)10-4-2-9(12)3-5-10/h2-7H,12H2,1H3,(H,13,14,15)

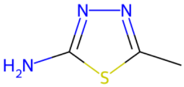
CC1=NC(=NC=C1)NS(=O)(=O)C2=CC=C(C=C2)N

# 127-79-7

264 g/mol pH 5.2 – 5.2 Insoluble

2-amino-5-methyl-1,3,4-thiadiazole IP -0.5

$C_3H_5N_3S$



NO. 086 Bioactivity Building Block Molecule

5-methyl-1,3,4-thiadiazol-2-amine

InChI=1S/C3H5N3S/c1-2-5-6-3(4)7-2/h1H3,(H2,4,6)

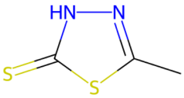
CC1=NN=C(S1)N

# 108-33-8

115 g/mol pH 7.0 – 6.3 Soluble

5-methyl-1,3,4-thiadiazole-2-thiol IP 3.6

$C_3H_4N_2S_2$



NO. 087 Agrochemical Building Block Molecule

5-methyl-3H-1,3,4-thiadiazole-2-thione

InChI=1S/C3H4N2S2/c1-2-4-5-3(6)7-2/h1H3,(H,5,6)

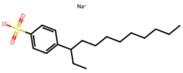
CC1=NNC(=S)S1

# 29490-19-5

132 g/mol pH 4.4 – 4.6 Insoluble

Sodium dodecyl benzenesulfonate IP -2.4

$C_{18}H_{30}O_3SNa$



NO. 088 Detergent Surfactant Molecule

Sodium n-dodecylbenzenesulfonate

InChI=1S/C18H30O3S.Na/c1-2-3-4-5-6-7-8-9-10-14-17-21-22(19,20)18-15-12-11-13-16-18;/h11-13,15-16H,2-10,14,17H2,1H3

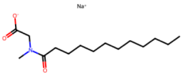
CCCCCCCCCC(CC)C1=CC=C(C=C1)S(=O)(=O)[O-].[Na+]

# 25155-30-0

349 g/mol pH 7.3 – 7.0 Soluble

Sodium lauroylsarcosine IP -1.1

$C_{15}H_{28}NO_3Na$



NO. 089 Foaming Agent Molecule

Sodium (N-methyldodecanamido)acetate

InChI=1S/C15H29NO3.Na/c1-3-4-5-6-7-8-9-10-11-12-14(17)16(2)13-15(18)19;/h3-13H2,1-2H3,(H,18,19);/q;+1/p-1

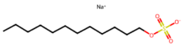
CCCCCCCCCCCC(=O)N(C)CC(=O)[O-].[Na+]

# 137-16-6

298 g/mol pH 5.5 – 6.1 Soluble

Sodium dodecylsulfate IP 0.1

$C_{12}H_{25}O_4SNa$



NO. 090 Versatile Surfactant Molecule

Sodium dodecyl sulfate

InChI=1S/C12H26O4S.Na/c1-2-3-4-5-6-7-8-9-10-11-12-16-17(13,14)15;/h2-12H2,1H3,(H,13,14,15);/q;+1/p-1

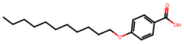
CCCCCCCCCCCCOS(=O)(=O)[O-].[Na+]

# 151-21-3

288 g/mol pH 5.8 – 5.9 Soluble

4-(undecyloxy)benzoic acid IP -1.0

$C_{18}H_{28}O_3$



NO. 091 Liquid Crystal Assembly Molecule

4-undecoxybenzoic acid

InChI=1S/C18H28O3/c1-2-3-4-5-6-7-8-9-10-15-21-17-13-11-16(12-14-17)18(19)20/h11-14H,2-10,15H2,1H3,(H,19,20)

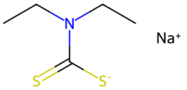
CCCCCCCCCCCOC1=CC=C(C=C1)C(=O)O

15872-44-3

292 g/mol pH 7.3 – 7.2 Insoluble

Sodium diethyldithiocarbomate IP 9.0

$C_5H_{10}NS_2Na$



NO. 092 Heavy Metal Poisoning Antidote Molecule

Sodium N,N-diethylcarbamodithioate

InChI=1S/C5H11NS2.Na/c1-3-6(4-2)5(7)8;/h3-4H2,1-2H3,(H,7,8);/q;+1/p-1

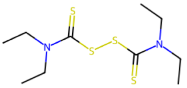
CCN(CC)C(=S)[S-].[Na+]

148-18-5

171 g/mol pH 8.1 – 8.2 Soluble

Tetraethylthiuram disulfide IP 9.6

$C_{10}H_{20}N_2S_4$



NO. 093 Alcohol Deterrent Molecule

Diethylcarbamothioylsulfanyl N,N-diethylcarbamodithioate

InChI=1S/C10H20N2S4/c1-5-11(6-2)9(13)15-16-10(14)12(7-3)8-4/h5-8H2,1-4H3

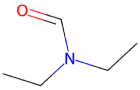
CCN(CC)C(=S)SSC(=S)N(CC)CC

97-77-8

296 g/mol pH 5.9 – 5.1 Insoluble

Diethylformamide IP -1.6

$C_5H_{11}NO$



NO. 094 MOF Production Molecule

N,N-diethylformamide

InChI=1S/C5H11NO/c1-3-6(4-2)5-7/h5H,3-4H2,1-2H3

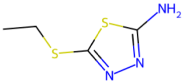
CCN(CC)C=O

617-84-5

144 g/mol pH 7.0 – 7.4 Soluble

2-amino-5-ethylthio-1,3,4-thiadiazole IP 1.7

$C_7H_5O_2Na$



NO. 095 Bioactive Intermediate Molecule

5-ethylsulfanyl-1,3,4-thiadiazol-2-amine

InChI=1S/C4H7N3S2/c1-2-8-4-7-6-3(5)9-4/h2H2,1H3,(H2,5,6)

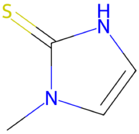
CCSC1=NN=C(S1)N

25660-70-2

161 g/mol pH 4.5 – 5.1 Soluble

2-mercapto-1-methylimidazole IP 6.3

$C_4H_6N_2S$



NO. 096 Hyperthyroidism Treatment Molecule

3-methyl-1H-imidazole-2-thione

InChI=1S/C4H6N2S/c1-6-3-2-5-4(6)7/h2-3H,1H3,(H,5,7)

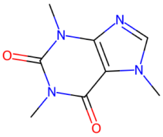
CN1C=CNC1=S

60-56-0

114 g/mol pH 6.5 – 6.5 Soluble

Caffeine IP -1.0

$C_8H_{10}N_4O_2$



NO. 097 Coffee Molecule

1,3,7-trimethylpurine-2,6-dione

InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

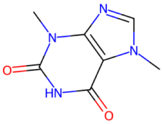
CN1C=NC2=C1C(=O)N(C(=O)N2C)C

58-08-2

194 g/mol pH 6.0 – 6.4 Soluble

Theobromine IP -0.2

$C_7H_8N_4O_2$



NO. 098 Chocolate Molecule

3,7-dimethylpurine-2,6-dione

InChI=1S/C7H8N4O2/c1-10-3-8-5-4(10)6(12)9-7(13)11(5)2/h3H,1-2H3,(H,9,12,13)

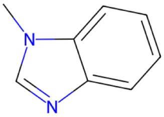
CN1C=NC2=C1C(=O)NC(=O)N2C

83-67-0

180 g/mol pH 5.5 – 5.4 Soluble

1-methylbenzimidazole IP -1.4

C_8H_8N



NO. 099 Solar Cell Additive Molecule

3-amino-1H-1,2,4-triazole-5-carboxylic acid

InChI=1S/C8H8N2/c1-10-6-9-7-4-2-3-5-8(7)10/h2-6H,1H3

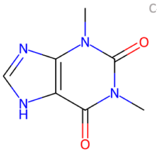
CN1C=NC2=CC=CC=C21

1632-83-3

132 g/mol pH 5.0 – 5.6 Soluble

Theophylline IP 1.4

$C_7H_8N_4O_2$



NO. 100 Tea Molecule

1,3-dimethyl-7H-purine-2,6-dione

InChI=1S/C7H8N4O2/c1-10-5-4(8-3-9-5)6(12)11(2)7(10)13/h3H,1-2H3,(H,8,9)

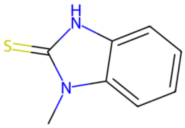
☺ CN1C2=C(C(=O)N(C1=O)C)NC=N2

58-55-9

180 g/mol pH 6.4 – 6.1 Soluble

1-methyl-1-H-benzimidazole-2-thiol IP 9.4

$C_8H_8N_2S$



NO. 101 Catalysis Coordination Molecule

3-methyl-1H-benzimidazole-2-thione

InChI=1S/C8H8N2S/c1-10-7-5-3-2-4-6(7)9-8(10)11/h2-5H,1H3,(H,9,11)

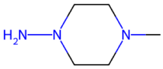
☺ CN1C2=CC=CC=C2NC1=S

2360-22-7

164 g/mol pH 7.2 – 6.7 Soluble

1-amino-4-methylpiperazine IP 0.2

$C_5H_{13}N_3$



NO. 0102 Anti-infection Drug Precursor Molecule

4-methylpiperazin-1-amine

InChI=1S/C5H13N3/c1-7-2-4-8(6)5-3-7/h2-6H2,1H3

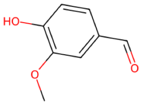
☺ CN1CCN(CC1)N

6928-85-4

115 g/mol pH 8.9 – 7.9 Soluble

Vanillin IP 3.8

$C_8H_8O_3$



NO. 103 Vanilla Molecule

4-hydroxy-3-methoxybenzaldehyde

InChI=1S/C8H8O3/c1-11-8-4-6(5-9)2-3-7(8)10/h2-5,10H,1H3

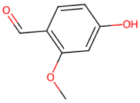
☺ COC1=C(C=CC(=C1)C=O)O

121-33-5

152 g/mol pH 4.9 – 5.1 Soluble

4-hydroxy-2-methoxybenzaldehyde IP 1.3

$C_8H_8O_3$



NO. 104 Fragrance Molecule

4-hydroxy-2-methoxybenzaldehyde

InChI=1S/C8H8O3/c1-11-8-4-7(10)3-2-6(8)5-9/h2-5,10H,1H3

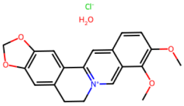
☺ COC1=C(C=CC(=C1)O)C=O

18278-34-7

152 g/mol pH 5.6 – 5.9 Soluble

Berberine chloride IP 0.1

$C_{20}H_{18}NO_4Cl$



NO. 105 Anti Inflammatory Chinese Herb Molecule

Berberine chloride

InChI=1S/C20H18NO4.ClH/c1-22-17-4-3-12-7-16-14-9-19-18(24-11-25-19)8-13(14)5-6-21(16)10-15(12)20(17)23-2;/h3-4,7-10H,5-6,11H2,1-2H3;1H/q+1;/p-1

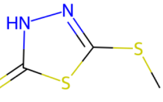
☺ COC1=C(C2=C[N+]3=C(C=C2C=C1)C4=CC5=C(C=C4CC3)OCO5)OC.O.[Cl-]

633-65-8

372 g/mol pH 5.3 – 6.1 Insoluble

5-methylthio-1,3,4-thiadiazole-2-thiol IP 0.9

$C_3H_4N_2S_3$



NO. 106 Pharmaceutical Precursor Molecule

5-methylsulfanyl-3H-1,3,4-thiadiazole-2-thione

InChI=1S/C3H4N2S3/c1-7-3-5-4-2(6)8-3/h1H3,(H,4,6)

☺ CSC1=NNC(=S)S1

6264-40-0

164 g/mol pH 4.1 – 4.2 Soluble

C

PYTHON LIBRARY FOR MACHINE LEARNING MODEL TRAINING AND ANALYSIS

The library created for the machine learning methodology of Chapter 5 is presented below.

```
import os
import pandas as pd
import datamol as dm
import numpy as np
import pandas as pd
import warnings
import os
import seaborn as sns
from collections import Counter
import ast
import time
import pickle

#rdkit
from rdkit import Chem
from rdkit.Chem import Descriptors
from rdkit.ML.Descriptors import MoleculeDescriptors

#sci-kit learn
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.feature_selection import VarianceThreshold
from sklearn.preprocessing import StandardScaler, MinMaxScaler,
    ↪ PowerTransformer
from sklearn.model_selection import RandomizedSearchCV, cross_val_score,
    ↪ KFold
from sklearn.decomposition import PCA
from sklearn.metrics import mean_squared_error, mean_absolute_error,
    ↪ r2_score
from sklearn.svm import SVR
from sklearn.model_selection import learning_curve

# matplotlib
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import seaborn as sns
import umap

# other specific imports
from probatus.feature_elimination import ShapRFECV
from verstack.stratified_continuous_split import scsplit
from bayes_opt import BayesianOptimization
from xgboost import XGBRegressor


