ARTICLE   **OPEN**

Check for updates

High temperature oxidation of corrosion resistant alloys from machine learning

Christopher D. Taylor [ID][1 ✉] and Brett M. Tossey [ID][1 ✉]

Parabolic rate constants, k_p, were collected from published reports and calculated from corrosion product data (sample mass gain or corrosion product thickness) and tabulated for 75 alloys exposed to temperatures between ~800 and 2000 K (~500–1700 °C; 900–3000 °F). Data were collected for environments including lab air, ambient and supercritical carbon dioxide, supercritical water, and steam. Materials studied include low- and high-Cr ferritic and austenitic steels, nickel superalloys, and aluminide materials. A combination of Arrhenius analysis, simple linear regression, supervised and unsupervised machine learning methods were used to investigate the relations between composition and oxidation kinetics. The supervised machine learning techniques produced the lowest mean standard errors. The most significant elements controlling oxidation kinetics were Ni, Cr, Al, and Fe, with Mo and Co composition also found to be significant features. The activation energies produced from the machine learning analysis were in the correct distributions for the diffusion constants for the oxide scales expected to dominate in each class.

INTRODUCTION

High-temperature oxidation is a significant contributor to the failure of energy generation system components, affecting combustion systems (turbines and engines), heat exchangers, and furnaces[1]. The rate of high-temperature oxidation is most commonly mitigated through appropriate materials selection and environmental controls. The materials selection process is advised through a combination of field experience, long-term exposure as well as shorter-term accelerated tests performed in the laboratory. Imposing environmental controls ultimately involves limiting performance and/or requiring additional quality control steps over gas streams, temperature and pressure monitoring, as well as regular or risk-based maintenance and inspection.

Alloy design strategies typically consist of incorporating elements that readily form oxides having good rate-limiting barrier properties. These include the elements titanium, aluminum, and chromium. Small amounts of other reactive elements, such as hafnium, can assist in improving the integrity of the barrier films by promoting scale adhesion to the matrix and/or binding impurities, such as sulfur[2]. Whereas it is desirable to form kinetically limiting barrier oxides, the change in ratios of alloying elements near the surface as a result of oxidation can cause undesirable changes in the near-surface microstructure, potentially compromising other properties such as creep resistance, through the dissolution of strengthening precipitate phases or changes in, for example, the content of grain boundary chromium carbides that inhibit secondary creep[3,4].

To improve the efficiency of fossil energy generation systems, next-generation designs are targeting higher temperatures, placing additional demands on materials. In particular, systems using supercritical CO_2 as the heat exchange fluid are targeting pressures (200 bar) and temperatures (~1100 °C) that are projected to require extremely corrosion-resistant materials with a target parabolic rate constant for oxidation of 5×10^{-13} g^2/cm^4/s units[5]. Candidate materials being explored to meet these targets are most frequently in the class of nickel-based superalloys[6]. Science-based approaches using computational modeling as a complement to synthesis and characterization of new materials provide an opportunity to further improve on the properties of these materials.

Computationally assisted materials design (CAMD) is entering a second wave of innovation. The first wave corresponded to the development of multiphysics modeling methods: thermodynamic approaches such as CALPHAD[7,8], phase-field models[9], and atomistic simulations[10] that predict the properties of materials systems from first principles. In such approaches, various schemes were put in place to combine computational models in serial or parallel to allow synchronistic modeling of atomistic, mesoscale, and continuum processes with methods like density functional theory[11], kinetic Monte Carlo[12] and finite element approaches[13]. The second and contemporary wave of CAMD includes and builds upon this science-based foundation by adding the potent opportunities offered by the application of machine learning and data informatics[14–16]. This opportunity was recognized early on by the Materials Genome Initiative[17], and has since spawned several other initiatives such as the Materials Project[18]. An organizing principle for conducting informatics projects related to corrosion of materials was produced by Taylor in 2015 and used a hierarchical approach to organize the various sources of knowledge and information related to problems in corrosion science and engineering[19]. In the present work, the focus is on applying a materials informatics approach to compiling and visualizing the data on the oxidation rate constants of candidate high-temperature materials, and then developing a machine learning model for the classification and prediction of materials according to high-temperature oxidation rate constant.

Prior attempts to predict the rate of high-temperature oxidation for alloys based on their composition have included genetic and/or regression algorithms based on notional ideas of the functional dependencies of oxide layer transport properties on composition. Most approaches rely in some way on the Wagner model for oxide growth[20], which predicts a $t^{1/2}$ power law for corrosion, in which case the corrosion performance of a material can be quantified by a characteristic parabolic rate constant which may vary by

[1]DNV GL USA Inc, Dublin, OH, USA. ✉email: Christopher.Taylor@dnv.com; Brett.Tossey@dnv.com

C.D. Taylor and B.M. Tossey

temperature and chemical environment for any given material[21]. In this form, the square of the weight change per unit area is linear with respect to time, with the proportionality constant k_p:

$$\left(\frac{\Delta W}{A}\right)^2 = k_p t \tag{1}$$

The parabolic rate constant, k_p, is often given in units of $g^2/cm^4/s$. Microstructural as well as composition factors will also influence the corrosion performance of the alloy and alter the parabolic rate constant.

We note here that k_p may mask a multiplicity of processes in addition to mass gained due to the chemical oxidation of alloy elements. For example, if oxides formed are volatile, then k_p calculated in this way will underestimate the extent of oxidation since it does not take into account the mass loss when materials volatilize. Likewise when the oxide spalls and may not be recovered in post-exposure examination. Oxide-film breakdown is another important failure mechanism that is not fully represented by the parabolic rate constant and in such cases the combination of thermal, mechanical and corrosion effects will determine the ultimate materials limitations in service.

Chatterjee et al.[22] introduced a kinetic model for the high-temperature oxidation of nickel-base alloys based on a modified Wagner's oxidation model. Namely, the parabolic rate constant was related to the partial pressure gradient of oxygen across the oxide film, via the diffusivities of cations and anions according to the equation:

$$k_p = \int_{p_{O_2(I)}}^{p_{O_2(II)}} \left[\frac{aD^*(M)}{f_M} + \frac{D^*(O)}{f_O}\right] d\ln p_{O_2} \tag{2}$$

here $p_{O_2(I)}$ and $p_{O_2(II)}$ are the partial pressures of oxygen at the metal/oxide and oxide/gas interfaces, a is a stoichiometry coefficient, $D^*(M)$ and $D^*(O)$ are the tracer diffusivities of cations and anions in the oxide and f_M and f_O are correlation factors for self-diffusion. The authors then used a genetic algorithm to parameterize a model for oxide growth that was formulated under the assumption of a chromia/alumina solid solution phase. In particular, the parabolic rate constants were fitted against the composition of the alloys, which were inputs to a thermodynamic model based on CALPHAD that provided the partial pressure of oxygen at the metal/oxide interface. The model was significantly focused on nickel-base alloys, primarily alloyed with Cr and/or Al, and the physics of the model was limited to the diffusivity of O, Cr and Al defects in the oxide film. Hence, the role of other alloying elements was not directly taken into account.

Sato and co-workers in 2011 outlined a predictive model for oxidation kinetics[23], beginning with an expression for the rate constant that is proportional to the free energy of oxide formation:

$$k_p = -\frac{\sigma_t t_{ion} t_{el}}{z_c^2 z_a^2 e} \Delta G_f \tag{3}$$

where σ_t is the conductivity of the oxide, t_{ion} and t_{el} are transport numbers for ion and electrons, respectively, z_c and z_a are the valencies of the cation and anion, and e is the electronic charge. Consequently, the authors then related the various transport and valence terms to the free anion vacancy concentration in the oxide, resulting in an expression for the rate constant that should be proportional to the free energy and the oxygen anion vacancy concentration:

$$k_p \propto V_{\ddot{o}} \Delta G_f \tag{4}$$

Given that vacancy concentrations are not readily determined, the authors estimated the defect concentration of vacancies from the effective valence of the oxide phase, Val_t^{eff}, as determined by the atomic fraction of elements having a valence different from the parent oxide phase (in their study it was Al_2O_3). Thus the vacancy concentration was assumed to be proportional to Val_t^{eff},

computed by:

$$Val_t^{eff} \propto \sum_{n=i} (z_i - z_{Al}) c_i^\gamma \tag{5}$$

where z_i is the valence of cations i that may be co-present with Al as defects in the oxide phase (assumed to be Ni and Cr), and c_i^γ is the composition of that element i in the alloy matrix.

Bensch developed an oxidation-creep model for the nickel-based superalloy material René N5[24]. The diffusion kinetic code DICTRA was used to simulate the oxidation front of the alloy[25]. Within this sophisticated multiphysics model, the oxidation was proposed to modify the near-surface concentration of the oxidized elements, which were also the precipitation hardening forming elements (Ti, Al). Based on this outgoing flux of the oxidized element(s), two microstructurally distinct zones emerged in the near-surface region: a precipitate free zone, and a precipitate depleted zone. Accordingly, there was predicted to be a decrease in mechanical properties in those two zones. Ultimately, a model was developed as shown linking these components (oxidation, diffusion, precipitate depletion, mechanical property degradation) together to predict the creep response of the material due to oxidation.

Bensch further developed the model to take into account variations in the oxidation rate constant due to changes in the alloy composition[26]. In this model the rate constant was expressed with an equation of the form:

$$k_p = c_2 \exp(c_3 U) \tag{6}$$

where U is essentially an activation energy, related to the energy to form a unit of Al_2O_3, and is influenced by the alloy composition by the expression (which includes Eq. 5):

$$U = \sum_{i=1}^{N} [(z_i - z_{Al}) \cdot x_i^\gamma] \cdot (\Delta G_f + c_1) \tag{7}$$

The underlying physical principle behind adopting this expression for U is that elements with a higher valence state (z_i) than Al should increase the trapping state for defects responsible for oxide growth. c_1, c_2, and c_3 are fitting constants, as developed by Sato and co-workers and described above[23]. ΔG_f is the free energy of aluminum oxide formation, and x_i^γ is the composition of element i in the γ-phase. The model was applied to alloys SCA425 + and Rene N5, and the formula for U only considered Ni and Cr as dopants in the oxide scale. As in the earlier work by Bensch et al.[24], the goal of modeling the oxidation was in part to understand the near-surface depletion of the strengthening precipitates induced by the oxidation. We will revisit this expression in the work that follows, as we extend the model to a very broad class of high-temperature alloys and provide some further theoretical basis for the identity of the variables U and the parameters c_1, c_2, and c_3.

Carter, Gleeson, and Young in 1996 solved for the near-surface depletion profile of the more reactive primary scale-forming element by assuming a flux condition associated with the parabolic oxide film growth and provided an analytical solution[27]. The method was applied to Ni–Si alloy and used to infer the dissolution of the Ni_3Si precipitates in the near-surface region.

Perez-Gonzalez et al.[28] studying the oxidation of Haynes alloy 282 and reported parabolic rate constants. Microscopic analysis indicated that the oxide scale consisted of an outer TiO_2 and an inner Cr_2O_3 scale, with an internal oxidation zone of mixed titanium and aluminum oxides. Despite the heterogeneity of the oxide composition and structure, the parabolic rate law still applied and should be considered a composite function for the mass change as a function of time across the various oxides forms. Similar findings were reported by Liang et al.[29] under flowing air and steam exposures, with an acceleration in the corrosion rate observed to occur under flow and in the presence of water vapor.

Pint et al.[5] studied the oxidation properties of a variety of alloys under ambient and supercritical CO_2 conditions at 700 and 750 C, including alloy 25, 740H, 625, and 282. The mass change with oxidation for very long times of exposure was highest for alloy 25 (Fe-based alloy), and decreased in effect from 282 to 740H to 625, which was the most resistant to oxidation. Over shorter times, the order of mass increase was 282 > 25 > 625 > 740. The oxides formed in CO_2 at 1 bar or 300 bar were similar to those formed in air.

Oleksak et al.[30] studied the oxidation behavior of several Fe and Ni commercial alloys in high-temperature environments containing CO_2, H_2O, O_2, and SO_2. A variety of alloys were studied including 617, 230, 740H, 347H, Grade 22 and 91, 310S, 625 and 282; i.e., spanning low-Cr ferritic, high-Cr austenitic steels, as well as Ni alloys. The materials were exposed to temperatures of 700 and 750 C for up to 2000 h. Ni alloys were shown to be superior, followed by the austenitic (high Cr) steels, and inferior were the low Cr ferritic steels.

Chyrkin (2015 and 2018) studied the oxidation characteristics of alloy 602CA, a chromium-rich, carbide strengthened wrought Ni-base alloy used in many high-temperature applications[31,32]. The alloy was exposed to temperatures from 800 C to 1200 C. Oxidation rate constants were reported as a function of temperature. Oxidation resistance at the high temperatures was attributed to the formation of alumina scale, which formed in conjunction with a depletion of the Al in the near-surface region of the microstructure. Beyond 1100 C the pre-formed alumina oxide film breaks down, resulting in only internal Al oxidation. The influence of oxidation on carbide stability was also investigated, with Cr depletion driving the dissolution of carbides in the near-surface region. Both the depletion in Al and Cr would be expected to limit alloy lifetimes due to its effect on alloy strengthening mechanisms.

Hou (2003) collected oxidation kinetics data on a large number of aluminides (iron, chromium, nickel, and platinum aluminides in various combinations and some incorporating reactive elements)[2]. The data was collected and visualized using the Arrhenius plot format (1000/T vs k_p in g^2/cm^4s). Hou observed that the effect of alloying and shifting the majority base element (Fe, Cr or Ni) on transport rates seemed small. At the same time, oxidation rates of the Fe- and Ni-based alloys were reported to be about one order of magnitude higher than the Pt-based ones.

Due to the significance of chromium in high-temperature alloy development, and the important role of chromia films in providing a barrier to corrosion, Hallstrom et al. studied the kinetics and microstructure of oxidation of chromium metal. Oxidation results were reported at 600, 625, and 700 C[33]. They compared the extent to which grain boundary as well as in-grain transport contributed to oxide growth mechanisms. The contribution of grain boundary diffusion indicated that the size of the oxide grains and thermally activated boundary diffusion rates would be significant in overall oxidation performance of chromium.

Christofidou et al. studied the oxidation performance of nine Ni–Co–Al–Ti–Cr alloys at 800 C[34]. The alloys were primarily Ni–Co alloys, with Cr varying from around 10–20% and around 2–8% of Al and Ti blended in to form a series of model alloys. A fit was made to determine the parabolic rate constant to the mass change upon oxidation for 1000 h. Microscopy (secondary electron imaging) was used to look at the elemental distribution in the scales. Ni–17.5Co–7.5Al–2.5Ti–20Cr was found to be the most resistant to oxidation. For a given Cr concentration, increasing Ni and Al had the most effect on reducing oxidation. The formation of γ' phase was noticed to enhance the chromium content of the matrix, increasing corrosion resistance.