```

```

def process_string(input_string):
    # Split a string by backslash and return the last item without its
    ↪ last 4 characters.
    parts = input_string.split(’\\’)

    # Extract the last item and remove its last 4 characters
    last_item = parts[-1]
    processed_string = last_item[:-4]

    return processed_string

def combine_strings_with_underscore(string_a, string_b):
    # Combine two strings with an underscore between them.

    combined_string = f"{string_a}_WITH_{string_b}"
    return combined_string

def preprocessing(df, fill_nan = True, var_drop = True, corr_drop = True,
    ↪ scale = False, scaler = ’min_max’):
    # Extract the "SMILES" column
    smiles_column = df[’SMILES’]

    # Drop the "SMILES" column for further processing
    df = df.drop(’SMILES’, axis=1)

    if fill_nan is True:
        # Fill missing values with column medians
        df = df.fillna(df.median())

    if var_drop is True:
        # Drop features with low variance, < 0.1
        vt = VarianceThreshold(threshold=0.1)
        df = pd.DataFrame(vt.fit_transform(df), columns=df.columns[vt.
    ↪ get_support()])

    if corr_drop is True:
        # Drop features with highly correlated features, > 0.8
        correlated_features = set()
        correlation_matrix = df.corr()
        for i in range(len(correlation_matrix.columns)):
            for j in range(i):
                if abs(correlation_matrix.iloc[i, j]) > 0.8:
                    colname = correlation_matrix.columns[i]
                    correlated_features.add(colname)
        df = df.drop(correlated_features, axis=1)

    if scale is True:
        # # Scale features:
```

```
        if scaler == 'min_max':
            scaler_func = MinMaxScaler()
        elif scaler == 'standard':
            scaler_func = StandardScaler()
        elif scaler == 'power':
            scaler_func = PowerTransformer()
        df[:] = scaler_func.fit_transform(df).tolist()

    # Add the "SMILES" column back to the left side of the DataFrame
    df = pd.concat([smiles_column, df], axis=1)

    return df

def stratified_split(df, target_column, test_size=0.11, random_state=42, return_dataset = False):
    df_stratified = df.copy()
    train, test = scsplit(df_stratified, stratify = df_stratified[target_column],
                        test_size=test_size,
                        train_size = 1-test_size,
                        random_state=42)

    # add a column to the original dataframe to indicate whether the row is in the train or test set
    df_stratified['split'] = 'validation'
    df_stratified.loc[df_stratified.index.isin(train.index), 'split'] = 'train'

    df_stratified.groupby(by="split")[target_column].plot.density(legend=True)

    # record the indices of the train and test sets
    train_idx = df_stratified[df_stratified['split'] == 'train'].index
    test_idx = df_stratified[df_stratified['split'] == 'validation'].index

    print(f'Train set size: {len(train_idx)}')
    print(f'Validation set size: {len(test_idx)}')

    # sanity check to make sure no items are shared between the train and test sets
    common_elements = train_idx.intersection(test_idx)
    if len(common_elements) == 0:
        print("No common elements between train and validation sets.")
    else:
        print("There are common elements between train and validation sets.")

    if return_dataset:
        return df_stratified, train_idx, test_idx
    else:
        return train_idx, test_idx

def visualize_chemspace(df_target: pd.DataFrame, df_features: pd.DataFrame
    ↪ , split_names, target_column, mol_col = "SMILES", size_col=None,
    ↪ output_file=None):
    features = df_features.copy().drop(columns=[mol_col])
    target = df_target.copy()

    # embedding of the feature space into 2 UMAP dimensions
    embedding = umap.UMAP(random_state=42).fit_transform(features)
    target["UMAP_0"], target["UMAP_1"] = embedding[:, 0], embedding[:, 1]

    # Save each plot as a separate file
    for i, split_name in enumerate(split_names):
        plt.figure()
        sns.scatterplot(data=target, x="UMAP_0", y="UMAP_1", style=
    ↪ split_name, hue=target_column, alpha=0.7, palette="viridis", legend
    ↪ ="brief")
        plt.title(f"UMAP Embedding of compounds for {split_name}")
        plt.legend(loc=’upper left’, bbox_to_anchor=(1, 1))
        plt.tight_layout()
        if output_file:
            plt.savefig(output_file)
        plt.close()

def recursive_feature_elimination(descriptor_df, target_df, num_features
    ↪ =10, selection_steps=100, random_seed=42):
    """
    Perform Recursive Feature Elimination (RFE) using a Random Forest
    ↪ Regressor for feature selection.

    Parameters:
    - descriptor_df (pd.DataFrame): DataFrame containing molecular
    ↪ descriptors, including the "SMILES" column.
    - target_df (pd.DataFrame): DataFrame containing the target variable
    ↪ for regression.
    - num_features (int): Number of features to select.
    - selection_steps (int): Number of random selections to perform for
    ↪ stability.
    - random_seed (int): Seed for reproducibility.

    Returns:
    - selected_df (pd.DataFrame): DataFrame containing the selected
    ↪ features along with the "SMILES" column.
    """

    # Extract the "SMILES" column
    smiles_column = descriptor_df[’SMILES’]

    # Exclude the first column with SMILES from calculations
    descriptor_df = descriptor_df.iloc[:, 1:]
    target_df = target_df.iloc[:, 1:]

    selected_features = []

    np.random.seed(random_seed)
    random_ints = np.random.randint(low=1000, size=selection_steps)

    for i in random_ints:
        rfe = RFE(estimator=RandomForestRegressor(random_state=i, n_jobs
    ↪ =-1), n_features_to_select=num_features)
        rfe.fit_transform(descriptor_df, target_df.values.ravel())
        selected_features.append(str(list(rfe.get_support(indices=True))))

    data = Counter(selected_features)
    features = data.most_common(1)[0][0]
    selected_features = ast.literal_eval(features)

    selected_df = descriptor_df.iloc[:, selected_features]
    selected_df = pd.concat([smiles_column, selected_df], axis=1)

    print('The following descriptors have been selected:', list(
    ↪ selected_df.columns)[1:])
    return selected_df

def shapley_feature_elimination(descriptor_df, target_df, disp_fig = True,
    ↪   reporting = False):

    # Extract the "SMILES" column
    smiles_column = descriptor_df['SMILES']

    # Exclude the first column with SMILES from calculations
    descriptor_df = descriptor_df.iloc[:, 1:]
    target_df = target_df.iloc[:, 1:]

    # Prepare model and parameter search space
    clf = RandomForestRegressor(random_state=42)

    param_grid = {
            'n_estimators': [10, 100, 1000],
            'max_depth': [None, 2, 5, 10, 20],
            'min_samples_split': [2, 5, 10],
            'min_samples_leaf': [1, 2, 4, 8],
        }

    search = RandomizedSearchCV(clf, param_grid)

    # Run feature elimination
    shap_elimination = ShapRFECV(
        clf=search, step=0.1, cv=5, scoring='neg_root_mean_squared_error', n_jobs=-1, random_state=42)
    report = shap_elimination.fit_compute(descriptor_df, target_df.to_numpy().ravel(), check_additivity=False)

    # Get features with best performance, no matter how many features
    best_features = list(shap_elimination.get_reduced_features_set(num_features="best", return_type="feature_names"))
    print(f"The selected {len(best_features)} best features are: {best_features}")

    if disp_fig is True:
        performance_plot = shap_elimination.plot()

    # save
    selected_df = descriptor_df[best_features]
    df_best_features = pd.concat([smiles_column, selected_df], axis=1)

    if reporting is True:
        return df_best_features, report
    return df_best_features

# Analyze correlations between features and target

def pearson_correlations(descriptor_df, target_df):
    merged_df = pd.merge(descriptor_df, target_df, on='SMILES')

    # Extract features and target values
    X = merged_df.iloc[:, 1:-1]  # Assuming features start from the second column
    y = merged_df.iloc[:, -1]

    corr_matrix = merged_df.corr(numeric_only=True)
    top_10 = corr_matrix.iloc[:, -1].sort_values(ascending=False)[1:11]
    bottom_10 = corr_matrix.iloc[:, -1].sort_values(ascending=True)[0:10]

    print("Pearson Correlations")
    print("Top 10:")
    print(top_10)
    print(" ")
    print("Bottom 10:")
    print(bottom_10)

    return corr_matrix

from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.decomposition import PCA
```

```

def perform_pca(target_df, feature_df, scale=False):
    """
    Perform PCA analysis on feature dataset and generate relevant plots.

    Parameters:
    - target_df: DataFrame with ’SMILES’ and target column
    - feature_df: DataFrame with ’SMILES’ and feature columns

    Returns:
    - pca_result: DataFrame containing PCA results
    """

    # Merge the target and feature datasets on ’SMILES’
    merged_df = pd.merge(target_df, feature_df, on=’SMILES’, how=’inner’)

    # Extract the target column for separate analysis
    target_column = merged_df.columns[1]

    # Extract feature columns for PCA
    feature_columns = merged_df.columns[2:]

    if scale is True:
        # Scale the feature columns
        feature_data = MinMaxScaler().fit_transform(merged_df[
    ↪ feature_columns])
    else:
        feature_data = merged_df[feature_columns]

    # Perform PCA
    pca = PCA()
    principal_components = pca.fit_transform(feature_data)

    # Create a DataFrame with PCA results
    pca_result = pd.DataFrame(data=principal_components, columns=[f’PC{i}’
    ↪  for i in range(1, pca.n_components_ + 1)])
    pca_result[’SMILES’] = merged_df[’SMILES’]
    pca_result[target_column] = merged_df[target_column]

    return pca_result, pca

def plot_pca_results(fig_type, pca_result, pca, target_column, directory,
    ↪ threshold=0.):
    # Plot Explained Variance Ratio, 2D PCA, and 3D PCA in the same
    ↪ subplot
    fig = plt.figure(figsize=(15, 18))

    # Explained Variance Ratio
    plt.subplot(3, 2, (1,2))
    plt.bar(range(1, len(pca.explained_variance_ratio_) + 1), pca.
    ↪ explained_variance_ratio_)
    plt.xlabel('Principal Components')
    plt.ylabel('Explained Variance Ratio')
    plt.title('Explained Variance Ratio')

    # 2D PCA
    plt.subplot(3, 2, 3)
    plt.scatter(pca_result['PC1'], pca_result['PC2'], c=pca_result[
    ↪ target_column], cmap='viridis')
    plt.title('2D PCA Continous Target')

    plt.subplot(3, 2, 4)
    plt.scatter(pca_result['PC1'][pca_result[target_column] > threshold],
                pca_result['PC2'][pca_result[target_column] > threshold],
    ↪ c='red', label=f'{target_column} > {threshold}')
    plt.scatter(pca_result['PC1'][pca_result[target_column] <= threshold],
                pca_result['PC2'][pca_result[target_column] <= threshold],
    ↪  c='blue', label=f'{target_column} <= {threshold}')
    plt.legend()
    plt.title(f'2D PCA Categorical Target')

    # 3D PCA
    ax1 = fig.add_subplot(3, 2, 5, projection='3d')
    scatter_3d = ax1.scatter(pca_result['PC1'], pca_result['PC2'],
    ↪ pca_result['PC3'], c=pca_result[target_column], cmap='viridis')
    cbar = plt.colorbar(scatter_3d, ax=plt.gca())
    cbar.set_label(target_column)
    ax1.set_title('3D PCA Continous Target')

    ax2 = fig.add_subplot(3, 2, 6, projection='3d')
    ax2.scatter(pca_result['PC1'][pca_result[target_column] > threshold],
                pca_result['PC2'][pca_result[target_column] > threshold],
                pca_result['PC3'][pca_result[target_column] > threshold],
    ↪ c='red', label=f'{target_column} > {threshold}')
    ax2.scatter(pca_result['PC1'][pca_result[target_column] <= threshold],
                pca_result['PC2'][pca_result[target_column] <= threshold],
                pca_result['PC3'][pca_result[target_column] <= threshold],
    ↪  c='blue', label=f'{target_column} <= {threshold}')
    ax2.legend()
    ax2.set_title(f'3D PCA Categorical Target')

    # save figure
    fig_name = os.path.join(directory, fig_type)
    plt.savefig(f'{fig_name}_PCA.png')
    # plt.show()

def process_pca(fig_type, target_df, feature_df, target_column,
    ↪ save_directory, threshold=0.):
```

```
    result, pca = perform_pca(target_df, feature_df)

    plot_pca_results(fig_type, result, pca, target_column, directory =
    ↪ save_directory,
                     threshold=threshold # change this as needed
                     )
    return result, pca

def plot_histograms(df, save_directory):

    # Identify numeric columns
    numeric_cols = df.select_dtypes(include=[’number’]).columns

    # Set the number of subplot rows and columns
    n_rows = len(numeric_cols) // 2 + len(numeric_cols) % 2
    n_cols = 5 if len(numeric_cols) > 1 else 1

    # Create a figure to hold the subplots
    fig, axes = plt.subplots(n_rows, n_cols, figsize=(n_cols*5, n_rows*5))

    # Flatten axes array if more than one subplot
    if len(numeric_cols) > 1:
        axes = axes.flatten()
    else:
        axes = [axes]

    # Generate histograms
    for ax, col in zip(axes, numeric_cols):
        df[col].hist(ax=ax, bins=10, grid=False)
        # ax.set_title(f’Histogram of {col}’)
        ax.set_xlabel(col)
        ax.set_ylabel(’Frequency’)

    # Hide unused subplots
    for ax in axes[len(numeric_cols):]:
        ax.set_visible(False)

    # Adjust layout
    plt.tight_layout()

    # Save the figure
    plt.savefig(os.path.join(save_directory, ’feature_histograms.png’))

def learning_curves(estimator, X, y, scoring = ’
    ↪ neg_root_mean_squared_error’):
    train_sizes, train_scores, test_scores = learning_curve(
        estimator, X, y, train_sizes=np.linspace(0.1, 1.0, 40), cv=10,
        scoring=scoring, shuffle=True, n_jobs=-1)
    return train_sizes, train_scores, test_scores
```

```
def optimized_random_forest_model(X_train, y_train, X_test, y_test, shuffl
    ↪ =False):
    # Define the model
    def init_model(params):
        model = RandomForestRegressor(n_estimators=int(params['
    ↪ n_estimators']),
                                      max_depth=int(params['max_depth']),
                                      min_samples_split=int(params['
    ↪ min_samples_split']),
                                      max_features=params['max_features'],
                                      random_state=42,
                                      n_jobs=-1)
        return model

    start = time.time()
    # Define the objective function for optimization
    def rf_bo(n_estimators, max_depth, min_samples_split, max_features):
        params = {'n_estimators': n_estimators, 'max_depth': max_depth,
                  'min_samples_split': min_samples_split, 'max_features':
    ↪ max_features}
        model = init_model(params)
        cv = KFold(n_splits=10, shuffle=shuffl, random_state=42 if shuffl
    ↪ else None)
        return cross_val_score(model, X_train, y_train, cv=cv, scoring="
    ↪ neg_root_mean_squared_error", n_jobs=-1).mean()

    # Define the bounds for hyperparameters
    param_bounds = {'n_estimators': (10, 1000), 'max_depth': (1, 50), '
    ↪ min_samples_split': (2, 25), 'max_features': (0.1, 0.999)}

    # Run Bayesian Optimization
    optimizer = BayesianOptimization(f=rf_bo, pbounds=param_bounds,
    ↪ random_state=42, verbose=0, allow_duplicate_points=True)
    optimizer.maximize(init_points=40, n_iter=60)

    # Train and test for the best hyperparameters
    best_hyp_model = init_model(optimizer.max['params'])
    train_sizes, train_scores, test_scores = learning_curves(
    ↪ best_hyp_model, X_train, y_train)
    learning_curve_metrics = {'train_sizes': train_sizes, 'train_scores':
    ↪ train_scores, 'test_scores': test_scores}

    # Fit model for validation predictions
    final_model = best_hyp_model.fit(X_train, y_train)

    # Predictions
    y_fit = final_model.predict(X_train)
    y_pred = final_model.predict(X_test)
```

```python

    # Metrics from the final model
    metrics = {
        'r2_train': r2_score(y_train, y_fit),
        'rmse_train': mean_squared_error(y_train, y_fit, squared=False),
        'mae_train': mean_absolute_error(y_train, y_fit),
        'r2_test': r2_score(y_test, y_pred),
        'rmse_test': mean_squared_error(y_test, y_pred, squared=False),
        'mae_test': mean_absolute_error(y_test, y_pred),
        'q2_test': 1 - mean_squared_error(y_test, y_pred) / np.var(y_test)
    }

    # Print metrics
    for metric, value in metrics.items():
        print(f"{metric}: {round(value, 2)}")

    print(f'\nOptimization took {round((time.time() - start)/60, 1)} minutes')

    return final_model, metrics, learning_curve_metrics

def optimized_svm_model(X_train, y_train, X_test, y_test, shuffl=False):
    def init_model(params):
        # Exponentiate parameters to transform them back from log scale to original scale
        # log is necessary for searches in parameter spaces more evenly across vastly different scales, such as between 0.01-0.1 and 1-100
        model = SVR(C=np.exp(params['log_C']), epsilon=np.exp(params['log_epsilon']), gamma=np.exp(params['log_gamma']))
        return model

    start = time.time()
    # Define the objective function for optimization using transformed parameters
    def svm_bo(log_C, log_epsilon, log_gamma):
        params = {'log_C': log_C, 'log_epsilon': log_epsilon, 'log_gamma': log_gamma}
        model = init_model(params)
        cv = KFold(n_splits=10, shuffle=shuffl, random_state=42 if shuffl else None)
        return cross_val_score(model, X_train, y_train, cv=cv, scoring="neg_root_mean_squared_error", n_jobs=-1).mean()

    # Bounds for hyperparameters, now on a log scale
    param_bounds = {
        'log_C': (np.log(0.001), np.log(1000)),
        'log_epsilon': (np.log(0.001), np.log(10)),
        'log_gamma': (np.log(0.0001), np.log(100))
    }

    # Run Bayesian Optimization
    optimizer = BayesianOptimization(f=svm_bo, pbounds=param_bounds,
    ↪ random_state=42, verbose=0, allow_duplicate_points=True)
    optimizer.maximize(init_points=40, n_iter=60)

    # Train and test for the best hyperparameters
    best_hyp_model = init_model(optimizer.max['params'])
    train_sizes, train_scores, test_scores = learning_curves(
    ↪ best_hyp_model, X_train, y_train)
    learning_curve_metrics = {'train_sizes': train_sizes, 'train_scores':
    ↪ train_scores, 'test_scores': test_scores}

    # Fit model for validation predictions
    final_model = best_hyp_model.fit(X_train, y_train)

    # Predictions
    y_fit = final_model.predict(X_train)
    y_pred = final_model.predict(X_test)

    # Metrics from the final model
    metrics = {
        'r2_train': r2_score(y_train, y_fit),
        'rmse_train': mean_squared_error(y_train, y_fit, squared=False),
        'mae_train': mean_absolute_error(y_train, y_fit),
        'r2_test': r2_score(y_test, y_pred),
        'rmse_test': mean_squared_error(y_test, y_pred, squared=False),
        'mae_test': mean_absolute_error(y_test, y_pred),
        'q2_test': 1 - mean_squared_error(y_test, y_pred) / np.var(y_test)
    }

    # Print metrics
    for metric, value in metrics.items():
        print(f"{metric}: {round(value, 2)}")

    print(f'\nOptimization took {round((time.time() - start)/60, 1)}
    ↪ minutes')

    return final_model, metrics, learning_curve_metrics

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import PolynomialFeatures

def optimized_knn_model(X_train, y_train, X_test, y_test, shuffl=False):
    def init_model(params):
        # Create a pipeline with polynomial features, and KNN
        model = Pipeline([
            ('poly', PolynomialFeatures(degree=2, include_bias=False)),
            ('knn', KNeighborsRegressor(n_neighbors=int(params['
     ↪ n_neighbors']),
                                        weights=params['weights'],
                                        metric=params['metric']))
        ])
        return model

    start = time.time()

    # Define the objective function for optimization
    def knn_bo(n_neighbors, weights, metric):
        # Map continuous weights to discrete values
        weights_options = ['uniform', 'distance']
        metric_options = ['euclidean', 'manhattan', 'minkowski']
        weights_mapped = weights_options[int(weights)]
        metric_mapped = metric_options[int(metric)]

        params = {'n_neighbors': n_neighbors, 'weights': weights_mapped, '
     ↪ metric': metric_mapped}
        model = init_model(params)
        cv = KFold(n_splits=10, shuffle=shuffl, random_state=42 if shuffl
     ↪ else None)
        return cross_val_score(model, X_train, y_train, cv=cv, scoring="
     ↪ neg_root_mean_squared_error", n_jobs=-1).mean()

    # Bounds for hyperparameters
    param_bounds = {
        'n_neighbors': (1, 10),  # Assuming a reasonable max number of
     ↪ neighbors
        'weights': (0, 1),  # 0 for 'uniform', 1 for 'distance'
        'metric': (0, 1)  # 0 for 'euclidean', 1 for 'manhattan'
    }

    # Run Bayesian Optimization
    optimizer = BayesianOptimization(f=knn_bo, pbounds=param_bounds,
     ↪ random_state=42, verbose=0, allow_duplicate_points=True)
    optimizer.maximize(init_points=40, n_iter=60)

    # Map optimized parameters back to their discrete values
    optimized_params = optimizer.max['params']
    optimized_params['weights'] = ['uniform', 'distance'][int(
     ↪ optimized_params['weights'])]
    optimized_params['metric'] = ['euclidean', 'manhattan', 'minkowski'][
     ↪ int(optimized_params['metric'])]

    # Train and test for the best hyperparameters
    best_hyp_model = init_model(optimized_params)
    train_sizes, train_scores, test_scores = learning_curves(
     ↪ best_hyp_model, X_train, y_train)
    learning_curve_metrics = {'train_sizes': train_sizes, 'train_scores':
    ↪ train_scores, 'test_scores': test_scores}

    # Fit model for validation predictions
    final_model = best_hyp_model.fit(X_train, y_train)

    # Predictions
    y_fit = final_model.predict(X_train)
    y_pred = final_model.predict(X_test)

    # Metrics from the final model
    metrics = {
        'r2_train': r2_score(y_train, y_fit),
        'rmse_train': mean_squared_error(y_train, y_fit, squared=False),
        'mae_train': mean_absolute_error(y_train, y_fit),
        'r2_test': r2_score(y_test, y_pred),
        'rmse_test': mean_squared_error(y_test, y_pred, squared=False),
        'mae_test': mean_absolute_error(y_test, y_pred),
        'q2_test': 1 - mean_squared_error(y_test, y_pred) / np.var(y_test)
    }

    # Print metrics
    for metric, value in metrics.items():
        print(f"{metric}: {round(value, 2)}")

    print(f'\nOptimization took {round((time.time() - start)/60, 1)}
    ↪ minutes')

    return final_model, metrics, learning_curve_metrics

def optimized_xgb_model(X_train, y_train, X_test, y_test, shuffl=False):
    def init_model(params):
        model = XGBRegressor(
            n_estimators=int(params['n_estimators']),
            max_depth=int(params['max_depth']),
            learning_rate=params['learning_rate'],
            subsample=params['subsample'],
            colsample_bytree=params['colsample_bytree'],
            gamma=params['gamma'],
            reg_alpha=np.exp(params['log_reg_alpha']),
            reg_lambda=np.exp(params['log_reg_lambda']),
            random_state=42,
            n_jobs=-1
        )
        return model

    start = time.time()

    # Define the objective function for optimization
    def xgb_bo(n_estimators, max_depth, learning_rate, subsample,
    ↪ colsample_bytree, gamma, log_reg_alpha, log_reg_lambda):
        params = {
            'n_estimators': n_estimators,
            'max_depth': max_depth,
            'learning_rate': learning_rate,
            'subsample': subsample,
            'colsample_bytree': colsample_bytree,
            'gamma': gamma,
            'log_reg_alpha': log_reg_alpha,
            'log_reg_lambda': log_reg_lambda
        }
        model = init_model(params)
        cv = KFold(n_splits=10, shuffle=shuffl, random_state=42 if shuffl
    ↪ else None)
        return cross_val_score(model, X_train, y_train, cv=cv, scoring="
    ↪ neg_root_mean_squared_error", n_jobs=-1).mean()

    # Bounds for hyperparameters
    param_bounds = {
        'n_estimators': (100, 1000),
        'max_depth': (2, 10),
        'learning_rate': (0.01, 0.1),
        'subsample': (0.1, 1.0),
        'colsample_bytree': (0.1, 1.0),
        'gamma': (0.1, 1.0),
        'log_reg_alpha': (-3, 2),
        'log_reg_lambda': (-3, 2)
    }

    # Run Bayesian Optimization
    optimizer = BayesianOptimization(f=xgb_bo, pbounds=param_bounds,
    ↪ random_state=42, verbose=0, allow_duplicate_points=True)
    optimizer.maximize(init_points=40, n_iter=60)

    # Train and test for the best hyperparameters
    best_hyp_model = init_model(optimizer.max['params'])
    train_sizes, train_scores, test_scores = learning_curves(
    ↪ best_hyp_model, X_train, y_train)
    learning_curve_metrics = {'train_sizes': train_sizes, 'train_scores':
    ↪ train_scores, 'test_scores': test_scores}

    # Fit model for validation predictions
    final_model = best_hyp_model.fit(X_train, y_train,
                                     early_stopping_rounds=10, eval_set=[(
    ↪ X_test, y_test)], verbose=False)

    # Predictions
    y_fit = final_model.predict(X_train)
    y_pred = final_model.predict(X_test)

    # Metrics from the final model
    metrics = {
        ’r2_train’: r2_score(y_train, y_fit),
        ’rmse_train’: mean_squared_error(y_train, y_fit, squared=False),
        ’mae_train’: mean_absolute_error(y_train, y_fit),
        ’r2_test’: r2_score(y_test, y_pred),
        ’rmse_test’: mean_squared_error(y_test, y_pred, squared=False),
        ’mae_test’: mean_absolute_error(y_test, y_pred),
        ’q2_test’: 1 - mean_squared_error(y_test, y_pred) / np.var(y_test)
    }

    # Print metrics
    for metric, value in metrics.items():
        print(f"{metric}: {round(value, 2)}")

    print(f’\nOptimization took {round((time.time() - start)/60, 1)}
    ↪ minutes’)

    return final_model, metrics, learning_curve_metrics

def save_model_and_metrics(model, metrics, save_directory, model_name):
    # Save the model with pickle
    model_path = os.path.join(save_directory, f’{model_name}.pkl’)
    with open(model_path, ’wb’) as file:
        pickle.dump(model, file)
    print(f"Model saved to {model_path}")

    # Convert metrics dictionary to a pandas DataFrame
    metrics_df = pd.DataFrame([metrics])

    # Save the metrics DataFrame to an Excel file
    metrics_file_path = os.path.join(save_directory, f’{model_name}
    ↪ _metrics.xlsx’)
    metrics_df.to_excel(metrics_file_path, index=False)
    print(f"Optimized model and metrics for {model_name} saved to {
    ↪ metrics_file_path}")

def save_learning_curves(learning_curve_metrics, save_directory,
    ↪ model_name):
    # Convert the learning curve metrics to a pandas DataFrame
    train_sizes, train_scores, test_scores = learning_curve_metrics.values
    ↪ ()
    # Convert the results into a DataFrame
    learning_curve_df = pd.DataFrame({
        ’Train Sizes’: train_sizes,
        ’Train Scores’: train_scores.mean(axis=1),
        ’Test Scores’: test_scores.mean(axis=1)
    })
    # Save the learning curve DataFrame to an Excel file
    learning_curve_file_path = os.path.join(save_directory, f'{model_name}_learning_curves.xlsx')
    learning_curve_df.to_excel(learning_curve_file_path, index=False)
    print(f"Learning curve for {model_name} saved to {learning_curve_file_path}")

def plot_learning_curves(learning_curve_metrics, save_directory, model_name, error=True):
    train_sizes, train_scores, test_scores = learning_curve_metrics.values()
    # Calculate means and standard deviations
    train_scores_mean = -train_scores.mean(axis=1)
    train_scores_std = train_scores.std(axis=1)
    test_scores_mean = -test_scores.mean(axis=1)
    test_scores_std = test_scores.std(axis=1)

    # Create plot
    plt.figure()
    if error is True:
        plt.fill_between(train_sizes, train_scores_mean - train_scores_std,
                        train_scores_mean + train_scores_std, alpha=0.1, color='r')
        plt.fill_between(train_sizes, test_scores_mean - test_scores_std,
                        test_scores_mean + test_scores_std, alpha=0.1, color='b')
    plt.plot(train_sizes, train_scores_mean, 'o-', color='r', label='Train')
    plt.plot(train_sizes, test_scores_mean, 'o-', color='b', label='Test')
    plt.xlabel('Training set size')
    plt.ylabel('RMSE')
    plt.legend(loc='best')

    # Save the plot
    plt.savefig(os.path.join(save_directory, f'{model_name}_learning_curves.png'))
    # Display the plot
    # plt.show()

def prediction_plot(model, metrics, X_train, y_train, X_test, y_test, save_directory, model_name):
    plt.figure()
    # Getting predictions
    train_predictions = model.predict(X_train)
    test_predictions = model.predict(X_test)

    # Plotting train and validation data
    train_scatter = plt.scatter(y_train, train_predictions, label='Train')
    train_color = train_scatter.get_facecolor()[0]  # Get color of the first trace
    plt.scatter(y_test, test_predictions, label='Validation')

    # Determine the range for the diagonal line
    combined_values = np.concatenate([y_train.to_numpy().ravel(), train_predictions.ravel(), y_test.to_numpy().ravel(), test_predictions.ravel()])
    min_val, max_val = combined_values.min(), combined_values.max()

    # Diagonal line indicating perfect predictions
    plt.plot([min_val, max_val], [min_val, max_val], ls="--", c="gray", alpha=0.5)

    # Labels and legend
    plt.xlabel('Actual')
    plt.ylabel('Predicted')
    plt.legend()

    # Adjust the limits and aspect ratio to make x and y axes have equal units
    plt.xlim(min_val, max_val)
    plt.ylim(min_val, max_val)
    # plt.axis('equal')

    # Add R^2 and MAE metrics to the plot
    textstr = '\n'.join((
        f'R2: {metrics["r2_train"]:.2f}',
        f'MAE: {metrics["mae_train"]:.2f}',
    ))
    # Position the text on the plot; adjust the position as necessary
    plt.gca().text(0.85, 0.1, textstr, transform=plt.gca().transAxes, fontsize=10, verticalalignment='top', color=train_color)

    # Ensure the save_directory exists
    os.makedirs(save_directory, exist_ok=True)

    plt.savefig(os.path.join(save_directory, f'{model_name}_prediction_plot.png'))
    # Display the plot
    # plt.show()

def summarise_learning_curves(root_dir):
    # DataFrame to collect all summary data
    summary_df = [['scaling', 'features', 'model', 'cross_validation_rmse']]

    # Define the directory structure
    scalings = ['minmax', 'noscale', 'power', 'standard']
```

```python
    features = ['all_features', 'RFE_selected_10', 'shap_selected_best']
    models = ['RF_learning_curves', 'SVR_learning_curves', 'KNN_learning_curves', 'XGB_learning_curves']

    # Traverse the directory structure
    for scaling in scalings:
        for feature in features:
            for model in models:
                file_path = os.path.join(root_dir, scaling, feature, model + '.xlsx')
                if os.path.exists(file_path):
                    # Load the Excel file
                    data = pd.read_excel(file_path)
                    # Extract the last value from the 'Test Scores' column
                    last_test_score = data['Test Scores'].iloc[-1]
                    # Append the data to the DataFrame
                    summary_df.append([scaling, feature, model.replace('_learning_curves', ''), last_test_score])

    # Save the summary DataFrame to a new Excel file
    summary_df = pd.DataFrame(summary_df[1:], columns=summary_df[0])
    summary_df.to_excel(os.path.join(root_dir, 'cv_summary.xlsx'), index=False)
    sorted_summary_df = summary_df.sort_values(by='cross_validation_rmse', ascending=False)
    sorted_summary_df.to_excel(os.path.join(root_dir, 'sorted_cv_summary.xlsx'), index=False)

    return sorted_summary_df



def compare_models(df_target_location, df_descriptors_location):
    # target generated from experiments
    df_target = pd.read_csv(df_target_location)
    # descriptors calculated with cheminformatics tools
    df_descriptors = pd.read_csv(df_descriptors_location)
    # descriptors determined experimentally, to be added to computationally generated descriptor datasets
    exp_descriptors = pd.read_csv(r'C:\Users\black\Projects\Molecule-Discovery\data\features\ph_sol_exp_features.csv')

    # merge computational descriptors with experimentally measured features
    merged_descriptors = pd.merge(df_descriptors, exp_descriptors, on='SMILES', how='inner')

    # save directory and titles
    saved_directory = r"C:\Users\black\Projects\Molecule-Discovery\
    ↪ notebooks\feature_selection\feature-target_comparisons"

    # specify the target column based on the file name
    head, tail = os.path.split(df_target_location)
    # dictionary mapping file names to target columns
    file_to_column = {
        'Rp_24h.csv': '<Rp> (kOhm cm2)',
        'Rp_avg.csv': '<Rp> (kOhm cm2)',
        'Rp_EIS24h.csv': '|Z| (kOhm cm2)',
        'IP_24h.csv': 'IP (dB)',
        'IP_avg.csv': 'IP (dB)',
        'IP_EIS24h.csv': '|Z| (kOhm cm2)',
        'IE_24h.csv': 'IE (%)',
        'IE_avg.csv': 'IE (%)',
        'IE_EIS24h.csv': '|Z| (kOhm cm2)'
    }
    # mapping
    if tail in file_to_column:
        target_column = file_to_column[tail]

    # later for use as threshold in PCA and others
    if target_column == '<Rp> (kOhm cm2)':
        threshold = 30000.
    elif target_column == 'IP (dB)':
        threshold = 3.
    elif target_column == 'IE (%)':
        threshold = 50.
    elif target_column == '|Z| (kOhm cm2)':
        threshold = 0.

    # target-feature combined directory for automatic saving
    feature_target_combination = combine_strings_with_underscore(
    ↪ process_string(df_descriptors_location),process_string(
    ↪ df_target_location))
    directory = os.path.join(saved_directory, feature_target_combination)
    os.makedirs(directory, exist_ok=True)

    # feature prepreprocessing with different scaling
    noscale_descriptors = preprocessing(merged_descriptors ,scale=False)
    minmax_scaled_descriptors = preprocessing(merged_descriptors ,scale=
    ↪ True, scaler='min_max')
    standard_scaled_descriptors = preprocessing(merged_descriptors ,scale=
    ↪ True, scaler='standard')
    power_scaled_descriptors = preprocessing(merged_descriptors ,scale=
    ↪ True, scaler='power')

    print(r'Total features after preprocessing:', noscale_descriptors.
    ↪ shape[1]-1)

    # make directories for each scaling
    noscale_directory = os.path.join(directory, 'noscale')
    minmax_directory = os.path.join(directory, 'minmax')
    standard_directory = os.path.join(directory, 'standard')
    power_directory = os.path.join(directory, 'power')
    os.makedirs(noscale_directory, exist_ok=True)
    os.makedirs(minmax_directory, exist_ok=True)
    os.makedirs(standard_directory, exist_ok=True)
    os.makedirs(power_directory, exist_ok=True)

    # read previously performed stratified split for consistent train/
    ↪ validation split
    df_stratified = pd.read_csv(os.path.join(saved_directory, '
    ↪ stratified_split.csv'))
    df_stratified = pd.merge(df_target, df_stratified, on='SMILES')

    # saving UMAP visualizations for each scaling
    visualize_chemspace(df_stratified, noscale_descriptors, split_names=["
    ↪ split"], target_column=target_column, output_file=os.path.join(
    ↪ noscale_directory, 'UMAP_chemspace.png'))
    visualize_chemspace(df_stratified, minmax_scaled_descriptors,
    ↪ split_names=["split"], target_column=target_column, output_file=os.
    ↪ path.join(minmax_directory, 'UMAP_chemspace.png'))
    visualize_chemspace(df_stratified, standard_scaled_descriptors,
    ↪ split_names=["split"], target_column=target_column, output_file=os.
    ↪ path.join(standard_directory, 'UMAP_chemspace.png'))
    visualize_chemspace(df_stratified, power_scaled_descriptors,
    ↪ split_names=["split"], target_column=target_column, output_file=os.
    ↪ path.join(power_directory, 'UMAP_chemspace.png'))

    # pearson correlations between selected features and target
    pearson_correlations(noscale_descriptors, df_target).to_excel(os.path.
    ↪ join(noscale_directory, 'pearson_correlations.xlsx'))
    pearson_correlations(minmax_scaled_descriptors, df_target).to_excel(os
    ↪ .path.join(minmax_directory, 'pearson_correlations.xlsx'))
    pearson_correlations(standard_scaled_descriptors, df_target).to_excel(
    ↪ os.path.join(standard_directory, 'pearson_correlations.xlsx'))
    pearson_correlations(power_scaled_descriptors, df_target).to_excel(os.
    ↪ path.join(power_directory, 'pearson_correlations.xlsx'))

    # feature selection with RFE and Shapley
    descriptor_list = [noscale_descriptors, minmax_scaled_descriptors,
    ↪ standard_scaled_descriptors, power_scaled_descriptors]
    directory_list = [noscale_directory, minmax_directory,
    ↪ standard_directory, power_directory]

    selected_features_10_list = []
    shap_best_list = []
    for descriptor, directory in zip(descriptor_list, directory_list):
```

```
    # RFE
    selected_features_10 = recursive_feature_elimination(descriptor,
↪ df_target, num_features=10)
    selected_features_10.to_excel(os.path.join(directory, '
↪ RFE_selected_features.xlsx'))
    selected_features_10_list.append(selected_features_10)
    # Shapley
    shap_best, report = shapley_feature_elimination(descriptor,
↪ df_target, reporting=True)
    shap_best.to_excel(os.path.join(directory, '
↪ shapley_selected_features.xlsx'))
    report.to_excel(os.path.join(directory, 'shapley_report.xlsx'))
    shap_best_list.append(shap_best)

feature_lists = [descriptor_list, selected_features_10_list,
↪ shap_best_list]

# PCA analysis for each scaling
for descriptor_lists, save_names in zip([descriptor_list,
↪ selected_features_10_list], ["all_features", "10_RFE-features"]):
    for descriptor, directory in zip(descriptor_lists, directory_list)
↪ :
        if len(perform_pca(df_target, descriptor)[0].columns) > 3:
            process_pca(save_names, df_target, descriptor,
↪ target_column, save_directory=directory, threshold=threshold)

# plotting histograms for selected features
for descriptor, directory_item in zip(selected_features_10_list,
↪ directory_list):
    plot_histograms(descriptor, directory_item)

# model training and evaluation

i = -1
for feature_list in feature_lists:
    for descriptor, feature_directory in zip(feature_list,
↪ directory_list):
        i = i+1
        print(f"***Training set {i+1} out of 12***")
        # stratified test and validation split
        X_train = descriptor[df_stratified['split']=='train'].drop(
↪ columns=['SMILES'], inplace=False)
        y_train = df_target[df_stratified['split']=='train'].drop(
↪ columns=['SMILES'], inplace=False)
        X_test = descriptor[df_stratified['split']=='validation'].drop
↪ (columns=['SMILES'], inplace=False)
        y_test = df_target[df_stratified['split']=='validation'].drop(
↪ columns=['SMILES'], inplace=False)

```

```
        # Defining regression models and their corresponding names
        models = {
            ’RF’: optimized_random_forest_model,
            ’SVR’: optimized_svm_model,
            ’KNN’: optimized_knn_model,
            ’XGB’: optimized_xgb_model
        }

        # Iterating through models, training, saving metrics, and
↪ plotting predictions
        for model_name, model_func in models.items():
            print(f"{model_name} model: {[’all_features’, ’
↪ RFE_selected_10’, ’shap_selected_best’][(i) // 4]}-{[’noscale’, ’
↪ minmax’, ’standard’, ’power’][i % 4]}")
            # create directories for model saving
            feat_directory = os.path.join(feature_directory, [’
↪ all_features’, ’RFE_selected_10’, ’shap_selected_best’][(i) // 4])
            os.makedirs(feat_directory, exist_ok=True)
            # model training
            model, metrics, learning_curve_metrics = model_func(
↪ X_train, y_train, X_test, y_test, shuffl=True)
            # saving metrics and plots
            save_model_and_metrics(model, metrics, feat_directory,
↪ model_name=model_name)
            prediction_plot(model, metrics, X_train, y_train, X_test,
↪ y_test, feat_directory, model_name=model_name)
            save_learning_curves(learning_curve_metrics,
↪ feat_directory, model_name)
            plot_learning_curves(learning_curve_metrics,
↪ feat_directory, model_name=model_name)

# summarise learning curves
directory = os.path.join(saved_directory, feature_target_combination)
summarise_learning_curves(directory)
```