Cruchley compared the oxidation behavior of RR1000 to pure chromium, Inconel 718, Astroloy, Waspaloy, Inconel 720, Haynes 75, Alloy 230, and Rene 95[35]. An effective activation energy of 286 kJ/mol was reported for the RR1000 alloy, comparable to the

other chromia forming alloys reviewed in their paper (250–300 kJ/mol). Titania crystallites were observed on the outermost part of the scale. A γ' denuded zone was observed near the surface, as in similar work on complex precipitation-hardened alloys exposed to an oxidizing environment.

Schiek et al.[36] studied the oxidation kinetics of 602CA exposed to various gas environments characterized by their Ar-O_2-H_2O ratios at 800 °C. Alumina and chromia scales were observed to different extents and alumina could be internal or external depending on the temperature. Chromia growth rate was believed to be influenced by the presence of titanium, which could surface from the film as metallic nodules.

Collectively, the above-reviewed papers provide a significant data set of high-temperature oxidation properties of alloys spanning a diverse range of materials: nickel-based superalloys, aluminides, low-Cr ferritic steels, and higher-Cr austenitic steels. To facilitate the present and future development of CAMD of alloys with high-temperature oxidation resistance, we have collected this data and evaluated several models for predicting the parabolic rate constants published in the literature (or in some cases, derived from the provided oxidation measurements) as a function of composition. As will be described in the methods section, the basic form of the models constructed was drawn in part by the works of Chatterjee, Sato, and Bensch[22,23,26], and provides further insight into the role of alloying elements in influencing the susceptibility of materials to high-temperature corrosion.

RESULTS AND DISCUSSION

Visualization of materials data

Parabolic rate constants, k_p, were collected either directly from published reports, or calculated from corrosion product data (sample mass gain or corrosion product thickness) and tabulated for 75 alloys exposed to temperatures between ~800 and 2000 K (~500–1700 °C; 900–3000 °F). Data were collected for environments including lab air, ambient and supercritical carbon dioxide, supercritical water, and steam. When the environment was not directly reported, it was assumed that laboratory air was used.

As can be seen in the data spread, most of the data-points collected corresponded to (or were assumed to be, when information was unavailable) laboratory air. As further data are collected in our ongoing efforts in this area, increasing the diversity of environments and more granularity regarding, for example, partial pressure of different gas components, will be an objective. However, the process of data curation remains largely manual and, therefore, time-consuming, and, so for the purposes of this manuscript, the focus will be on exploring the roles of alloy composition and temperature on oxidation kinetics, and the influence of environment will be considered to be within the spread of data-points for any given material. Typically, the role of the environment can change the oxidation rate constant by a factor of one order of magnitude, in the order of steam/water >lab air >CO_2.

Arrhenius model selection using linear regression

The Arrhenius plots used to infer reaction kinetics are generally characterized by a slope and an intercept. Visual examination of Fig. 1 suggests that the data to the left of the plot have a steeper slope (corresponding generally to aluminide alloys), whereas the central set of data-points have a shallower slope (nickel-based superalloys), and low-Cr alloys potentially an even shallower slope, although the data-points are sparse in that region (top right). Generically, the form of the rate constant as a function of temperature should take the Arrhenius form:

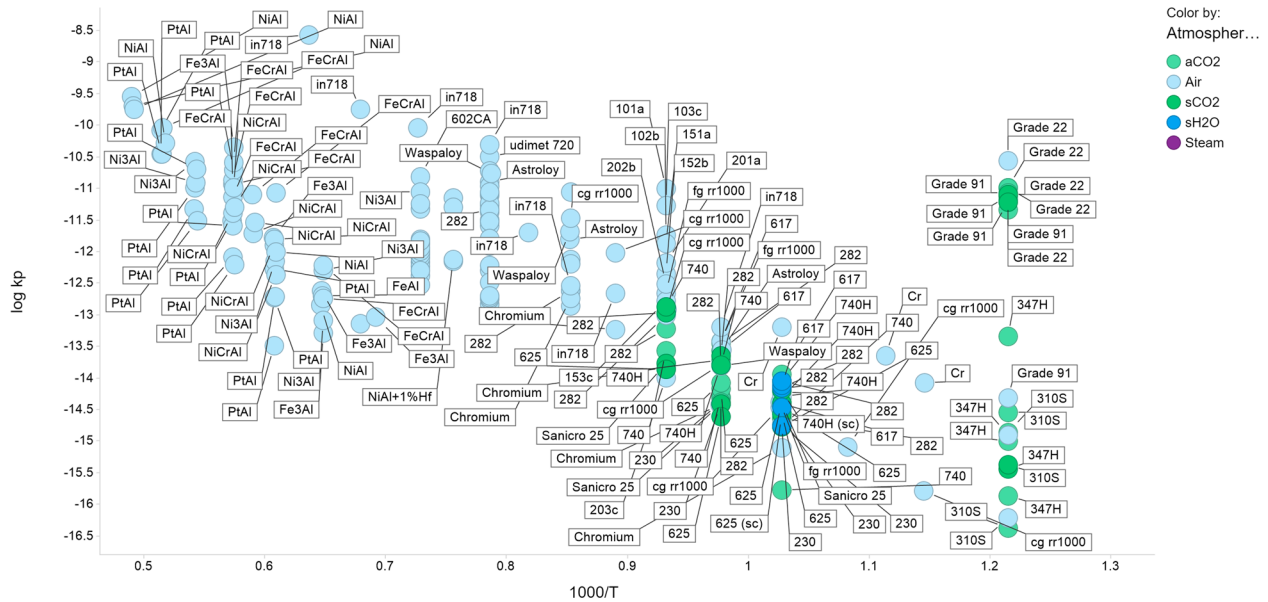$$\text{Log}_{10}k_p = \log_{10}A + m(1000/T) \tag{8}$$

Fig. 1 Scatterplot of parabolic rate constants vs. temperature. Arrhenius plot ($x = 1000/T$, with T in Kelvin; and $y = \log_{10}k_p$, with k_p in units of $g^2/cm^4/s$) for the collection of alloys curated from the high-temperature oxidation literature. For alloys where no environment is explicitly reported in the paper, it is assumed the environment is air. aCO_2 is ambient pressure CO_2, and sCO_2 is supercritical CO_2. sH_2O is supercritical water.

A simple and subjective visual examination of the data suggests that there may be a common pre-factor $\log_{10}A$ which lies somewhere around zero. Since, within Arrhenius theory, the pre-factor relates generically to an attempt frequency, it seems reasonable that this value for the y-intercept may be common, or nearly common among all materials classes, whereas the activation energy (expressed by -m) will vary depending on the rate-limiting step for oxide growth and will be a function of materials composition and microstructure.

In the work of Sato et al. and Bensch et al. the activation energy for oxide growth was associated with the content of "dopants" in the alumina scale, and thus fitted to a function of the alloy composition variables, with a focus only on the role of Ni, Cr, and Al. Chatterjee et al.'s model likewise focused on transport[23,26] in the oxide and focused on the influence of Cr and Al content on the diffusivity[22]. For this reason, we attempt similar linear regression methods for the composition of the alloys against the value of m, for different selections of $c = \log_{10}A$ varied between 0 and +20. Effectively deciding on a value of c maps each data-point in Fig. 1 to a unique, temperature-independent value of m, given by the expression:

$$m = (\log_{10}k_p - c)/(1000/T) \tag{9}$$

This method has the advantage of providing multiple estimators of m for the same material independently of temperature but still remaining sensitive to changes in the environment and variations in the alloy composition, as well as any other variations in methodology, measurement or numerical accuracy.