BIBLIOGRAPHY

1. Winkler, D. A. *et al.* Impact of inhibition mechanisms, automation, and computational models on the discovery of organic corrosion inhibitors. *Progress in Materials Science,* 101392 (2024).
2. Popov, B. N. *Corrosion engineering: principles and solved problems* (Elsevier, 2015).
3. Raabe, D., Tasan, C. C. & Olivetti, E. A. Strategies for improving the sustainability of structural metals. *Nature* **575,** 64–74 (2019).
4. Waldman, J. *Rust: The longest war* (Simon and Schuster, 2015).
5. Koch, G. H. *et al. Corrosion cost and preventive strategies in the United States* tech. rep. (United States. Federal Highway Administration, 2016).
6. Koch, G. H., Brongers, M. P., Thompson, N. G., Virmani, Y. P., Payer, J. H., *et al. International Measures of Prevention, Application, and Economics of Corrosion Technologies Study* tech. rep. (NACE International EEB. IMPACT, 2002). http://impact.nace.org/documents/Nace-International-Report.pdf.
7. Sax, N. I., Bruce, R. D. & Durham, W. F. *Dangerous properties of industrial materials* (Van Nostrand Reinhold New York, 1975).
8. The European Parliament and the Council. *Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, amending Directive 199* 2006.
9. LaPuma, P. T., Fox, J. M. & Kimmel, E. C. Chromate concentration bias in primer paint particles. *Regulatory Toxicology and Pharmacology* **33,** 343–349 (2001).
10. Den Braver-Sewradj, S. P. *et al.* Occupational exposure to hexavalent chromium. Part II. Hazard assessment of carcinogenic effects. *Regulatory Toxicology and Pharmacology* **126,** 105045 (2021).
11. Hessel, E. V., Staal, Y. C., Piersma, A. H., den Braver-Sewradj, S. P. & Ezendam, J. Occupational exposure to hexavalent chromium. Part I. Hazard assessment of non-cancer health effects. *Regulatory Toxicology and Pharmacology* **126,** 105048 (2021).
12. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews* **16,** 3–50 (1996).
13. Mullard, A. *et al.* The drug-maker's guide to the galaxy. *Nature* **549,** 445–447 (2017).
14. Micklus, A. & Muntner, S. Biopharma deal-making in 2015: changing the pharma landscape. *Nature Reviews Drug Discovery* **15,** 78–80 (2016).

15. Pappu, A. & Paige, B. Making Graph Neural Networks Worth It for Low-Data Molecular Machine Learning. *arXiv.* eprint: 2011.12203. http://arxiv.org/abs/2011.12203 (2020).

16. Stanley, M. *et al.* FS-Mol: A Few-Shot Learning Dataset of Molecules. *NeurIPS.* https://github.com/microsoft/FS-Mol/ (2021).

17. Wei, J. *et al.* Machine learning in materials science. *InfoMat* **1,** 338–358 (2019).

18. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **3,** 54 (2017).

19. Jablonka, K. M., Schwaller, P. & Ortega-guerrero, A. Is GPT-3 all you need for low-data discovery in chemistry ? *ChemRxiv,* 1–32 (2023).

20. Gubernatis, J. & Lookman, T. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Physical Review Materials* **2,** 120301 (2018).

21. Moosavi, S. M., Jablonka, K. M. & Smit, B. The role of machine learning in the understanding and design of materials. *Journal of the American Chemical Society* **142,** 20273–20287 (2020).

22. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559,** 547–555 (2018).

23. Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. *Journal of Materiomics* **3,** 159–177 (2017).

24. Morgan, D. & Jacobs, R. Opportunities and challenges for machine learning in materials science. *Annual Review of Materials Research* **50,** 71–103 (2020).

25. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620,** 47–60. ISSN: 14764687 (2023).

26. Achar, S. K. & Keith, J. A. Small Data Machine Learning Approaches in Molecular and Materials Science. *Chemical Reviews* **124,** 13571–13573 (2024).

27. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials* **4,** 25 (2018).

28. Xu, P., Ji, X., Li, M. & Lu, W. Small data machine learning in materials science. *npj Computational Materials* **9,** 42 (2023).

29. Abolhasani, M. & Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis* **2,** 483–492 (2023).

30. Lo, S. *et al.* Review of low-cost self-driving laboratories in chemistry and materials science: the "frugal twin" concept. *Digital Discovery.* ISSN: 2635098X (2024).

31. Li, X. *et al.* Predicting corrosion inhibition efficiencies of small organic molecules using data-driven techniques. *npj Materials Degradation* **7.** ISSN: 23972106 (2023).

32. Würger, T. *et al.* Exploring structure-property relationships in magnesium dissolution modulators. *npj Materials Degradation* **5,** 1–10. ISSN: 23972106 (2021).

33. Winkler, D. A. *et al.* Towards chromate-free corrosion inhibitors: Structure-property models for organic alternatives. *Green Chemistry* **16,** 3349–3357. ISSN: 14639270 (2014).

34. Sutojo, T. *et al.* A machine learning approach for corrosion small datasets. *npj Materials Degradation* **7.** ISSN: 23972106 (2023).

35. Coelho, L. B. *et al.* Reviewing machine learning of corrosion prediction in a data-oriented perspective. *npj Materials Degradation* **6.** ISSN: 23972106 (2022).

36. Li, S.-.-S. *et al.* Development and applications of aluminum alloys for aerospace industry. *Journal of Materials Research and Technology* (2023).

37. Hughes, A. E. *et al.* High strength Al-alloys: microstructure, corrosion and principles of protection. *Recent trends in processing and degradation of aluminium alloys* **1,** 223–262 (2011).

38. *Aircraft Aluminium* https://aircraft-aluminium.com/. Accessed: 2025-01-09.

39. Hughes, A. E., Parvizi, R. & Forsyth, M. Microstructure and corrosion of AA2024. *Corrosion Reviews* **33,** 1–30. ISSN: 03346005 (2015).

40. Birbilis, N. & Buchheit, R. G. Electrochemical characteristics of intermetallic phases in aluminum alloys: an experimental survey and discussion. *Journal of the Electrochemical Society* **152,** B140 (2005).

41. Yasakau, K., Zheludkevich, M. & Ferreira, M. Role of intermetallics in corrosion of aluminum alloys. Smart corrosion protection. *Intermetallic matrix composites,* 425–462 (2018).

42. Kosari, A. *et al.* Dealloying-driven local corrosion by intermetallic constituent particles and dispersoids in aerospace aluminium alloys. *Corrosion Science* **177,** 108947. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2020.108947 (2020).

43. Boag, A. *et al.* How complex is the microstructure of AA2024-T3? *Corrosion Science* **51,** 1565–1568. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2009.05.001 (2009).

44. Xu, Y., Thompson, G. & Wood, G. Mechanism of anodic film growth on aluminium. *Transactions of the IMF* **63,** 98–103 (1985).

45. Carroll, W. & Breslin, C. Stability of passive films formed on aluminium in aqueous halide solutions. *British Corrosion Journal* **26,** 255–259 (1991).

46. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68,** 314–319 (2013).

47. Verma, C., Ebenso, E. E. & Quraishi, M. A. Corrosion inhibitors for ferrous and non-ferrous metals and alloys in ionic sodium chloride solutions: A review. *Journal of Molecular Liquids* **248,** 927–942. ISSN: 01677322. https://doi.org/10.1016/j.molliq.2017.10.094 (2017).

48. Feliu, S. *et al.* Characterisation of porous and barrier layers of anodic oxides on different aluminium alloys. *Journal of Applied Electrochemistry* **37,** 1027–1037 (2007).

49. Feliu Jr, S., Bartolomé, M. J., González, J., López, V. & Feliu, S. Passivating oxide film and growing characteristics of anodic coatings on aluminium alloys. *Applied Surface Science* **254,** 2755–2762 (2008).

50. Olgiati, M., Denissen, P. J. & Garcia, S. J. When all intermetallics dealloy in AA2024-T3: Quantifying early stage intermetallic corrosion kinetics under immersion. *Corrosion Science* **192,** 109836 (2021).

51. Boag, A. *et al.* Stable pit formation on AA2024-T3 in a NaCl environment. *Corrosion Science* **52,** 90–103 (2010).

52. Zhang, X., Zhou, X., Hashimoto, T. & Liu, B. Localized corrosion in AA2024-T351 aluminium alloy: Transition from intergranular corrosion to crystallographic pitting. *Materials Characterization* **130,** 230–236 (2017).

53. Zhang, W. & Frankel, G. Transitions between pitting and intergranular corrosion in AA2024. *Electrochimica Acta* **48,** 1193–1210 (2003).

54. Zhou, X., Luo, C., Hashimoto, T., Hughes, A. & Thompson, G. Study of localized corrosion in AA2024 aluminium alloy using electron tomography. *Corrosion Science* **58,** 299–306 (2012).

55. Kosari, A. *et al.* In-situ nanoscopic observations of dealloying-driven local corrosion from surface initiation to in-depth propagation. *Corrosion Science* **177,** 108912. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2020.108912 (2020).

56. Boag, A., Hughes, A., Glenn, A., Muster, T. & McCulloch, D. Corrosion of AA2024-T3 Part I: Localised corrosion of isolated IM particles. *Corrosion Science* **53,** 17–26 (2011).

57. Hughes, A. *et al.* Corrosion of AA2024-T3 Part II: Co-operative corrosion. *Corrosion Science* **53,** 27–39 (2011).

58. Glenn, A. *et al.* Corrosion of AA2024-T3 part III: propagation. *Corrosion Science* **53,** 40–50 (2011).

59. Hughes, A. *et al.* Co-operative corrosion phenomena. *Corrosion science* **52,** 665–668 (2010).

60. Jakab, M., Presuel-Moreno, F. & Scully, J. Effect of molybdate, cerium, and cobalt ions on the oxygen reduction reaction on AA2024-T3 and selected intermetallics: experimental and modeling studies. *Journal of the Electrochemical Society* **153,** B244 (2006).

61. Paz Martínez-Viademonte, M., Abrahami, S. T., Hack, T., Burchardt, M. & Terryn, H. A review on anodizing of aerospace aluminum alloys for corrosion protection. *Coatings* **10,** 1106 (2020).

62. Gharbi, O., Thomas, S., Smith, C. & Birbilis, N. Chromate replacement: what does the future hold? *npj Materials Degradation* **2,** 23–25. ISSN: 23972106. http://dx.doi.org/10.1038/s41529-018-0034-5 (2018).

63. Al-Amiery, A. A., Isahak, W. N. R. W. & Al-Azzawi, W. K. Corrosion inhibitors: natural and synthetic organic inhibitors. *Lubricants* **11,** 174 (2023).

64. Dariva, C. G. & Galio, A. F. in *Developments in Corrosion Protection* (ed Aliofkhazraei, M.) 365–378 (InTech, London, UK, 2014).

65. Andreatta, F. & Fedrizzi, L. in *Active Protective Coatings* (eds Hughes, A. E., Mol, J. M. C., Zheludkevich, M. L. & Buchheit, R. G.) 233rd ed., 59–84 (Springer Netherlands, Dordrecht, 2016). http://link.springer.com/10.1007/978-94-017-7540-3_4.

66. Montemor, M. F. in *Active Protective Coatings* (eds Hughes, A. E., Mol, J. M. C., Zheludkevich, M. L. & Buchheit, R. G.) 233rd ed., 107–137 (Springer Netherlands, Dordrecht, 2016).

67. Haynes, W. M. *CRC handbook of chemistry and physics* (CRC press, 2016).

68. Kovačević, N. & Kokalj, A. Chemistry of the interaction between azole type corrosion inhibitor molecules and metal surfaces. *Materials Chemistry and Physics* **137,** 331–339 (2012).

69. Straub, B. F. *Organotransition Metal Chemistry. From Bonding to Catalysis. Edited by John F. Hartwig.* 2010.

70. Milošev, I. *et al.* The effect of molecular structure of imidazole-based compounds on corrosion inhibition of Cu, Zn, and Cu-Zn alloys. *Corrosion Science* **240,** 112328 (2024).

71. Kokalj, A. On the alleged importance of the molecular electron-donating ability and the HOMO–LUMO gap in corrosion inhibition studies. *Corrosion Science* **180,** 109016. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2020.109016 (2021).

72. Kokalj, A. & Costa, D. Model study of penetration of Cl- ions from solution into organic self-assembled-monolayer on metal substrate: trends and modeling aspects. *Journal of The Electrochemical Society* **168,** 071508 (2021).

73. Gašparič, L., Poberžnik, M. & Kokalj, A. DFT study of hydrogen bonding between metal hydroxides and organic molecules containing N, O, S, and P heteroatoms: clusters vs. surfaces. *Chemical Physics* **559,** 111539 (2022).

74. Zhu, J. & Hihara, L. Corrosion of continuous alumina-fibre reinforced Al–2 wt.% Cu–T6 metal–matrix composite in 3.15 wt.% NaCl solution. *Corrosion Science* **52,** 406–415 (2010).

75. Beverskog, B. & Puigdomenech, I. Revised Pourbaix diagrams for copper at 25 to 300 C. *Journal of the Electrochemical Society* **144,** 3476 (1997).

76. Kokalj, A. Corrosion inhibitors: physisorbed or chemisorbed? *Corrosion Science* **196,** 109939 (2022).

77. Curkovic, H. O., Stupnisek-Lisac, E. & Takenouti, H. The influence of pH value on the efficiency of imidazole based corrosion inhibitors of copper. *Corrosion Science* **52,** 398–405 (2010).

78. Finšgar, M., Lesar, A., Kokalj, A. & Milošev, I. A comparative electrochemical and quantum chemical calculation study of BTAH and BTAOH as copper corrosion inhibitors in near neutral chloride solution. *Electrochimica Acta* **53,** 8287–8297. ISSN: 00134686 (2008).