Using the TIBCO Spotfire package we follow a similar procedure to Sato et al. and Bensch et al. to fit the activation energy term m to a linear function of the Ni, Cr and Al composition of the alloys[23,26]. We perform fits for various values of the pre-factor term, varying c from 0 to +20, since it was noticed that the R^2 coefficient appeared to improve as the value for c was varied in the positive direction. By eye it appeared that a common intercept for the various linear clusters perceived in the plot would be somewhere around 0. However, the linear regression fitting appeared to favor higher (more positive) values of c, reaching a local maximum at $c = 8$. The results of the regression fitting are shown in Supplementary Information Fig. 1. In a following section

we have provided a physics-based derivation of the pre-factor c, and then show different choices of c influence the machine learning and other outcomes of this analysis.

The linear regression methods were varied to include other potential alloying elements of significance: Si, Fe, Mo, yet the regression was not significantly improved. The most significant alloy elements included in this data analysis were ranked according to their t-statistic as Al > Ni > Cr (Supplementary Information Table 1). In Supplementary Information Fig. 2, we plot the rate constants and color them according to the values of m derived for the low intercept case ($c = 0$), and (b) the high intercept case, which is the optimal value from Supplementary Information Fig. 1 ($c = 8$). In Supplementary Information Fig. 3 we present parity plots on the predictor variable m for the various selections of c.

To assess the validity of this method for selecting different choices for the intercept c before proceeding to further analysis, we consider the physical interpretation and compare to the methods of Bensch et al.[26]. and Chatterjee et al.[22].

Modifying the equation provided by Bensch et al.[26], the parabolic rate constant can be expressed as

$$\log k_p = \log c_2 + (\log e)c_3 U \tag{10}$$

here c_2 and c_3 are fitting parameters. The intercept in the present model should be approximately equivalent to Bensch's value for $\log_{10} c_2$. The c_2 values are given as 10^{-2}, 10^{-5}, and 10^{-45} for three attempts to fit[26]. Thus, the Bensch model would indicate lower c values around -2 to -5 compared to those obtained by the fitting procedure in Supplementary Information Fig. 1 (we assume that 10^{-45} is most likely to be an extreme outlier). The Arrhenius slope in the Bensch model should relate to the parameter $c_3 U$ (Bensch's data is at 1000 °C and 900 °C, and since there is no temperature dependence in their equation, $c_3 U$ incorporates both a temperature effect, as well as an adjustment between ΔG_f and the kinetic parameter). According to Bensch's work, at 1000 °C, U is around 50 kJ/mol for the Ni superalloys, and c_3 is on the order of -0.5. Making the equivalency from the Bensch model to the

functional form of the Arrhenius equation:

$$\log_{10} e \Delta E_a / RT = \log_{10} e\, c_3 U \tag{11}$$

$$c_3 U \approx -0.5 * 50 = -E_a / 8.314 / 1273 \tag{12}$$

$$E_a \approx 260\, kJ/mol \tag{13}$$

This value for the activation energy is consistent with activation energies for chromia forming alloys in the literature. Chen et al. determined activation energies of 270, 300, and 250 kJ mol ~ for Astroloy, Waspaloy, and Udimet 720, compared to 255 for Cr in chromia[37]. Tawancy et al. report activation energies for alumina formers of about 396 kJ/mol, comparable to 382 kJ/mol for growth of α-Al$_2$O$_3$ on a number of Ni-Al-Cr alloys[38] and 388 kJ/mol for alloy MA 956[38]. Chyrkin et al. determine an activation energy of ~360 kJ/mole in the oxide scale on 602CA compared to that for Al diffusion in the alloy (264 kJ/mol)[32]. For NiO, the diffusivity of cations is reviewed by Stott and a value of 257 kJ/mol seen to be consistent with a wide range of data, although others prefer the lower value around 170 kJ/mol[39]. Cruchley et al.[35] survey several alloys and determine activation energies of 286 kJ/mol for RR1000, a little higher than for chromia ~250 kJ/mol[35]. Taylor's study of RR1000 found an activation energy significantly higher for RR1000, 326,740 kJ/mol indicating that other elements than Cr are being oxidized, likely Al given that the alloy has 3 wt.% aluminum[40]. Young's textbook tabulates activation energies for Fe, Ni, Cr and Co as 164 (700–1000 °C), 120 (600–1100 °C) & 240 (1100–1400 °C), 243 (980–1200 °C), and 230 (800–950 °C) & 150 (950–1150 °C) kJ/mol, respectively[1]. For alumina scales, the activation energy can vary between 300–400 kJ/mol depending on the extent of grain-boundary diffusion and the role of other alloying elements[1].

The primary significance of Ni, Cr, and Al, over the presence of other alloying elements in determining the best-fit model for the parabolic rate constant is also consistent with the prior models developed by Chatterjee, Sato, and Bensch[22–24,26].

Since the intercept term c is related to the Arrhenius pre-factor, another argument for the rough order of magnitude for the prefactor associated with k_p can be made using some other logical arguments. The attempt frequency is related to the natural vibration frequency of atoms and is given by the value $k_B T/h$ which is around 2×10^{13} /s at 1000 K. To convert to units of g^2/cm^4 we need to multiply by the square of the mass/area density of the atoms responsible for oxide growth. The molar mass of an oxygen atom is 16 g/mol. The surface density corresponds to ~1 site per unit cell which is on the order of ~0.5 nm, so the pre-factor for oxidation in the units of g^2/cm^4/s should be given by the expression:

$$A = \frac{k_B T}{h} \left(\frac{16g/mol}{6 \times 10^{23}\, atoms/mol} \frac{1\, atom}{(0.5 \times 10^{-7} cm)^2} \right)^2 \tag{14}$$

For $T \sim 1000$ °C this would yield a value of c of -2.6, which is generally consistent with the values determined for c_2 by Bensch (10^{-2} to 10^{-5})[23,26], and the apparent intercept which seems closer to 0 or slightly <2 as determined "by eye". In the following section, we will show that values of c close to -3 provide lower mean absolute errors for a variety of machine learning approaches applied to the data set, although values close to zero provide for activation energies more in line with those observed for the various alloy classes studied in this work.

Supervised learning techniques

To improve the predictivity of the models and further explore relations between alloying elements and oxidation behavior, we next considered the application of alternative regression methods based on supervised machine learning. Whereas the linear regression models assumed in part B and prior work may be

Table 1. Mean-squared error (MSE), root mean-squared error (RMSE), mean absolute error (MAE) and R^2 coefficient for the various machine learning approaches used to fit the Arrhenius slope for the parabolic rate constant as a function of materials composition.

	MSE	RMSE	MAE	R^2
Low c (c = 0)				
Linear (Ni, Cr, Al)[c]	1.69	1.30	0.90	0.79
Linear[c]	0.99	1.00	0.75	0.88
NN	0.81	0.90	0.64	0.80
SVM	1.03	1.02	0.78	0.75
KNN	0.78	0.90	0.61	0.81
Tree	0.82	0.90	0.64	0.80
Random forest	0.72	0.85	0.61	0.83
High c (c = 8)				
Linear (Ni, Cr, Al)	4.66	2.16	1.62	0.83
Linear	2.73	1.65	1.24	0.90
NN	2.41	1.55	1.00	0.91
SVM	3.90	1.97	1.51	0.85
KNN	2.12	1.46	0.95	0.92
Tree	2.29	1.51	0.99	0.91
Random forest	2.09	1.45	0.95	0.92
Physical (c = −3)				
Linear (Ni, Al, Cr)[a]	1.29	1.13	0.83	0.70
Linear[b]	0.78	0.88	0.66	0.81
NN	0.81	0.90	0.64	0.80
SVM	1.03	1.02	0.78	0.75
KNN	0.78	0.88	0.61	0.81
Tree	0.82	0.90	0.64	0.80
Random forest	0.72	0.85	0.61	0.83

[a]Initial linear fits were done on Ni, Cr, and Al compositions.
[b]Linear fits performed in orange/python using all composition data.
[c]Tested on train data only.

expected to approximately apply within the constraints for a limited class or similar classes of alloy (i.e., with small variations in alloy compositions and a common base metal), across different alloy classes (aluminides, Ni-based CRAs, chromia formers and ferritic vs austenitic steels) the relationship between alloy composition and oxidation rate constant would be expected to be non-linear. We therefore consider approaches such as artificial neural networks[41] and support vector machines (SVM)[42], since these apply to non-linear systems. We also consider classification based methods, such as k-nearest neighbor[43], which should favor finding the rate constants based on similarities in composition, as well as tree-based methods such as tree and random forest[44,45], based on the intuitive concept that alloys in similar 'family trees' may be expected to have similar oxidation rate constants. The models were tested and scored using five-fold cross-validation.