79. Taheri, P. *et al.* On the importance of time-resolved electrochemical evaluation in corrosion inhibitor-screening studies. *npj Materials Degradation* **4,** 1–4. ISSN: 23972106 (2020).

80. Özkan, C. *et al.* Laying the experimental foundation for corrosion inhibitor discovery through machine learning. *npj Materials Degradation* **8,** 21. ISSN: 2397-2106. https://www.nature.com/articles/s41529-024-00435-z (Feb. 2024).

81. Chyżewski, E. & Evans, U. R. The Classification of Anodic and Cathodic Inhibitors. *Transactions of The Electrochemical Society* **76,** 215. ISSN: 00964743 (1939).

82. Hey, A., Tansley, S. & Tolle, K. *The fourth paradigm: data-intensive scientific discovery* (Microsoft Research Redmond, WA, 2009).

83. Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science. *APL Materials* **4,** 1–10. ISSN: 2166532X. http://dx.doi.org/10.1063/1.4946894 (2016).

84. Frankel, G. S. & McCreery, R. L. Inhibition of Al Alloy Corrosion by Chromates. *The Electrochemical Society Interface* **10,** 34–38. ISSN: 1064-8208. https://iopscience.iop.org/article/10.1149/2.F06014IF (Dec. 2001).

85. Kendig, M. W. & Buchheit, R. G. Corrosion inhibition of aluminum and aluminum alloys by soluble chromates, chromate coatings, and chromate-free coatings. *Corrosion* **59,** 379–400. ISSN: 00109312 (2003).

86. Ilevbare, G. O. Inhibition of pitting corrosion on aluminum alloy 2024-T3: Effect of soluble chromate additions vs chromate conversion coating. *Corrosion* **56,** 227–242. ISSN: 00109312 (2000).

87. Yasakau, K. A., Zheludkevich, M. L., Lamaka, S. V. & Ferreira, M. G. Mechanism of corrosion inhibition of AA2024 by rare-earth compounds. *Journal of Physical Chemistry B* **110,** 5515–5528. ISSN: 15206106 (2006).

88. Matter, E. A., Kozhukharov, S., Machkova, M. & Kozhukharov, V. Comparison between the inhibition efficiencies of Ce(III) and Ce(IV) ammonium nitrates against corrosion of AA2024 aluminum alloy in solutions of low chloride concentration. *Corrosion Science* **62,** 22–33. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2012.03.039 (2012).

89. Kosari, A. *et al.* Editors' Choice—Dealloying-Driven Cerium Precipitation on Intermetallic Particles in Aerospace Aluminium Alloys. *Journal of The Electrochemical Society* **168,** 041505. ISSN: 0013-4651 (2021).

90. Markley, T. A., Forsyth, M. & Hughes, A. E. Corrosion protection of AA2024-T3 using rare earth diphenyl phosphates. *Electrochimica Acta* **52,** 4024–4031. ISSN: 00134686 (2007).

91. Lopez-Garrity, O. & Frankel, G. S. Corrosion Inhibition of Aluminum Alloy 2024-T3 by Sodium Molybdate. *Journal of The Electrochemical Society* **161,** C95–C106. ISSN: 0013-4651 (2014).

92. Jakab, M. A., Presuel-Moreno, F. & Scully, J. R. Effect of Molybdate, Cerium, and Cobalt Ions on the Oxygen Reduction Reaction on AA2024-T3 and Selected Intermetallics. *Journal of The Electrochemical Society* **153,** B244. ISSN: 00134651 (2006).

93. Kannan, B., Glover, C. F., McMurray, H. N., Williams, G. & Scully, J. R. Performance of a Magnesium-Rich Primer on Pretreated AA2024-T351 in Full Immersion: a Galvanic Throwing Power Investigation Using a Scanning Vibrating Electrode Technique. *Journal of The Electrochemical Society* **165,** C27–C41. ISSN: 0013-4651 (2018).

94. Collazo, A., Nóvoa, X. R. & Pérez, C. The role of Mg2+ ions in the corrosion behaviour of AA2024-T3 aluminium alloys immersed in chloride-containing environments. *Electrochimica Acta* **124,** 17–26. ISSN: 00134686. http://dx.doi.org/10.1016/j.electacta.2013.10.130 (2014).

95. Santucci, R. J. & Scully, J. R. Mechanistic Framework for Understanding pH-Induced Electrode Potential Control of AA2024-T351 by Protective Mg-Based Pigmented Coatings. *Journal of The Electrochemical Society* **167,** 131514. ISSN: 0013-4651 (2020).

96. Kosari, A. *et al.* Laterally-resolved formation mechanism of a lithium-based conversion layer at the matrix and intermetallic particles in aerospace aluminium alloys. *Corrosion Science* **190,** 109651. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2021.109651 (2021).

97. Visser, P., Gonzalez-Garcia, Y., Mol, J. M. C. & Terryn, H. Mechanism of Passive Layer Formation on AA2024-T3 from Alkaline Lithium Carbonate Solutions in the Presence of Sodium Chloride. *Journal of The Electrochemical Society* **165,** C60–C70. ISSN: 0013-4651 (2018).

98. Visser, P., Meeusen, M., Gonzalez-Garcia, Y., Terryn, H. & Mol, J. M. C. Electrochemical Evaluation of Corrosion Inhibiting Layers Formed in a Defect from Lithium-Leaching Organic Coatings. *Journal of The Electrochemical Society* **164,** C396–C406. ISSN: 0013-4651 (2017).

99. Marinescu, M. Recent advances in the use of benzimidazoles as corrosion inhibitors. *BMC Chemistry* **13,** 1–21. ISSN: 2661801X. https://doi.org/10.1186/s13065-019-0655-y (2019).

100. Xhanari, K. *et al.* Green corrosion inhibitors for aluminium and its alloys: A review. *RSC Advances* **7,** 27299–27330. ISSN: 20462069 (2017).

101. Zheludkevich, M. L., Yasakau, K. A., Poznyak, S. K. & Ferreira, M. G. Triazole and thiazole derivatives as corrosion inhibitors for AA2024 aluminium alloy. *Corrosion Science* **47,** 3368–3383. ISSN: 0010938X (2005).

102. Recloux, I. *et al.* Stability of benzotriazole-based films against AA2024 aluminium alloy corrosion process in neutral chloride electrolyte. *Journal of Alloys and Compounds* **735,** 2512–2522. ISSN: 09258388. https://doi.org/10.1016/j.jallcom.2017.11.346 (2018).

103. Verma, C., Quraishi, M. A. & Ebenso, E. E. Quinoline and its derivatives as corrosion inhibitors: A review. *Surfaces and Interfaces* **21.** ISSN: 24680230 (2020).

104. Snihirova, D., Lamaka, S. V., Taheri, P., Mol, J. M. C. & Montemor, M. F. Comparison of the synergistic effects of inhibitor mixtures tailored for enhanced corrosion protection of bare and coated AA2024-T3. *Surface and Coatings Technology* **303,** 342–351. ISSN: 02578972. http://dx.doi.org/10.1016/j.surfcoat.2015.10.075 (2016).

105. Mohammadi, I., Shahrabi, T., Mahdavian, M. & Izadi, M. Sodium diethyldithiocarbamate as a novel corrosion inhibitor to mitigate corrosion of 2024-T3 aluminum alloy in 3.5 wt NaCl solution. *Journal of Molecular Liquids* **307,** 112965. ISSN: 01677322. https://doi.org/10.1016/j.molliq.2020.112965 (2020).

106. Prakashaiah, B. G., Vinaya Kumara, D., Anup Pandith, A., Nityananda Shetty, A. & Amitha Rani, B. E. Corrosion inhibition of 2024-T3 aluminum alloy in 3.5 NaCl by thiosemicarbazone derivatives. *Corrosion Science* **136,** 326–338. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2018.03.021 (2018).

107. Harvey, T. G. *et al.* The effect of inhibitor structure on the corrosion of AA2024 and AA7075. *Corrosion Science* **53,** 2184–2190. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2011.02.040 (2011).

108. Lamaka, S. V., Zheludkevich, M. L., Yasakau, K. A., Montemor, M. F. & Ferreira, M. G. High effective organic corrosion inhibitors for 2024 aluminium alloy. *Electrochimica Acta* **52,** 7231–7247. ISSN: 00134686 (2007).

109. Xhanari, K. & Finšgar, M. Organic corrosion inhibitors for aluminum and its alloys in chloride and alkaline solutions: A review. *Arabian Journal of Chemistry* **12,** 4646–4663. ISSN: 18785352 (2019).

110. Popoola, L. T. Organic green corrosion inhibitors (OGCIs): A critical review. *Corrosion Reviews* **37,** 71–102. ISSN: 03346005 (2019).

111. Zhou, B., Wang, Y. & Zuo, Y. Evolution of the corrosion process of AA 2024-T3 in an alkaline NaCl solution with sodium dodecylbenzenesulfonate and lanthanum chloride inhibitors. *Applied Surface Science* **357,** 735–744. ISSN: 01694332 (2015).

112. Meeusen, M. *et al.* A Complementary Electrochemical Approach for Time-Resolved Evaluation of Corrosion Inhibitor Performance. *Journal of The Electrochemical Society* **166,** C3220–C3232. ISSN: 0013-4651 (2019).

113. Visser, P., Terryn, H. & Mol, J. M. C. On the importance of irreversibility of corrosion inhibitors for active coating protection of AA2024-T3. *Corrosion Science* **140,** 272–285. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2018.05.037 (2018).

114. White, P. A. *et al.* Towards materials discovery: Assays for screening and study of chemical interactions of novel corrosion inhibitors in solution and coatings. *New Journal of Chemistry* **44,** 7647–7658. ISSN: 13699261 (2020).

115. White, P. A. *et al.* A new high-throughput method for corrosion testing. *Corrosion Science* **58,** 327–331. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2012.01.016 (2012).

116. Taylor, S. & Chambers, B. THE DISCOVERY OF NON-CHROMATE CORROSION INHIBITORS FOR AEROSPACE ALLOYS USING HIGH-THROUGHPUT SCREENING METHODS. *Corrosion Reviews* **25,** 571–590. ISSN: 2191-0316. https://www.degruyter.com/document/doi/10.1515/CORRREV.2007.25.5-6.571/html (Dec. 2007).

117. Muster, T. H. *et al.* A rapid screening multi-electrode method for the evaluation of corrosion inhibitors. *Electrochimica Acta* **54,** 3402–3411. ISSN: 00134686 (2009).

118. Muster, T. H. *et al.* A review of high throughput and combinatorial electrochemistry. *Electrochimica Acta* **56,** 9679–9699. ISSN: 00134686. http://dx.doi.org/10.1016/j.electacta.2011.09.003 (2011).

119. García, S. J. *et al.* The influence of pH on corrosion inhibitor selection for 2024-T3 aluminium alloy assessed by high-throughput multielectrode and potentiodynamic testing. *Electrochimica Acta* **55,** 2457–2465. ISSN: 00134686 (2010).

120. Chambers, B. D. & Taylor, S. R. High-throughput assessment of inhibitor synergies on aluminum alloy 2024-T3 through measurement of surface copper enrichment. *Corrosion* **63,** 268–276. ISSN: 00109312 (2007).

121. Lamaka, S. V. *et al.* Comprehensive screening of Mg corrosion inhibitors. *Corrosion Science* **128,** 224–240. ISSN: 0010938X (2017).

122. Feiler, C. *et al.* In silico screening of modulators of magnesium dissolution. *Corrosion Science* **163,** 108245. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2019.108245 (2020).

123. Zabula, A. V. *et al.* Screening of molecular lanthanide corrosion inhibitors by a high-throughput method. *Corrosion Science* **165,** 108377. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2019.108377 (2020).

124. White, P. A. *et al.* High-throughput channel arrays for inhibitor testing: Proof of concept for AA2024-T3. *Corrosion Science* **51,** 2279–2290. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2009.06.038 (2009).

125. Visser, P. *et al.* Li leaching from Li carbonate-primer: Transport pathway development from the scribe edge of a primer/topcoat system. *Progress in Organic Coatings* **158,** 106284. ISSN: 03009440. https://doi.org/10.1016/j.porgcoat.2021.106284 (2021).

126. Moraes, C. V., Santucci, R. J., Scully, J. R. & Kelly, R. G. Finite Element Modeling of Chemical and Electrochemical Protection Mechanisms Offered by Mg-Based Organic Coatings to AA2024-T351. *Journal of The Electrochemical Society* **168,** 051505. ISSN: 0013-4651. http://dx.doi.org/10.1149/1945-7111/abfab8%20https://iopscience.iop.org/article/10.1149/1945-7111/abfab8 (May 2021).

127. Binggeli, M., Shen, T.-H. & Tileli, V. Simulating Current Distribution of Oxygen Evolution Reaction in Microcells Using Finite Element Method. *Journal of The Electrochemical Society* **168,** 106508. ISSN: 0013-4651. https://iopscience.iop.org/article/10.1149/1945-7111/ac2ebf (Oct. 2021).

128. Obot, I. B., Macdonald, D. D. & Gasem, Z. M. Density functional theory (DFT) as a powerful tool for designing new organic corrosion inhibitors: Part 1: An overview. *Corrosion Science* **99,** 1–30. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2015.01.037 (2015).

129. Kokalj, A. & Costa, D. *Molecular modeling of corrosion inhibitors* 332–345. ISBN: 9780128098943. http://dx.doi.org/10.1016/B978-0-12-409547-2.13444-4 (Elsevier, 2018).

130. Kokalj, A. *et al.* Simplistic correlations between molecular electronic properties and inhibition efficiencies: Do they really exist? *Corrosion Science* **179,** 108856. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2020.108856 (2021).

131. Luo, X. *et al.* Computational simulation and efficient evaluation on corrosion inhibitors for electrochemical etching on aluminum foil. *Corrosion Science* **187,** 109492. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2021.109492 (2021).

132. Costa, D., Ribeiro, T., Cornette, P. & Marcus, P. DFT modeling of corrosion inhibition by organic molecules: Carboxylates as inhibitors of aluminum corrosion. *Journal of Physical Chemistry C* **120,** 28607–28616. ISSN: 19327455 (2016).

133. Milošev, I. *et al.* Electrochemical, Surface-Analytical, and Computational DFT Study of Alkaline Etched Aluminum Modified by Carboxylic Acids for Corrosion Protection and Hydrophobicity. *Journal of The Electrochemical Society* **166,** C3131–C3146. ISSN: 0013-4651 (2019).

134. Milošev, I. *et al.* Editors' Choice—The Effect of Anchor Group and Alkyl Backbone Chain on Performance of Organic Compounds as Corrosion Inhibitors for Aluminum Investigated Using an Integrative Experimental-Modeling Approach. *Journal of The Electrochemical Society* **167,** 061509. ISSN: 1945-7111 (2020).

135. Milošev, I. *et al.* The Effects of Perfluoroalkyl and Alkyl Backbone Chains, Spacers, and Anchor Groups on the Performance of Organic Compounds as Corrosion Inhibitors for Aluminum Investigated Using an Integrative Experimental-Modeling Approach. *Journal of The Electrochemical Society* **168,** 071506. ISSN: 0013-4651 (2021).

136. Winkler, D. A. *et al.* Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors. *Corrosion Science* **106,** 229–235. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2016.02.008 (2016).

137. Würger, T. *et al.* Data science based Mg corrosion engineering. *Frontiers in Materials* **6,** 1–9. ISSN: 22968016 (2019).

138. Schiessler, E. J. *et al.* Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning models. *npj Computational Materials* **7,** 39–41. ISSN: 20573960 (2021).

139. Galvão, T. L., Novell-Leruth, G., Kuznetsova, A., Tedim, J. & Gomes, J. R. Elucidating Structure-Property Relationships in Aluminum Alloy Corrosion Inhibitors by Machine Learning. *Journal of Physical Chemistry C* **124,** 5624–5635. ISSN: 19327455 (2020).

140. Galvão, T. L. P. *et al.* CORDATA : an open data management web application to select corrosion inhibitors, 4–7 (2022).

141. *RDKit: Open-source cheminformatics* https://www.rdkit.org. Accessed: 2023-04-05.

142. *TURBOMOLE V?, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from*
http://www.turbomole.com.

143. *Impedance Spectroscopy: Theory, Experiment, and Applications* Second (eds Barsoukov, E. & Macdonald, J. R.) ISBN: 9780471647492. https://onlinelibrary.wiley.com/doi/book/10.1002/0471716243 (John Wiley & Sons, Apr. 2005).

144. Scully, J. R. Polarization resistance method for determination of instantaneous corrosion rates. *Corrosion* **56,** 199–217. ISSN: 00109312 (2000).

145. Pourbaix, M. Atlas of electrochemical equilibria in aqueous solutions. *NACE* (1966).

146. Thabtah, F., Hammoud, S., Kamalov, F. & Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Information Sciences* **513,** 429–441. ISSN: 00200255 (2020).

147. Benesty, J., Chen, J., Huang, Y. & Cohen, I. in *Springer Topics in Signal Processing* 1–4 (2009). ISBN: 9783642002960. https://link.springer.com/10.1007/978-3-642-00296-0_7%20http://link.springer.com/10.1007/978-3-642-00296-0_5.

148. Kelly, R. G., Scully, J. R., Shoesmith, D. & Buchheit, R. G. *Electrochemical Techniques in Corrosion Science and Engineering* ISBN: 9780203909133. https://www.taylorfrancis.com/books/9780203909133 (CRC Press, New York, Sept. 2002).

149. Verma, C., Verma, D. K., Ebenso, E. E. & Quraishi, M. A. Sulfur and phosphorus heteroatom-containing compounds as corrosion inhibitors: An overview. *Heteroatom Chemistry* **29.** ISSN: 10981071 (2018).

150. Rani, B. E. & Basu, B. B. J. Green inhibitors for corrosion protection of metals and alloys: An overview. *International Journal of Corrosion* **2012.** ISSN: 16879325 (2012).

151. Neupane, S. *et al.* Study Of Mercaptobenzimidazoles As Inhibitors For Copper Corrosion: Down to the Molecular Scale. *Journal of The Electrochemical Society* **168,** 051504. ISSN: 0013-4651 (2021).

152. Kozlica, D. K., Kokalj, A. & Milošev, I. Synergistic effect of 2-mercaptobenzimidazole and octylphosphonic acid as corrosion inhibitors for copper and aluminium – An electrochemical, XPS, FTIR and DFT study. *Corrosion Science* **182.** ISSN: 0010938X (2021).

153. Wu, X., Wiame, F., Maurice, V. & Marcus, P. Molecular scale insights into interaction mechanisms between organic inhibitor film and copper. *npj Materials Degradation* **5,** 1–8. ISSN: 23972106. http://dx.doi.org/10.1038/s41529-021-00168-3 (2021).

154. Özçelik, R., van Tilborg, D., Jiménez-Luna, J. & Grisoni, F. Structure-based drug discovery with deep learning. *ChemBioChem* **202200776.** ISSN: 1439-4227. arXiv: 2212.13295. http://arxiv.org/abs/2212.13295%20https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cbic.202200776 (June 2023).

155. Harren, T., Matter, H., Hessler, G., Rarey, M. & Grebner, C. Interpretation of Structure-Activity Relationships in Real-World Drug Design Data Sets Using Explainable Artificial Intelligence. *Journal of Chemical Information and Modeling* **62,** 447–462. ISSN: 1549960X (2022).

156. Miyao, T., Kaneko, H. & Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x). *Journal of Chemical Information and Modeling* **56,** 286–299. ISSN: 1549960X (2016).

157. Lo, Y. C., Senese, S., Damoiseaux, R. & Torres, J. Z. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. *ACS Chemical Biology* **11,** 2244–2253. ISSN: 15548937 (2016).

158. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery* **16,** 949–959. ISSN: 1746045X. https://doi.org/10.1080/17460441.2021.1909567 (2021).

159. Amar, Y., Schweidtmann, A. M., Deutsch, P., Cao, L. & Lapkin, A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical Science* **10,** 6697–6706. ISSN: 20416539 (2019).

160. Obrezanova, O., Csányi, G., Gola, J. M. & Segall, M. D. Gaussian processes: A method for automatic QSAR modeling of ADME properties. *Journal of Chemical Information and Modeling* **47,** 1847–1857. ISSN: 1549960X (2007).

161. Moret, M. *et al.* Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications* **14.** ISSN: 20411723 (2023).

162. Edwards, D. A. Steric hindrance effects on surface reactions: Applications to BIAcore. *Journal of Mathematical Biology* **55,** 517–539. ISSN: 03036812 (2007).

163. Yun, L. *et al.* Evaluation and Optimization of Corrosion Inhibitor System. *IOP Conference Series Materials Science and Engineering* **729.** ISSN: 1757899X (2020).

164. Frankel, G. S. in *Active protective coatings: new-generation coatings for metals* 17–32 (2016). http://link.springer.com/10.1007/978-94-017-7540-3_2.

165. Erlebacher, J. An Atomistic Description of Dealloying. *Journal of The Electrochemical Society* **151,** C614. ISSN: 00134651 (2004).

166. Kolics, A., Besing, A. S., Baradlai, P., Haasch, R. & Wieckowski, A. Effect of pH on Thickness and Ion Content of the Oxide Film on Aluminum in NaCl Media. *Journal of The Electrochemical Society* **148,** B251. https://dx.doi.org/10.1149/1.1376118 (May 2001).

167. Lamaka, S. V. *et al.* Local pH and Its Evolution Near Mg Alloy Surfaces Exposed to Simulated Body Fluids. *Advanced Materials Interfaces* **5,** 1800169. ISSN: 21967350. https://onlinelibrary.wiley.com/doi/10.1002/admi.201800169 (Sept. 2018).

168. Winkler, D. A. Predicting the performance of organic corrosion inhibitors. *Metals* **7,** 1–8. ISSN: 20754701 (2017).

169. Özkan, C., Anusuyadevi, P. R., Visser, P., Taheri, P. & Mol, A. Factors to consider in the quest for organic alternatives to hexavalent chromium based corrosion inhibitors. *Corrosion Science,* 113183 (2025).

170. European Commission. *Commission Regulation (EU) No 143/2011 of 17 February 2011 amending Annex XIV to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals ('REACH'). Off. J. Eur. Union L244/ 6–L244/9* 2014.

171. European Chemicals Agency (ECHA). *List of substances included in Annex XIV of REACH ("Authorisation List").* https://echa.europa.eu/authorisation-list (2024).

172. European Chemicals Agency (ECHA). *List of substances included in Annex XVII of REACH ("Restriction List").* https://echa.europa.eu/substances-restricted-under-reach (2024).

173. European Commission. *Questions & Answers, REACH and Chromium(VI) substances* 2023. https://ec.europa.eu/docsroom/documents/56174 (2024).

174. European Commission. *Request to the European Chemicals Agency to prepare an Annex XV restriction dossier on certain chromium (VI) substances* 2023. https://echa.europa.eu/documents/10162/17233/restdp_chromium_vi_mandate_redacted_en.pdf/3f84e822-1ed3-1cf0-899c-3b04083b8fe5?t=1714720983205 (2024).

175. European Commission. *Amendment complementing the request of 27 September 2023 to the European Chemicals Agency to prepare an Annex XV restriction dossier on certain chromium(VI) substances* 2024. https://echa.europa.eu/documents/10162/17233/restdp_chromium_vi_mandate_amendment_redacted_en.pdf/7ce19029-2683-955f-65c3-d77fc11a2e43?t=1714721085864 (2024).

176. Visser, P., Liu, Y., Terryn, H. & Mol, J. M. C. Lithium salts as leachable corrosion inhibitors and potential replacement for hexavalent chromium in organic coatings for the protection of aluminum alloys. *Journal of Coatings Technology and Research* **13,** 557–566. ISSN: 15470091 (2016).

177. Visser, P., Terryn, H. & Mol, J. M. C. Active corrosion protection of various aluminium alloys by lithium-leaching coatings. *Surface and Interface Analysis* **51,** 1276–1287. ISSN: 10969918 (2019).

178. Hughes, A. E., Markley, T. A., Garcia, S. J. & Mol, J. M. C. Comparative study of protection of AA 2024-T3 exposed to rare earth salts solutions. *Corrosion Engineering Science and Technology* **49,** 674–687. ISSN: 17432782 (2014).

179. Gobara, M., Baraka, A., Akid, R. & Zorainy, M. Corrosion protection mechanism of Ce4+/organic inhibitor for AA2024 in 3.5% NaCl. *RSC Advances* **10,** 2227–2240. ISSN: 20462069 (2020).

180. Dai, J. *et al.* Cross-category prediction of corrosion inhibitor performance based on molecular graph structures via a three-level message passing neural network model. *Corrosion Science* **209,** 110780. ISSN: 0010938X. https://doi.org/10.1016/j.corsci.2022.110780 (2022).

181. Galvão, T. L., Ferreira, I., Maia, F., Gomes, J. R. & Tedim, J. DATACORTECH: artificial intelligence platform for the virtual screen of aluminum corrosion inhibitors. *npj Materials Degradation* **8.** ISSN: 23972106. http://dx.doi.org/10.1038/s41529-024-00489-z (2024).

182. Zhao, S. & Birbilis, N. Searching for chromate replacements using natural language processing and machine learning algorithms. *npj Materials Degradation* **7.** ISSN: 23972106 (2023).

183. Roberto, E. C. *et al.* The effect of type of self-assembled system and pH on the efficiency of corrosion inhibition of carbon-steel surfaces. *Progress in Organic Coatings* **76,** 1308–1315 (2013).

184. Vujičić, V. & Lovreček, B. A study of the influence of pH on the corrosion rate of aluminium. *Surface technology* **25,** 49–57 (1985).

185. Visser, P. *et al.* The chemical throwing power of lithium-based inhibitors from organic coatings on AA2024-T3. *Corrosion Science* **150,** 194–206. ISSN: 0010938X (2019).

186. Meeusen, M. *et al.* The effect of time evolution and timing of the electrochemical data recording of corrosion inhibitor protection of hot-dip galvanized steel. *Corrosion science* **173,** 108780 (2020).

187. Homborg, A. M. *et al.* Application of transient analysis using Hilbert spectra of electrochemical noise to the identification of corrosion inhibition. *Electrochimica Acta* **116,** 355–365 (2014).

188. Xia, L., Akiyama, E., Frankel, G. & McCreery, R. Storage and release of soluble hexavalent chromium from chromate conversion coatings equilibrium aspects of Cr VI concentration. *Journal of the Electrochemical Society* **147,** 2556 (2000).

189. Zhao, J. *et al.* Effects of chromate and chromate conversion coatings on corrosion of aluminum alloy 2024-T3. *Surface and Coatings Technology* **140,** 51–57 (2001).

190. Schmutz, P. & Frankel, G. Influence of dichromate ions on corrosion of pure aluminum and AA2024-T3 in NaCl solution studied by AFM scratching. *Journal of the Electrochemical Society* **146,** 4461 (1999).

191. Cole, I., Castillo-Robles, M., De Freitas Martins, E. & Ordejón, P. Molecular modeling applied to corrosion inhibition: a critical review. *npj Materials Degradation* **2.** ISSN: 2397-2106. http://dx.doi.org/10.1038/s41529-024-00478-2 (2024).

192. Desimone, M. P., Gordillo, G. & Simison, S. N. The effect of temperature and concentration on the corrosion inhibition mechanism of an amphiphilic amido-amine in CO2 saturated solution. *Corrosion Science* **53,** 4033–4043. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2011.08.009 (2011).

193. Qafsaoui, W., Kendig, M. W., Joiret, S., Perrot, H. & Takenouti, H. Ammonium pyrrolidine dithiocarbamate adsorption on copper surface in neutral chloride media. *Corrosion Science* **106,** 96–107. ISSN: 0010938X. http://dx.doi.org/10.1016/j.corsci.2016.01.029 (2016).

194. Visser, P., Terryn, H. & Mol, J. M. C. in *Active Protective Coatings* (eds Hughes, A. E., Mol, J. M. C., Zheludkevich, M. L. & Buchheit, R. G.) 315–372 (Springer Netherlands, Dordrecht, 2016).

195. Orazem, M. E. & Tribollet, B. *Electrochemical Impedance Spectroscopy* 2nd edition. ISBN: 978-1118527399 (John Wiley & Sons, Inc., 2017).

196. Bard, A. J., Faulkner, L. R. & White, H. S. *Electrochemical Methods: Fundamentals and Applications* 3rd edition (John Wiley & Sons, 2022).

197. Yasuda, H., Yu, Q. & Chen, M. Interfacial factors in corrosion protection: an EIS study of model systems. *Progress in Organic Coatings* **41,** 273–279. ISSN: 03009440. https://linkinghub.elsevier.com/retrieve/pii/S0300944001001424 (May 2001).

198. Lazanas, A. C. & Prodromidis, M. I. Electrochemical Impedance Spectroscopy A Tutorial. *ACS Measurement Science Au* **3,** 162–193. ISSN: 2694250X (2023).

199. Rosero-Navarro, N. C. *et al.* Optimization of hybrid sol-gel coatings by combination of layers with complementary properties for corrosion protection of AA2024. *Progress in Organic Coatings* **69,** 167–174. ISSN: 03009440. http://dx.doi.org/10.1016/j.porgcoat.2010.04.013 (2010).

200. Hsu, C. H. & Mansfeld, F. Technical Note: Concerning the Conversion of the Constant Phase Element Parameter Y 0 into a Capacitance. *CORROSION* **57,** 747–748. ISSN: 0010-9312. https://meridian.allenpress.com/corrosion/article/57/9/747/161974/Technical-Note-Concerning-the-Conversion-of-the (Sept. 2001).

201. R. Williams. *pKa Values in Water Compilation* 2022. https://organicchemistrydata.org/hansreich/resources/pka/pka_data/pka-compilation-williams.pdf (2024).

202. Saeidi, N., Harnisch, F., Presser, V., Kopinke, F. D. & Georgi, A. Electrosorption of organic compounds: State of the art, challenges, performance, and perspectives. *Chemical Engineering Journal* **471,** 144354. ISSN: 13858947. https://doi.org/10.1016/j.cej.2023.144354 (2023).

203. Laurent, C., Scenini, F., Monetta, T., Bellucci, F. & Curioni, M. The contribution of hydrogen evolution processes during corrosion of aluminium and aluminium alloys investigated by potentiodynamic polarisation coupled with real-time hydrogen measurement. *npj Materials Degradation* **1,** 1–7. ISSN: 23972106. http://dx.doi.org/10.1038/s41529-017-0011-4 (2017).

204. Kwiatkowski, L., Grobelny, M. & Konarski, P. Selection of processing parameters for the conversion coatings on high-strength aluminum alloys by cyclic voltammetry. *Materials Science* **50,** 13–22. ISSN: 1573885X (2015).

205. Özkan, C. *et al.* Quasi-stable adsorption as a stepping stone to stable corrosion inhibition. *Applied Surface Science,* 164060 (2025).

206. Bender, R. *et al.* Corrosion challenges towards a sustainable society. *Materials and corrosion* **73,** 1730–1751 (2022).

207. Aliofkhazraei, M. Corrosion inhibitors, principles and recent applications (2018).

208. Chaubey, N., Qurashi, A., Chauhan, D. S., Quraishi, M., *et al.* Frontiers and advances in green and sustainable inhibitors for corrosion applications: A critical review. *Journal of Molecular Liquids* **321,** 114385 (2021).

209. Ma, I. W., Ammar, S., Kumar, S. S., Ramesh, K. & Ramesh, S. A concise review on corrosion inhibitors: types, mechanisms and electrochemical evaluation studies. *Journal of Coatings Technology and Research,* 1–28 (2022).

210. Antonijevic, M. & Petrovic, M. Copper corrosion inhibitors. A review. *International journal of electrochemical science* **3,** 1–28 (2008).

211. Ahmed, M. A., Amin, S. & Mohamed, A. A. Current and emerging trends of inorganic, organic and eco-friendly corrosion inhibitors. *RSC advances* **14,** 31877–31920 (2024).

212. Hughes, A. E., Mol, J. M. C., Zheludkevich, M. L. & Buchheit, R. G. Active protective coatings. *Active Protective Coatings: New-generation Coatings for Metals, Springer Series in Materials Science* **233** (2016).

213. Hughes, A. E. *et al.* Corrosion inhibition, inhibitor environments, and the role of machine learning. *Corrosion and Materials Degradation* **3,** 672–693 (2022).

214. Ma, J. *et al.* Data-driven corrosion inhibition efficiency prediction model incorporating 2D–3D molecular graphs and inhibitor concentration. *Corrosion Science* **222,** 111420 (2023).

215. Gong, H., Fu, Z., Ma, L. & Zhang, D. Inhibitor_Mol_VAE: a variational autoencoder approach for generating corrosion inhibitor molecules. *npj Materials Degradation* **8,** 102 (2024).

216. Schiessler, E. J. *et al.* Searching the chemical space for effective magnesium dissolution modulators: a deep learning approach using sparse features. *npj Materials degradation* **7,** 74 (2023).

217. Feiler, C., Mei, D., Luthringer-Feyerabend, B., Lamaka, S. & Zheludkevich, M. Rational design of effective Mg degradation modulators. *Corrosion* **77,** 204–208 (2021).

218. Recloux, I. *et al.* Active and passive protection of AA2024-T3 by a hybrid inhibitor doped mesoporous sol–gel and top coating system. *Surface and Coatings Technology* **303,** 352–361 (2016).

219. Li, J. F. *et al.* Shell-isolated nanoparticle-enhanced Raman spectroscopy. *nature* **464,** 392–395 (2010).

220. *Chemicalize* https://chemicalize.com/. Accessed: 2025-02-04.