In Table 1 we plot the mean standard error, the root mean standard error, the mean absolute error, and the R^2 value for various machine learning approaches with three different assumptions for the intercept term c: the low c case ($c = 0$), the 'optimized' value or high c case ($c = +8$) and the physically estimated value case, $c = -3$ (which is also in the range of values parameterized for the analogous constant by Bensch[26]). As can be quickly appraised from the table, the non-linear methods of Random Forest, KNN and Tree give the lowest errors, and for the negative and low c cases, with the errors being significantly higher for the 'optimized' c case, despite this value leading to higher R^2 scores.
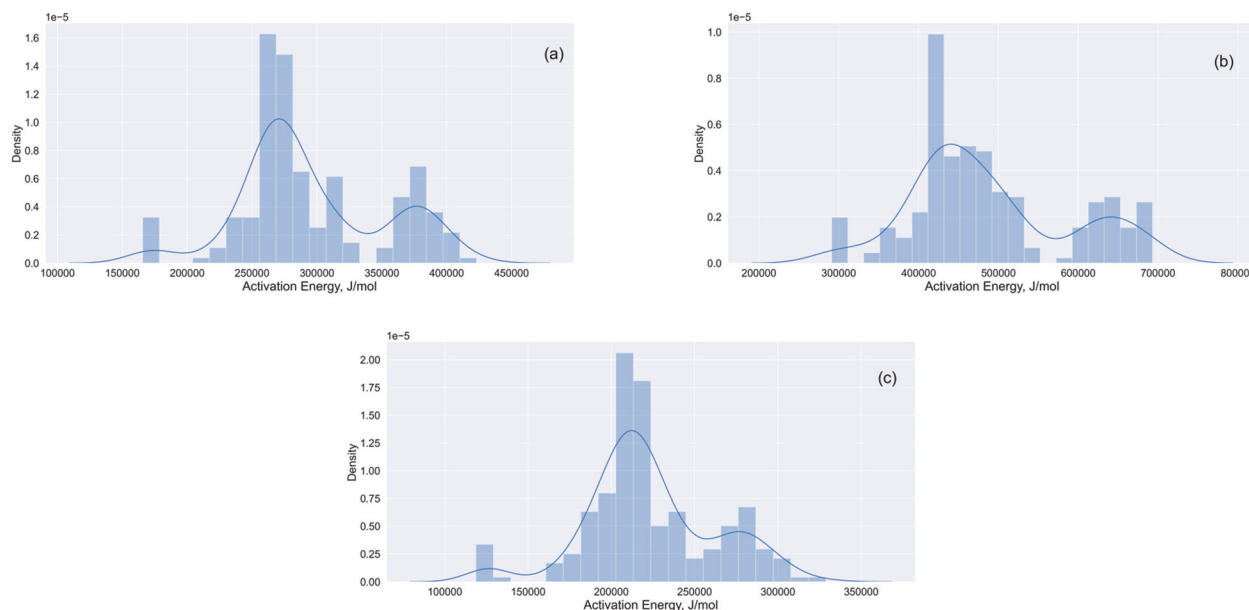
Fig. 2 Activation energies calculated with three different Arrhenius pre-factors. Activation energy distributions from the data set with different assumptions for the constant c which corresponds to the Arrhenius intercept. The distribution of activation energies is shown for the data when using the assumption of intercept (**a**) $c = 0$, (**b**) $c = +8$, (**c**) $c = -3$.

This more rigorous analysis reveals that the high c value of $c = 8$ leads to significantly larger errors in the linear and non-linear models. Hence the lower values of $c = 0$ and $c = -3$ lead to more accurate predictions for the parabolic rate constants, within the Arrhenius model.

To further investigate the veracity of the various model predictions, we plot the activation energies that can be readily derived from the slope m of the Arrhenius plot as follows:

$$E_a = -1000(8.314)m/\log_{10} e \, J/mol \qquad (15)$$

The distribution of activation energies from the data set collected for this work are provided for the low c ($c = 0$), high c ($c = 8$) and physical ($c = -3$) assumptions in Fig. 2. The activation energies for the $c = 0$ case most accurately align with the activation energies in the literature for iron oxide (~150–200 kJ/mol), chromia (250–300 kJ/mol) and alumina forming (300–400 kJ/mol) alloys[1]. The $c = 8$ assumption leads to activation energies that are ~2x higher and the $c = -3$ assumption leads to activation energies that are roughly 25% lower.

Given this finding, we focus on the $c = 0$ assumption for the remainder of the discussion and analysis in this paper, as it leads to the most realistic distribution of activation energies for the oxide scales when compared to the literature data.

The relative significance of the elements in the alloys to the oxidation kinetics was evaluated based on the rank correlation between element composition % and the value for the Arrhenius slope, m, obtained in the low c ($c = 0$) assumption. Both linear and random forest models (for which rankings were available) indicated that Ni, Cr and Al were in the top four most significant compositions.

The linear model lists Fe composition as the most significant variable, whereas the random forest lists Al as the most significant, and the fourth most significant as Mo, and then Fe as fifth. The other elements in the top 10 most significant are also given in Table 2. Elements in common to both methods (in addition to Fe, Ni, Cr and Al) include Co and Cu.

The decision tree regression model for the Arrhenius slope in the $c = 0$ case is shown in Fig. 3. As can be seen, the most significant split in the decision tree is Al, which differentiates at

Table 2. The ten most significant elements according to rank for linear and random forest.

Element	Linear Rank	Element	Random forest
Fe	1.547299	Al	0.396931
Ni	1.466135	Cr	0.163614
Cr	1.170996	Ni	0.086381
Al	1.059325	Mo	0.073809
Co	0.888269	Fe	0.064088
S	0.854328	C	0.043457
Pt	0.83646	Co	0.028014
N	0.667273	Mn	0.02771
Cu	0.653015	Cu	0.022063
Hf	0.410626	Pt	0.021459

the outset aluminides, having very steep Arrhenius slopes (high activation energies) from the nickel and iron-based superalloys. Subsequent splits differentiate between higher iron contents, low and high chromium steels, cobalt content, and then again according to the extent of aluminum versus nickel in the alloy. As can be seen visibly by the color differentiation, the right-most branch signals the aluminides, the center branch is the low-chrome ferritic steels, and the left-most branch divides according to stainless steels and nickel super alloys.