221. Southan, C. & Stracz, A. Extracting and connecting chemical structures from text sources using chemicalize. org. *Journal of cheminformatics* **5,** 1–10 (2013).

222. *pKa calculation | Chemaxon Docs* https://docs.chemaxon.com/display/lts-krypton/pka-calculation.md. Accessed: 2025-02-04.

223. Neese, F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2,** 73–78 (2012).

224. Neese, F. Software update: The ORCA program system—Version 5.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **12,** e1606 (2022).

225. Bursch, M., Mewes, J.-M., Hansen, A. & Grimme, S. Best-practice DFT protocols for basic molecular computational chemistry. *Angewandte Chemie International Edition* **61,** e202205735 (2022).

226. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of physical chemistry* **98,** 11623–11627 (1994).

227. Becke, A. D. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *The Journal of chemical physics* **96,** 2155–2160 (1992).

228. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical review B* **37,** 785 (1988).

229. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry* **32,** 1456–1465 (2011).

230. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **7,** 3297–3305 (2005).

231. Helmich-Paris, B., de Souza, B., Neese, F. & Izsák, R. An improved chain of spheres for exchange algorithm. *The Journal of Chemical Physics* **155** (2021).

232. Barone, V. & Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *The Journal of Physical Chemistry A* **102,** 1995–2001 (1998).

233. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *The Journal of Physical Chemistry B* **113,** 6378–6396 (2009).

234. Parr, R. G., Szentpály, L. v. & Liu, S. Electrophilicity index. *Journal of the American Chemical Society* **121,** 1922–1924 (1999).

235. Sadeghi, A. *et al.* Multiscale approach for simulations of Kelvin probe force microscopy with atomic resolution. *Physical Review B—Condensed Matter and Materials Physics* **86,** 075407 (2012).

236. Rahimi, E. *et al.* Morphological and surface potential characterization of protein nanobiofilm formation on magnesium alloy oxide: their role in biodegradation. *Langmuir* **38,** 10854–10866 (2022).

237. Liscio, A., Palermo, V. & Samori, P. Nanoscale quantitative measurement of the potential of charged nanostructures by electrostatic and Kelvin probe force microscopy: unraveling electronic processes in complex materials. *Accounts of chemical research* **43,** 541–550 (2010).

238. Lübben, J. F. *et al.* Tuning the surface potential of Ag surfaces by chemisorption of oppositely-oriented thiolated carborane dipoles. *Journal of colloid and interface science* **354,** 168–174 (2011).

239. Örnek, C., Leygraf, C. & Pan, J. On the Volta potential measured by SKPFM–fundamental and practical aspects with relevance to corrosion science. *Corrosion Engineering, Science and Technology* **54,** 185–198 (2019).

240. Kahn, A., Koch, N. & Gao, W. Electronic structure and electrical properties of interfaces between metals and π-conjugated molecular films. *Journal of Polymer Science Part B: Polymer Physics* **41,** 2529–2548 (2003).

241. Schmutz, P. & Frankel, G. Characterization of AA2024-T3 by scanning Kelvin probe force microscopy. *Journal of the Electrochemical Society* **145,** 2285 (1998).

242. Schmutz, P. & Frankel, G. Corrosion study of AA2024-T3 by scanning kelvin probe force microscopy and in situ atomic force microscopy scratching. *Journal of the Electrochemical Society* **145,** 2295 (1998).

243. Zhu, Y., Sun, K. & Frankel, G. Intermetallic phases in aluminum alloys and their roles in localized corrosion. *Journal of The Electrochemical Society* **165,** C807 (2018).

244. Finšgar, M. EQCM and XPS analysis of 1, 2, 4-triazole and 3-amino-1, 2, 4-triazole as copper corrosion inhibitors in chloride solution. *Corrosion Science* **77,** 350–359 (2013).

245. Artemenko, A. *et al. Reference XPS spectra of amino acids* in *IOP Conference Series: Materials Science and Engineering* **1050** (2021), 012001.

246. Graf, N. *et al.* XPS and NEXAFS studies of aliphatic and aromatic amine species on functionalized surfaces. *Surface Science* **603,** 2849–2860 (2009).

247. Stevens, J. S. *et al.* Quantitative analysis of complex amino acids and RGD peptides by X-ray photoelectron spectroscopy (XPS). *Surface and Interface Analysis* **45,** 1238–1246 (2013).

248. Chastain, J. & King Jr, R. C. Handbook of X-ray photoelectron spectroscopy. *Perkin-Elmer Corporation* **40,** 25 (1992).

249. Han, Y. F., Fu, T. & Shen, Y. Nanostructural C- Al- N thin films studied by x-ray photoelectron spectroscopy, Raman and high-resolution transmission electron microscopy. *Journal of Materials Research* **24,** 3321–3330 (2009).

250. Wrzosek, B. & Bukowska, J. Molecular structure of 3-amino-5-mercapto-1, 2, 4-triazole self-assembled monolayers on Ag and Au surfaces. *The Journal of Physical Chemistry C* **111,** 17397–17403 (2007).

251. Xia, Z., Baird, L., Zimmerman, N. & Yeager, M. Heavy metal ion removal by thiol functionalized aluminum oxide hydroxide nanowhiskers. *Applied Surface Science* **416,** 565–573 (2017).

252. Peak, D., Ford, R. G. & Sparks, D. L. An in situ ATR-FTIR investigation of sulfate bonding mechanisms on goethite. *Journal of colloid and interface science* **218,** 289–299 (1999).

253. Secco, E. A. Spectroscopic properties of SO4 (and OH) in different molecular and crystalline environments. I. Infrared spectra of Cu4 (OH) 6SO4, Cu4 (OH) 4OSO4, and Cu3 (OH) 4SO4. *Canadian journal of chemistry* **66,** 329–336 (1988).

254. Cabassi, F., Casu, B. & Perlin, A. S. Infrared absorption and Raman scattering of sulfate groups of heparin and related glycosaminoglycans in aqueous solution. *Carbohydrate Research* **63,** 1–11 (1978).

255. Kiefer, J., Stärk, A., Kiefer, A. L. & Glade, H. Infrared spectroscopic analysis of the inorganic deposits from water in domestic and technical heat exchangers. *Energies* **11,** 798 (2018).

256. Radha, A., Lander, L., Rousse, G., Tarascon, J. & Navrotsky, A. Thermodynamic stability and correlation with synthesis conditions, structure and phase transformations in orthorhombic and monoclinic Li 2 M (SO 4) 2 (M= Mn, Fe, Co, Ni) polymorphs. *Journal of Materials Chemistry A* **3,** 2601–2608 (2015).

257. Xi, S. *et al.* Micro-Raman study of thermal transformations of sulfide and oxysalt minerals based on the heat induced by laser. *Minerals* **9,** 751 (2019).

258. Meng, S., Zhao, Y., Xue, J. & Zheng, X. Environment-dependent conformation investigation of 3-amino-1, 2, 4-triazole (3-AT): Raman Spectroscopy and density functional theory. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **190,** 478–485 (2018).

259. Nowakowska-Langier, K. *et al.* Phase composition of copper nitride coatings examined by the use of X-ray diffraction and Raman spectroscopy. *Journal of Molecular Structure* **1165,** 79–83 (2018).

260. Rodríguez-Tapiador, M. I. *et al.* Impact of the rf power on the copper nitride films deposited in a pure nitrogen environment for applications as eco-friendly solar absorber. *Materials* **16,** 1508 (2023).

261. Deng, Y., Handoko, A. D., Du, Y., Xi, S. & Yeo, B. S. In situ Raman spectroscopy of copper and copper oxide surfaces during electrochemical oxygen evolution reaction: identification of CuIII oxides as catalytically active species. *Acs Catalysis* **6,** 2473–2481 (2016).

262. Li, X., Zhou, C., Jiang, G. & You, J. Raman analysis of aluminum nitride at high temperature. *Materials characterization* **57,** 105–110 (2006).

263. Wang, Y. *et al.* Qualitative and quantitative detection of corrosion inhibitors using surface-enhanced Raman scattering coupled with multivariate analysis. *Applied Surface Science* **568,** 150967 (2021).

264. Teng, D., Ma, J., Huang, Y., Zhang, X. & Zheng, R. *Investigate seawater and seawater anions' aqueous mixed solution by laser Raman spectroscopy* in *Semiconductor Lasers and Applications III* **6824** (2008), 130–139.

265. Chou, I.-M. & Wang, A. Application of laser Raman micro-analyses to Earth and planetary materials. *Journal of Asian Earth Sciences* **145,** 309–333 (2017).

266. Kloprogge, J. & Frost, R. Raman microscopy study of basic aluminum sulfate. *Journal of materials science* **34,** 4199–4202 (1999).

267. Hammer, B. & Nørskov, J. K. in *Advances in catalysis* 71–129 (Elsevier, 2000).

268. Qiu, Y. *et al.* Insight into synergistic corrosion inhibition of 3-amino-1, 2, 4-triazole-5-thiol (ATT) and NaF on magnesium alloy: Experimental and theoretical approaches. *Corrosion Science* **208,** 110618 (2022).

269. Damej, M. *et al.* Corrosion inhibition of brass 60Cu–40Zn in 3% NaCl solution by 3-amino-1, 2, 4-triazole-5-thiol. *Heliyon* **6** (2020).

270. Jiang, B., Jiang, S., Liu, X., Ma, A. & Zheng, Y. Corrosion inhibition performance of triazole derivatives on Copper-Nickel alloy in 3.5 wt.% NaCl solution. *Journal of Materials Engineering and Performance* **24,** 4797–4808 (2015).

271. Tassaoui, K. *et al.* Contribution to the corrosion inhibition of Cu–30Ni copper–nickel alloy by 3-amino-1, 2, 4-triazole-5-thiol (ATT) in 3% NaCl solution. Experimental and theoretical study (DFT, MC and MD). *Int J Corros Scale Inhib* **11,** 221–44 (2022).

272. Sherif, E.-S. M. Corrosion inhibition in chloride solutions of iron By 3-Amino-1, 2, 4-Triazole-5-Thiol and 1, 1'-Thiocarbonyldiimidazole. *International Journal of Electrochemical Science* **7,** 4834–4846 (2012).

273. Sherif, E.-S. M., Erasmus, R. & Comins, J. In situ Raman spectroscopy and electrochemical techniques for studying corrosion and corrosion inhibition of iron in sodium chloride solutions. *Electrochimica Acta* **55,** 3657–3663 (2010).

274. Udoh, I. I., Shi, H., Soleymanibrojeni, M., Liu, F. & Han, E.-H. Inhibition of galvanic corrosion in Al/Cu coupling model by synergistic combination of 3-Amino-1, 2, 4-triazole-5-thiol and cerium chloride. *Journal of Materials Science & Technology* **44,** 102–115 (2020).

275. Udoh, I. I., Shi, H., Liu, F. & Han, E.-H. Synergistic effect of 3-amino-1, 2, 4-triazole-5-thiol and cerium chloride on corrosion inhibition of AA2024-T3. *Journal of The Electrochemical Society* **166,** C185 (2019).

276. Sherif, E.-S. M. Electrochemical investigations on the corrosion inhibition of aluminum by 3-amino-1, 2, 4-triazole-5-thiol in naturally aerated stagnant seawater. *Journal of Industrial and Engineering Chemistry* **19,** 1884–1889 (2013).

277. Liu, G. *et al.* Understanding the hydrophobic mechanism of 3-hexyl-4-amino-1, 2, 4-triazole-5-thione to malachite by ToF-SIMS, XPS, FTIR, contact angle, zeta potential and micro-flotation. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **503,** 34–42 (2016).

278. Abd El-Ghaffar, M., Mohamed, M. & Elwakeel, K. Adsorption of silver (I) on synthetic chelating polymer derived from 3-amino-1, 2, 4-triazole-5-thiol and glutaraldehyde. *Chemical engineering journal* **151,** 30–38 (2009).

279. Kuznetsov, Y. I. *The role of irreversible adsorption in the protective action of volatile corrosion inhibitors* in *NACE CORROSION* (1998), NACE–98242.

280. Andreev, N. N. & Kuznetsov, Y. I. Physicochemical aspects of the action of volatile metal corrosion inhibitors. *Russian chemical reviews* **74,** 685 (2005).

281. Chiter, F. *et al.* DFT study of Cl- ingress into organic self-assembled monolayers on aluminum. *Journal of The Electrochemical Society* **170,** 071504 (2023).

282. Özkan, C. *et al.* Gaining scientific understanding with small data machine learning: explainable molecule representations and their consensus. *npj Materials Degradation* (2025).

283. Mater, A. C. & Coote, M. L. Deep learning in chemistry. *Journal of chemical information and modeling* **59,** 2545–2559 (2019).

284. Yano, J. *et al.* The case for data science in experimental chemistry: examples and recommendations. *Nature Reviews Chemistry* **6,** 357–370 (2022).

285. Karande, P., Gallagher, B. & Han, T. Y.-J. A strategic approach to machine learning for material science: how to tackle real-world challenges and avoid pitfalls. *Chemistry of Materials* **34,** 7650–7665 (2022).

286. Rodrigues, J. F., Florea, L., de Oliveira, M. C., Diamond, D. & Oliveira, O. N. Big data and machine learning for materials science. *Discover Materials* **1,** 1–27 (2021).

287. Chong, S. S., Ng, Y. S., Wang, H.-Q. & Zheng, J.-C. Advances of machine learning in materials science: Ideas and techniques. *Frontiers of Physics* **19,** 13501 (2024).

288. Choudhary, K. *et al.* Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **8,** 59 (2022).

289. Noh, J. *et al.* Inverse design of solid-state materials via a continuous representation. *Matter* **1,** 1370–1384 (2019).

290. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **4,** 268–276 (2018).

291. Reiser, P. *et al.* Graph neural networks for materials science and chemistry. *Communications Materials* **3,** 93 (2022).

292. Zhong, Z., Li, C.-T. & Pang, J. Hierarchical message-passing graph neural networks. *Data Mining and Knowledge Discovery* **37,** 381–408 (2023).

293. Li, K., DeCost, B., Choudhary, K., Greenwood, M. & Hattrick-Simpers, J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Computational Materials* **9,** 55 (2023).

294. Krenn, M. *et al.* On scientific understanding with artificial intelligence. *Nature Reviews Physics* **4,** 761–769 (2022).

295. Friederich, P., Krenn, M., Tamblyn, I. & Aspuru-Guzik, A. Scientific intuition inspired by machine learning-generated hypotheses. *Machine Learning: Science and Technology* **2,** 025027 (2021).

296. Zhong, X. *et al.* Explainable machine learning in materials science. *npj computational materials* **8,** 204 (2022).

297. Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E. & Hoffmann, H. *Explainability methods for graph convolutional neural networks* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), 10772–10781.

298. Jin, W., Barzilay, R. & Jaakkola, T. *Multi-objective molecule generation using interpretable substructures* in *International conference on machine learning* (2020), 4849–4859.

299. Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research* **3,** 597–607 (2022).

300. Vu, T.-S. *et al.* Towards understanding structure–property relations in materials with interpretable deep learning. *npj Computational Materials* **9,** 215 (2023).

301. Pham, T. H., Le, P. K., *et al.* A data-driven QSPR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques. *RSC advances* **14,** 11157–11168 (2024).

302. Ribeiro, M. T., Singh, S. & Guestrin, C. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016* (2016), 1135–1144.

303. Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games* **2,** 307–317 (1953).

304. Diao, Y., Yan, L. & Gao, K. Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features. *Materials & Design* **198,** 109326 (2021).

305. Noutahi, E. *et al. datamol-io/molfeat* version 0.9.4. 2023. https://github.com/datamol-io/molfeat.

306. Zherebtsov, D. *Verstack* https://github.com/DanilZherebtsov/verstack. 2020.

307. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html (2011).

308. Koops, R. *Probatus: Validation (like Recursive Feature Elimination for SHAP) of (multiclass) classifiers regressors and data used to develop them.* https://github.com/ing-bank/probatus. 2023.

309. Ho, T. K. *Random decision forests* in *Proceedings of 3rd international conference on document analysis and recognition* **1** (1995), 278–282.

310. Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20,** 273–297 (1995).

311. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46,** 175–185 (1992).

312. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, Aug. 2016), 785–794. ISBN: 9781450342322. https://dl.acm.org/doi/10.1145/2939672.2939785.

313. Nogueira, F. *Bayesian Optimization: Open source constrained global optimization tool for Python* 2014. https://github.com/bayesian-optimization/BayesianOptimization.

314. Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (O'reilly, 2019).

315. Gadaleta, D. *et al.* SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. *Journal of Cheminformatics* **11,** 1–16. ISSN: 17582946. https://doi.org/10.1186/s13321-019-0383-2 (2019).

316. U.S. National Archives and Records Administration. *Code of federal regulations, protection of environment, title 40, sec. 156.62* 2006. https://www.ecfr.gov/current/title-40/part-156/section-156.62.