The random forest model creates a "forest" of such decision trees and collectively uses the trees to make a composite prediction for the target variable, i.e., the Arrhenius slope in this case. A sample of the decision trees, presented in the Pythagorean tree representation is shown in Supplementary Information Fig. 4. The weighted signifcance of the different variables taken among the various "family trees" so constructed is given in Table 2. Figure 4 shows the parity plot for the random forest model, signifciantly less scatter than the linear models in Supplementary Information Fig. 3.
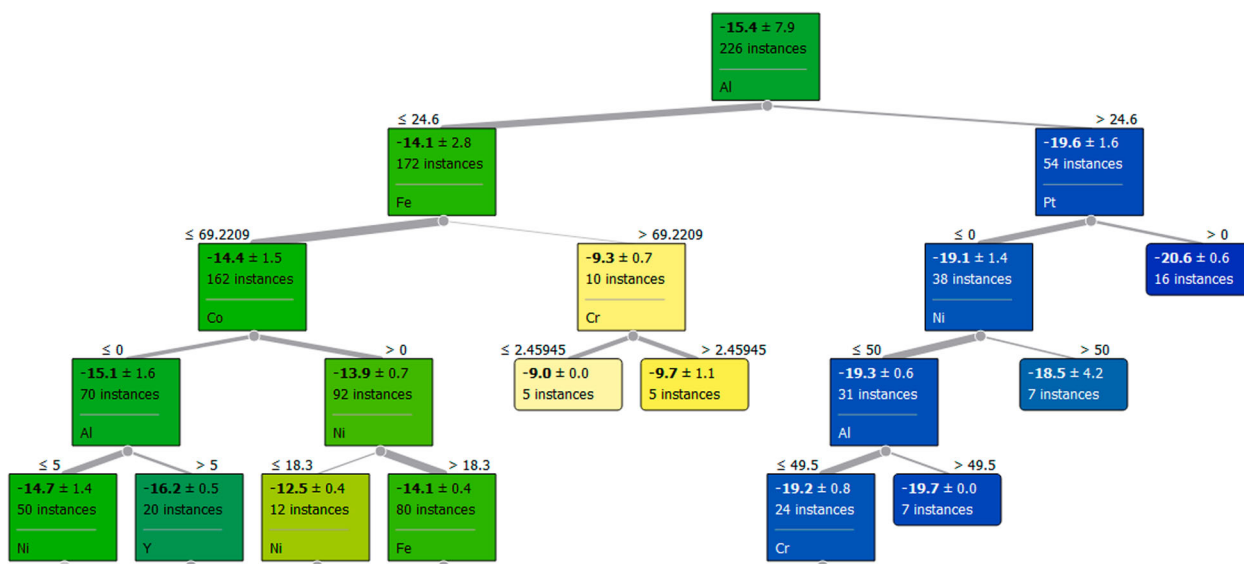
Fig. 3 Decision tree for activation energy proxy based on elemental composition. Decision tree for the determination of the Arrhenius slope based on the elemental composition.
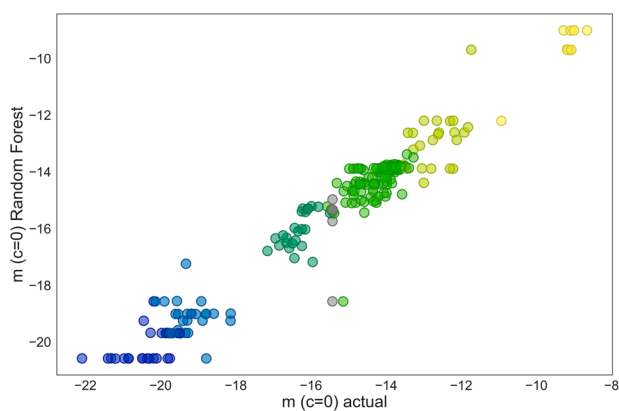

Fig. 4 Parity plot of random forest versus actual activation energy proxies. Parity plot of random forest (fitted) versus actual slope values using the intercept assumption that $c = 0$.

Fig. 5 Arrhenius slopes distributed according to k means clustering. Distribution of Arrhenius slopes according to k means clustering of the t-SNE two-dimensional projection of alloy composition space.

Unsupervised learning

Stochastic neighbor embedding according to the t-statistic (t-SNE) was used to explore the composition data set in an unsupervised fashion[46]. This technique allows for non-linear dimensionality reduction, which seeks to capture both local and global complexity of high dimensional data sets. It is essentially a non-linear extension of principal component analysis. The two-dimensional reduction of the composition data space that can be achieved using t-SNE is shown in Supplementary Information Fig. 5. There are two primary distributions of data as seen in the lower left and upper right quadrants of the two-dimensional plot. K-means analysis was used to further divide the data into clusters[47]. Six clusters were used for the analysis, as, from experience there should be at least 4 clusters of alloy classes present: ferritic and stainless steels, nickel superalloys, and aluminides. The t-SNE distribution clusters C6 and C5 closer together, C1 closer but separate, and C2–C4 separately. The Arrhenius slopes are grouped according to the clusters in Fig. 5, where it can be seen that C1 has the slope characteristic of the iron-based alloys (ferritic and stainless steels), C6 and C5 are the less corrosion-resistant nickel superalloys, C2 are the aluminides, and C3 and C4 span the aluminides and more resistant nickel superalloys. In Fig. 6 we
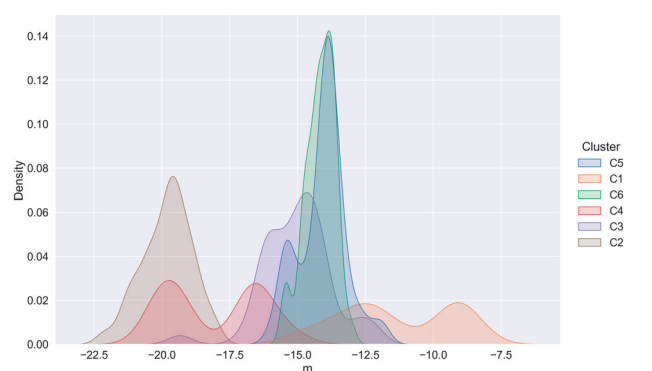
further break down these clusters according to their composition distributions, where these relations become further apparent.

- C1: high Fe content (ferritic and austenitic steels) alloys
- C2: the aluminides
- C3: higher Ni composition, chromia forming nickel superalloys without Al
- C4: higher Cr containing nickel superalloys with appreciable amounts of Al
- C5: lower Cr nickel superalloys with small amounts of Al
- C6: intermediate Cr nickel superalloys including some with Fe

The expected oxidation resistance based on compositions would be $C2 > C4 > C3 \sim C5 \sim C6 > C1$. The cluster analysis therefore provides some further classification of the nickel superalloys according to their extent of alloying with aluminum and chromium. As the database is further populated it is expected that such analyses can reveal insights as to the role of other alloy elements and may be used to guide composition optimization.

In this work we have collected data on parabolic rate constants for alloys spanning the classes of low- and high-Cr ferritic and austenitic steels, nickel superalloys, and aluminide materials and used a combination of Arrhenius analysis, simple linear regression, supervised and unsupervised machine learning methods to
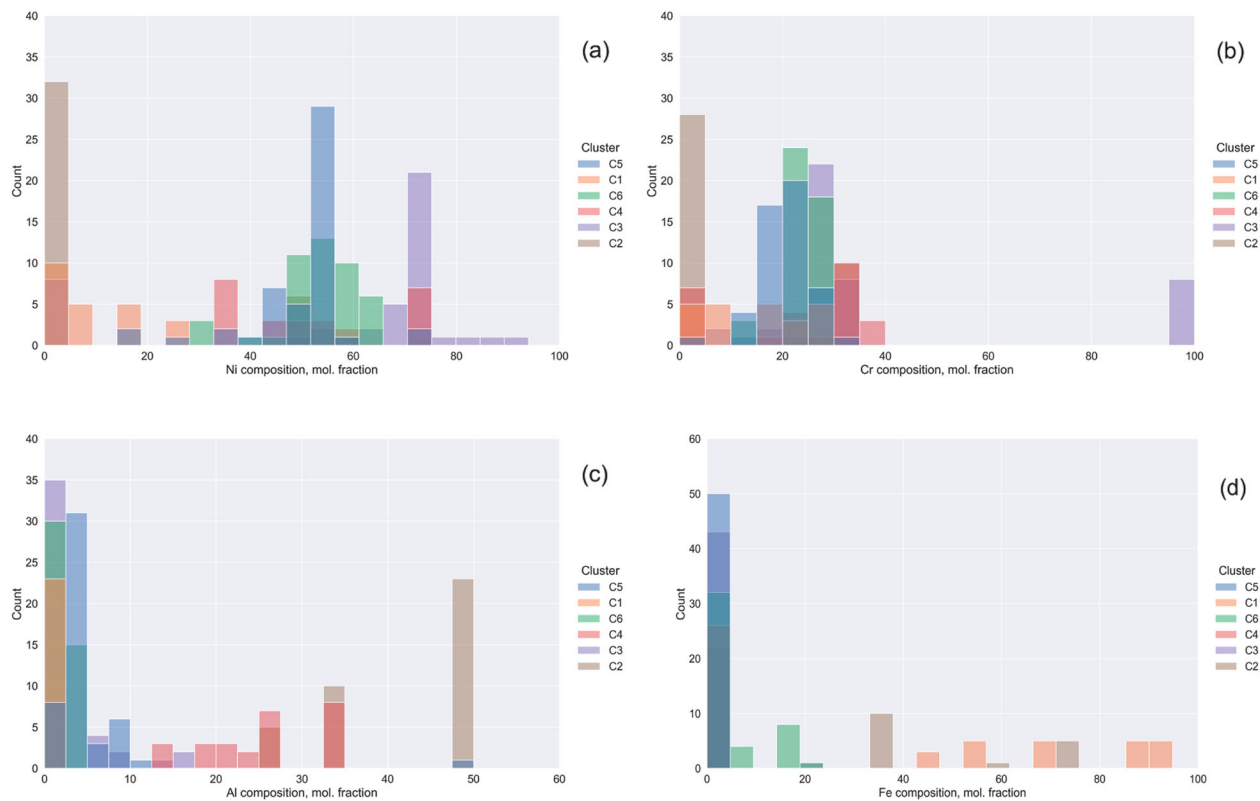

Fig. 6 Alloy distributions according to Ni, Cr, Al, and Fe content. Alloy distributions according to Ni, Cr, Al, and Fe content for the six clusters identified in the k-means analysis.

investigate the relations between composition and oxidation kinetics. The supervised machine learning techniques produced the lowest mean standard errors on the datasets when five-fold cross-validation was used. The most significant elements controlling oxidation kinetics were Ni, Cr, Al, and Fe, with Mo and Co also playing significant roles. The activation energies produced from the machine learning analysis were in the correct distributions for the diffusion constants for the oxide scales expected to dominate in each class. The database developed and the initial machine learning analyses described in this work should provide a basis for further investigations of the role of minority elements and variations in the environment for affecting the oxidation kinetics. The unsupervised learning provides an alternative way to group the composition data in the dataset through dimensionality, and it provides further granularity to the composition classes by further subdividing the nickel superalloys according to the relative richness of their aluminum, chromium or nickel content.

METHODS
Approach
High-temperature oxidation kinetics, in this work and most of those reviewed in this paper, are assumed to follow the Wagner kinetic model[20] where the rate of oxidation is characterized by the parabolic rate constant k_p, although it is known in many cases for highly corrosion-resistant alloys that growth may become sub-parabolic, and, in less resistant alloys and/or extreme service conditions, particularly cyclic temperature variations, involve successive events of oxide growth and spallation. The reason for utilizing parabolic oxide kinetics in the present work is that it provides the most expedient starting point for applying an informatics based approach: parabolic rate constants have been reported for a wide variety of alloys across numerous classes (austenitic, ferritic, intermetallic, and superalloys) and can be readily extracted by fitting $t^{1/2}$ curves to oxidation mass change data reported in the literature. Future improvements of this work, and indeed, the theory of high-temperature oxidation of alloys, can then add

successively more complex models. For example, one might consider in an alternate analysis using only the observable mass change or oxide film thickness, rather than the post-analytical rate constant, and other mechanisms relevant to materials could also apply, such as oxide film breakdown, spallation, and thermo-mechanical corrosion fatigue. However, these are beyond the scope of (a) this analysis, and (b) the types of data reported in the measurements collected for this work.

Data collection
Data on high-temperature oxidation was collected from a series of original reports and review papers[2,5,22,28–36,48]. 237 unique rate constants were obtained with a total of 75 different materials exposed to different environments and temperatures. The compositions for the materials were obtained from these papers and converted into atomic percent. The rate constants were converted into equivalent units of g²/cm⁴/s. Information was also collected, where available, on exposure environment, including temperature and gas composition.

Featurization
A machine learning approach was adopted to classify and model the materials based upon their parabolic oxidation rate constants using the techniques of data curation, data visualization, logistic regression[49], linear regression, random forest[45], single decision tree[44], artificial neural network[41], and k-nearest neighbor (KNN)[14,16,43]. To facilitate the classification process, an assumption was made to convert the parabolic rate constant into an equivalent activation energy for oxidation. This step was taken to remove the effect of temperature from the parabolic rate constants obtained for tests performed on the same alloy material but under different temperatures. Effectively, it was assumed that the parabolic rate constant was of the form:

$$k_p = A\exp(-\Delta E_a/RT) \tag{16}$$

in which case, the usual Arrhenius presentation would have the form:

$$\log_{10} k_p = c + m^*1000/T \tag{17}$$

i.e., taking the y-axis as the decadal logarithm of the parabolic rate constant and the x-axis as $1000/T$. Based on an initial visualization of the plot of $\log_{10}k_p$ versus $1000/T$ it seemed reasonable to assume that the intercept of the Arrhenius plot had a nearly common value for the pre-factor $c = \log_{10}A$. The value of the unknown parameter $\log_{10}A$ was varied to optimize the goodness of fit of the regression techniques attempted in this work. Whereas the environment may affect the pre-factor term, most of the experimental data-points collected appear to come from experiments performed in laboratory air and, furthermore, we assume that the environment is likely to have no more than approximately ±1 order of magnitude effect on the rate constants (and this assumption was borne out by casual inspection). For these reasons, and due to the limited nature of the data set in terms of diversity of environments studied, we do not include environmental effects on the rate constant in the present analysis, choosing to focus only on composition. As the data sets become richer it will become possible in the future to incorporate the environment as a 'feature' in the machine learning analysis.

Computational tools

The computational tools used to visualize and model the data include TIBCO Spotfire (https://www.tibco.com/products/tibco-spotfire), Orange v3.23 (https://orange.biolab.si/), and Python scikit-learn (https://scikit-learn.org/).

DATA AVAILABILITY

Data is available upon request to the corresponding author(s).

CODE AVAILABILITY

Codes are available upon request to the corresponding author(s).