317. Lundberg, S. M. & Lee, S.-I. in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 4765–4774 (Curran Associates, Inc., 2017). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

318. Allouche, A.-r. Software News and Updates Gabedit — A Graphical User Interface for Computational Chemistry Softwares. *Journal of computational chemistry* **32,** 174–182. ISSN: 1096-987X (2012).

319. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* **42,** 1273–1280. ISSN: 00952338 (2002).

320. Smith, D. H., Carhart, R. E. & Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *Journal of Chemical Information and Computer Sciences* **25,** 64–73. ISSN: 00952338 (1985).

321. Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Transactions on Computers* **C-22,** 1025–1034. ISSN: 00189340 (1973).

322. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50,** 742–754. ISSN: 1549-9596. arXiv: ci100050 [10.1021]. https://pubs.acs.org/doi/10.1021/ci100050t (May 2010).

323. York, J. T., Bar-Nahum, I. & Tolman, W. B. Copper–sulfur complexes supported by N-donor ligands: Towards models of the CuZ site in nitrous oxide reductase. *Inorganica chimica acta* **361,** 885–893 (2008).

324. Ganjoo, R. & Kumar, A. Current trends in anti-corrosion studies of surfactants on metals and alloys. *Journal of Bio-and Tribo-Corrosion* **8,** 1–35 (2022).

325. Abdelmonem, H., Al-Bonayan, A. M. & Fouda, A. E.-A. S. Some Surfactants as Corrosion Inhibitors for Carbon Steel in Acidic Solutions. *Surface Engineering and Applied Electrochemistry* **58,** 412–423 (2022).

326. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (John Wiley & Sons, 2009).

327. Barysz, M., Jashari, G., Lall, R. S., Srivastava, V. K. & Trinajstic, N. On the distance matrix of molecules containing heteroatoms. Chemical applications of topology and graph theory. *Studies in physical and theoretical chemistry* (1983).

328. Levitt, M. & Perutz, M. F. Aromatic rings act as hydrogen bond acceptors. *Journal of molecular biology* **201,** 751–754 (1988).

329. Huang, S. *et al.* Hydrogen Bond Induces Hierarchical Self-Assembly in Liquid-Crystalline Block Copolymers. *Macromolecular Rapid Communications* **39,** 1700783 (2018).

330. Sikder, A. & Ghosh, S. Hydrogen-bonding regulated assembly of molecular and macromolecular amphiphiles. *Materials Chemistry Frontiers* **3,** 2602–2616 (2019).

331. Manish Sud. *MACCS (Molecular ACCess System) documentation* 2024. http://www.mayachemtools.org/docs/modules/pdf/MACCSKeys.pdf.

332. Ren, C., Ma, L., Zhang, D., Li, X. & Mol, A. High-throughput experimental techniques for corrosion research: A review. *Materials Genome Engineering Advances* **1,** e20 (2023).

333. Yang, J. *et al.* Combinatorial discovery and investigation of the synergism of green amino acid corrosion inhibitors: Integrating high-throughput experiments and interpretable machine learning approach. *Corrosion Science* **245,** 112675 (2025).

334. Cornet, A., Homborg, A., Mol, J. M. C., *et al.* Corrosion protective performance evaluation of structural aircraft coatings in cyclic salt spray, outdoor and In-Service environments. *Engineering Failure Analysis,* 109566 (2025).

335. Scalfani, V. F., Patel, V. D. & Fernandez, A. M. Visualizing chemical space networks with RDKit and NetworkX. *Journal of Cheminformatics* **14,** 87 (2022).

List of Publications

Related to the dissertation

Self-sustaining non-toxic corrosion inhibition compositions for metallic substrates (Patent Application No. P100996NL00). *C. Özkan*, A. Kooijman, P. Taheri, J.M.C. Mol, Netherlands Patent Office, 2025.

Gaining scientific understanding with small data machine learning: explainable molecule representations and their consensus. *C. Özkan*, L. Sahlmann, T. Würger, C. Feiler, S.V. Lamaka, M.L. Zheludkevich, P. Taheri, J.M.C. Mol, npj Materials Degradation, 2025.

Quasi-stable adsorption as a stepping stone to stable corrosion inhibition. *C. Özkan*, A.M. Armaki, E. Rahimi, P.R. Anusuyadevi, H. Nie, Y. Hedberg, P. Taheri, J.M.C. Mol, Applied Surface Science, 2025.

Factors to consider in the quest for organic alternatives to hexavalent chromium based corrosion inhibitors. *C. Özkan*, P.R. Anusuyadevi, P. Visser, P. Taheri, J.M.C. Mol, Corrosion Science, 2025.

Impact of inhibition mechanisms, automation, and computational models on the discovery of organic corrosion inhibitors. D. A. Winkler, A. E. Hughes, *C. Özkan*, J.M.C. Mol, Tim Würger, C. Feiler, D. Zhang, S.V. Lamaka , Progress in Materials Science, 2024.

Laying the experimental foundation for corrosion inhibitor discovery through machine learning. *C. Özkan*, L. Sahlmann, C. Feiler, M. Zheludkevich, S. V. Lamaka, P. Sewlikar, A. Kooijman, P. Taheri, J.M.C. Mol, npj Materials Degradation, 2024.

Other works

Leveraging machine learning for the performance prediction of organic corrosion inhibitors for aluminium alloys. L. Sahlmann, *C. Özkan*, N. Konchakova, J.M.C. Mol, D. A. Winkler, S.V. Lamaka, M.L. Zheludkevich, C. Feiler, in-preparation for npj Materials Degradation.

Corrosion protection in a coating defect on AA2024-T3 by lithium carbonate inhibitor leaching: An experimentally validated FEM approach. N. Abdelrahman, N. Van den Steen, *C. Özkan*, C. Wang, P. Visser, S.V. Lamaka , S. Kallip, R. Böttcher, J.M.C. Mol, M.L. Zheludkevich, H. Terryn, T. Hauffman, M. Meeusen, Corrosion Science, 2025.

Artificial intelligence, machine learning, and big data for corrosion control – quo vadis. D.A. Winkler, A.E. Hughes, *C. Özkan*, J.M.C. Mol, T. Würger, C. Feiler, S.V. Lamaka, The Australasian Corrosion Association, 2023.

From experimental and computational inhibitor screening to advanced characterization of active protective coatings. *C. Özkan*, J.M.C. Mol, The Australasian Corrosion Association, 2023.

Properties of passive films formed on ferrite-martensite and ferrite-pearlite steel microstructures. A. Yilmaz, *C. Özkan*, J. Sietsma, Y. Gonzalez-Garcia, Metals, 2021.

List of Presentations

Talks

Gaining scientific understanding through small dataset machine learning.
DIFFER - TU/e Center for Computational Energy Research Seminar (Eindhoven, the Netherlands, 2025).

Using statistical models as materials characterization tools for scientific insight.
MaterialenNL - Materials Innovation Institute m2i Conference 2024 (Papendal, the Netherlands, 2024).

From statistical models to scientific insight in corrosion inhibition electrochemistry.
Aqueous Corrosion Gordon Research Seminar (New London, United States, 2024).

Machine learning and ’small’ data for hexavalent-chromium-free corrosion inhibition of airplanes.
Studiekern Corrosie 75 year Anniversary Event (Delft, the Netherlands, 2023).

Evaluating organic corrosion inhibitors as chromate replacements through electrochemical experiments.
EUROCORR 2023 Annual European Corrosion Congress (Brussels, Germany, 2023).

Time-resolved electrochemical studies of AA2024-T3 corrosion inhibition: sodium dichromate vs. high-performing organic inhibitors.
ASST2023 Aluminum Surface Science and Technology Symposium (Stockholm, Sweden, 2023).

‘Small data’s big potential for functional molecule discovery.
Delft University of Technology Materials Science & Engineering Department Performance & Recycling Talks (Online, 2022).

Time-resolved analysis of corrosion inhibitor layer irreversibility.
EUROCORR 2022 Annual European Corrosion Congress (Berlin, Germany, 2022).

Laying the experimental foundation for machine learning in corrosion inhibitor discovery.
Corrosion and Materials Degradation Web Conference (Online, 2022).

Towards linking electrochemical and machine learning based assessment of corrosion inhibitor efficiency.
AETOC 2022 12th International Workshop on Application of Electrochemical Techniques to Organic Coatings (Val di Fiemme, Italy, 2022).

POSTERS

From statistical models to scientific insight in corrosion inhibition electrochemistry.
Generative Modeling Summer School (Eindhoven, the Netherlands, 2024),
Aqueous Corrosion Gordon Research Conference (New London, United States, 2024).

Molecule discovery: Man vs. the Machine.
OIP-2023 Open Innovation and Collaborative Decision Making for Materials Design and Manufacturing Conference (Belval, Luxembourg, 2023),
GC-MAC Material Acceleration Consortium Summer School (Karlsruhe, Germany, 2023),
TU Delft Materials Science and Engineering Department Materials Technology Day 2023 Artificial Intelligence (Delft, the Netherlands, 2023).

Hexavalent-chromium-free corrosion inhibitor performance for active protective coatings.
Delft University of Technology Materials Science & Engineering Department Performance & Recycling Talks (Online, 2021).

Acknowledgements

I cannot name everyone for the wonderful suggestions, discussions, and all the critical feedback I got that resulted in this book. I'm lucky to have many wonderful, stubborn people in my life - not exclusively scientists - who make disagreeing with the most minute, inane things their hobby (I am looking at you "the scientifically right way to eat a tompoes") which helped tremendously. So I'll instead try to paint a picture of the characters of my story.

On the façade of our 3mE - sorry ME faculty, there stand four statues. Around the campus, many stories were told about their meaning, but for me, they have always represented the muses: inspirational goddesses of literature, arts, philosophy, and science. Even back in the day, science must have been so temperamental, filled with such uncertainty and doubt, that the Greeks needed support from their capricious divine patrons. After spending two years in masters, and four in the doctorate, looking at these statues every day as I entered the Materials Science and Engineering Department; now I see inspiration from a different light. Yes, the sudden inspiration that comes under the shower might be the muses of Anatolia, but the inspiration that makes you grow, that keeps you going each and every day - especially in such a long-winded and solitary project - that comes from the muses around you. I was lucky to say I had more than four in my life, and this is my attempt at squeezing all their awesomeness in couple of paragraphs, which unfortunately is bound to get lost in translation.

To start off, Aytaç I'd like to thank you for showing that science can be fun during my masters thesis, which set the direction for the rest of my academic adventures. Quite literally - if you were not in my masters defense committee Peyman, I think my life would've turned out quite different, and I'm very grateful that you gave my younger self the chance to develop into an independent researcher. During that time I learned a lot from you Arjan, and I'll always cherish our chats during trips to project meetings and conferences, which took me all around the world despite COVID. I'm extremely thankful and honoured that you all gave me this opportunity to grow into the shoes of a scientist.

I am so happy to have learned from and endeavoured together with the VIPCOAT consortium, especially for all the unique experiences around Europe, e.g., celebrating my 28th birthday with a tiny cake at a synchrotron night shift thanks to kindness of Claudia, (not) sniffing in deadly chromate compounds thanks to Peter and experimental support of AkzoNobel, gaining gamer cred of Tim and Christian by beating Dark Souls Bell Gargoyles in the evening of a conference over some whisky on a borrowed console... the experiences were truly unique. Thank you all.

For day-to-day, I was extremely lucky to spend time with you Agnieszka, not only because you truly cared every time I needed your support, but also because every case of spending time with you was a blast. I was truly fortunate to start this project with such an amazing student as you Parth (thanks to you I can say 100% of my masters students continued to get a PhD), and to end this project with you Giray (please finish your masters thesis at some point so I can say 100% of my students finished their degrees). Reina it was a gift to get to know your kind heart during so many awesome events we pulled off together - your temperament turned my dread of organising events into something that I look forward to. In between work, I always marvelled at the infinite ability of straw-splitting of my fellow PhD candidates and Postdocs: Jose, Arjun, Tim, Saurabh, Luis, Daniel, Philipp, Sajjad, Camila, Joep, Elsa, Eszter, Soroush, Amir, Khatereh, Jasper, Julia, Fabian, Alice, Keer, Gaojie, Prakash, Arjan, Ehsan, Prasaanth, - thank you all for the quotidian camaraderie. You guys rock.

And after work. Without precious weekends with you guys, whether it is forgetting how to serve at tennis, dealing with boardgame tantrums, or just simply goofing around, these last four years would have been a dull experience. Tess and Juul, sorry for laying siege to your home during COVID, but being stuck together was, to be honest, amazing. Deniz, I'm hopeful that our relationship will outlast Tartufo. Marco, never would have thought that Dutch courses would result in an Italian fratellino - you still need to teach us how to make a true Genovese focaccia. Barış, I promise to keep practicing bass so we can bring disco back. Doruk, please teach me drifting one day *if* I can manage to get a license. Berk, I hope we can keep bringing democracy to our citizens. Mustafa, I love you no matter how many times you force me into a bellydancer costume. Ece, I could not have asked for a better sister. Melis, Yiğit, and our newest member Mete, you guys manage to make Berlin even cooler, that's something. Utku, your trumpet taught me that persistence beats perfection, though your neighbors may disagree. Ekin, I will come to Lausanne one day if you can prove to me that it exists. Ali Ozan, my offer to be your stay-at-home husband when you become a millionaire still stands. İsmail, your saga from heartbreak to rave commander was truly marvelous. Serter, I'm so proud of you that after so many years of education you are finally pursuing your dreams of owning a bakkal. Gautham, I promise you one day you won't have to pull your punches during tennis (possibly after retirement). Jesper, without you in our lives we will get fat, so please continue being our active-life drill sergeant. Jelmer, I am hopeful that you'll complete your character arc and become the absolute crazy scientist that creates a time-machine as an accident (see Steins;Gate). Rohan and Kelsey, can you buy 5 kg. of assault rifles when visiting us, thanks! Ide, I'm waiting for the day you turn Chinese spy, it'll be so much fun. Ines, I've been a proud baby-mama watching you grow. Maud, please keep cooking us yummy food, and Daan, if we combine our powers we can start an uprising against them, trust me. Bas, Wendy and Sam, you have been so amazing opening your hearts from day one, I could not have asked for a more genuine Dutch family experience.

To Mom and Dad, as the first person among our people to study science (as far as we know), you need to know that the title of PhD stands for Doctor of Philosophy, not Doctor of Science, and I'd like to think that is for a reason. The word 'doctor' is derived from the Latin verb 'docere', meaning, to teach. 'Science' comes from Latin 'scientia', or 'to know'; but 'philosophy' translated from ancient Greek stands for 'the love of wisdom'. I remember my first example of philosophia when I was a child half your height, when you pushed books on me such as Alice Travels to Quantum Land without knowing what quantum is (I did not get it at all...). That was the first spark for me to learn that it was not about knowing, it was about the curiosity.

Bana dünyayı sevmeyi öğrettiğiniz için size sonsuza kadar minnettarım.
Sizi çok seviyorum.

This is one of the fruits of that curiosity that you nurtured so gently. And for that, I dedicate this book to you.

And last but not least, Robin. I love to joke that I'd have finished the PhD in 3 years if you were not distracting me. In fact I think I would not have finished it at all. Everything else turns to black without your colour. I love you.

Biography

Can Özkan is a materials scientist, engineer and researcher. He had been pursuing his PhD in Materials Science at the Delft University of Technology for the past 4 years. Can obtained his bachelor's degree in Civil Engineering (2018) with a specialisation in engineering materials and finite element modelling from Boğaziçi University (arguably the university with best views on earth), İstanbul, Turkey. Afterward, he continued his education in materials science with a master's in Materials Science & Engineering (2020) from TU Delft, Delft, the Netherlands. His master thesis studied the relationship between the microstructure and corrosion/passivation behavior of multi-phase high strength low alloy steels. After a brief period of working as a materials scientist in the industry on thermoelectric and battery anode materials development, he joined the Corrosion Technology and Electrochemistry group in May 2021 as a PhD candidate.

For the last 4 years, Can has been working on corrosion inhibitor discovery as an experimental scientist under the supervision of Dr. Peyman Taheri and Dr. Arjan Mol. His research aimed to improve the understanding of corrosion inhibition and self-healing mechanisms of coatings. He analyzed inorganic conversion layers and organic self-assembled thin films with surface analytical and electrochemical experiments at micro to nano-scale, density functional theory simulations, and predictive machine-learning relationships.

Next to his scientific work, he volunteered to be the PhD liaison of TUBalkain, student society of the Materials Science and Engineering Department of TU Delft, to promote the rights of graduate students. He was scientific communications officer and board member for yEFC, the youth branch of the European Federation of Corrosion. He assisted with the organization of AETOC24 (Application of Electrochemical Techniques to Organic Coatings Conference) as the scientific secretary, and created the 4TU.HTM - TU Delft joint workshop on the Role of Machine Learning in Molecular Discovery and Scientific Understanding.

Lately, he is slowly becoming better at scuba-diving and tennis. When he is not working he uses his experimentation skills to brew the perfect cup of coffee.