REFERENCES

1. Young, D. J. *High Temperature Oxidation and Corrosion of Metals*; 2nd Edn. (Elsevier, 2016).
2. Hou, P. Y. Impurity effects on alumina scale growth. *J. Am. Ceram. Soc.* **86**, 660 (2003).
3. Pandey, M. C. & Satyanarayana, D. V. V. Effect of gamma prime depletion on creep behavior of a nickel base superalloy (inconel alloy X-750). *Bull. Mater. Sci.* **19**, 1009 (1996).
4. Satyanarayana, D. V. V. & Malakondaiah, G. Effect of prior oxidation on the creep behavior of nickel and iron based superalloys. *Trans. Indian Inst. Met.* **62**, 223 (2009).
5. Pint, B. A., Unocic, K. A., Brese, R. G. & Keiser, J. R. Characterization of chromia scales formed in supercritical carbon dioxide. *Mater. High. Temp.* **35**, 39 (2017).
6. Darolia, R. Development of strong, oxidation and corrosion resistant nickel-based superalloys: critical review of challenges, progress and prospects. *Int. Mater. Rev.* **64**, 355 (2018).
7. Lukas, H. L., Fries, S. G. & Sundman, B. *Computational Thermodynamics: the CALPHAD Method*. Vol. 131 (Cambridge University Press, 2007).
8. Spencer, P. J. A brief history of CALPHAD. *CALPHAD* **32**, 1 (**2008**).
9. Wen, Y.-H., Chen, L.-Q. & Hawk, J. A. Phase-field modeling of corrosion kinetics under dual oxidants. *Model. Sim. Mater. Sci. Eng.* **20**, 035013 (2012).
10. Uberuaga, B. P., Andersson, D. A. & Stanek, C. R. Defect behavior in oxides: insights from modern atomistic simulation methods. *Curr. Opin. Solid State Mater. Sci.* **17**, 249 (2013).
11. Ke, H. & Taylor, C. D. Density functional theory: an essential partner in the integrated computational materials engineering approach to corrosion. *Corrosion* **75**, 708 (2019).
12. Seyeux, A., Maurice, V. & Marcus, P. Oxide film growth kinetics on metals and alloys. *J. Electrochem. Soc.* **160**, C189 (2013).
13. de Borst, R. Challenges in computational materials science: multiple scales, multi-physics and evolving discontinuities. *Comp. Mater. Sci.* **43**, 1 (2008).
14. Wei, J. et al. Machine learning in materials science. *InfoMat* **1**, 338 (2019).
15. Schleder, G. R. et al. to machine learning: recent approaches to materials science–a review. *J. Phys.: Mater.* **2**, 032001 (2019).
16. Chen, A., Zhang, X. & Zhou, Z. Machine learning: accelerating materials development for energy storage and conversion. *InfoMat* **2**, 553 (2020).
17. Holdren, J. P. *Materials Genome Initiative for Global Competitiveness*. (National Academy of Science and Technology, 2011).
18. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
19. Taylor, C. D. Corrosion informatics: an integrated approach to modelling corrosion. *Corr. Eng. Sci. Tech.* **50**, 490 (2015).
20. Wagner, C. Beitrag zur Theorie des Anlaufvorgangs. *Z. Phys. Chem.* **21B**, 25 (1933).
21. Young, D. J. Predicting internal oxidation: building on the Wagner model. *Mater. Sci. Forum* **696**, 1 (2011).
22. Chatterjee, A. et al. Kinetic modeling of high temperature oxidation of Ni-base alloys. *Comp. Mater. Sci.* **50**, 811 (2011).
23. Sato, A., Chiu, Y. L. & Reed, R. C. Oxidation of nickel-based single-crystal superalloys for industrial gas turbine applications. *Acta Mater.* **59**, 225 (2011).
24. Bensch, M. et al. Modelling and analysis of the oxidation influence on creep behaviour of thin-walled structures of the single-crystal nickel-base superalloy René N5 at 980 °C. *Acta Mater.* **58**, 1607 (2010).
25. Larsson, H. & Hoglund, L. Multiphase diffusion simulations in 1D using the DICTRA homogenization model. *Calphad* **33**, 495 (2009).
26. Bensch, M. et al. Modelling of high temperature oxidation of alumina-forming single-crystal nickel-base superalloys. *Acta Mater.* **60**, 5468 (2012).
27. Carter, P., Gleeson, B. & Young, D. J. Calculation of precipitate dissolution zone kinetics in oxidising binary two-phase alloys. *Acta Mater.* **44**, 4033 (1996).
28. Pérez-González, F. A., Garza-Montes-de Oca, N. F. & Colás, R. High temperature oxidation of the Haynes 282© nickel-based superalloy. *Oxid. Met.* **82**, 145 (2014).
29. Liang, Z., Singh, P. M., Zhao, Q. & Wang, Y. High temperature oxidation of newly developed alloy 282 in the flowing-air and steam condition at 900–1100 °C. *Oxid. Met.* **84**, 291 (2015).
30. Oleksak, R. P., Tylczak, J. H., Carney, C. S., Holcomb, G. R. & Doğan, Ö. N. High-temperature oxidation of commercial alloys in supercritical CO_2 and related power cycle environments. *JOM* **70**, 1527 (2018).
31. Chyrkin, A. et al. Modeling carbide dissolution in alloy 602 CA during high temperature oxidation. *Corr. Sci.* **96**, 32 (2015).
32. Chyrkin, A. et al. Stability of external α-Al_2O_3 scales on alloy 602 CA at 1100–1200 °C. *Oxid. Met.* **90**, 119 (2017).
33. Hallström, S., Halvarsson, M., Höglund, L., Jonsson, T. & Ågren, J. High temperature oxidation of chromium: Kinetic modeling and microstructural investigation. *Solid State Ion.* **240**, 41 (2013).
34. Christofidou, K. A., Jones, N. G., Hardy, M. C. & Stone, H. J. The oxidation behaviour of alloys based on the Ni–Co–Al–Ti–Cr system. *Oxid. Met.* **85**, 443 (2015).
35. Cruchley, S., Evans, H. E., Taylor, M. P., Hardy, M. C. & Stekovic, S. Chromia layer growth on a Ni-based superalloy: sub-parabolic kinetics and the role of titanium. *Corr. Sci.* **75**, 58 (2013).
36. Schiek, M. et al. Scale formation of alloy 602 CA during isothermal oxidation at 800–1100 °C in different types of water vapor containing atmospheres. *Oxid. Met.* **84**, 661 (2015).
37. Chen, J. H., Rogers, P. M. & Little, J. A. Oxidation behavior of several chromia-forming commercial nickel-base superalloys. *Oxid. Met.* **47**, 381 (1997).
38. Tawancy, H. M. On the behaviour of minor active elements during oxidation of selected Ni-base high-temperature alloys. *Mater. High. Temp.* **34**, 22 (2016).
39. Stott, F. H. The protective action of oxide scales in gaseous environments at high temperature. *Rep. Prog. Phys.* **50**, 861 (1987).
40. Taylor, M. P., Evans, H. E., Stekovic, S. & Hardy, M. C. The oxidation characteristics of the nickel-based superalloy, RR1000, at temperatures of 700–900 °C. *Mater. High. Temp.* **29**, 145 (2014).
41. Jain, A. K., Mao, J. & Mohiuddin, K. Artificial neural networks: a tutorial. *Computer* **29**, 31 (1996).
42. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273 (1995).
43. Cover, T. M. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13**, 21 (1967).
44. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81 (1986).
45. Breiman, L. Random forests. *Mach. Learn.* **45**, 5 (2001).
46. van der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579 (2008).
47. Kanungo, T. et al. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 881 (2002).
48. Cruchley, S. et al. Cautionary note on use of focused ion beam sectioning as technique for characterising oxidation damage in Ni based superalloys. *Mater. High. Temp.* **31**, 27 (2014).
49. Hilbe, J. M. *Logistic Regression Models* (CRC Press, 2009).

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-021-00184-3.

Correspondence and requests for materials should be addressed to C.D.T. or B.M.T.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.