Exploring New Ways to Classify Industries for Energy Analysis and Modeling

Liz Wachs,[1] Colin McMillan,[1] Gale Boyd,[2] and Matt Doolin[2]

1 National Renewable Energy Laboratory
2 Duke University

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-6A20-82957
October 2022

Exploring New Ways to Classify Industries for Energy Analysis and Modeling

Liz Wachs,[1] Colin McMillan,[1] Gale Boyd,[2] and Matt Doolin[2]

1 National Renewable Energy Laboratory
2 Duke University

Suggested Citation
Wachs, Liz, Colin McMillan, Gale Boyd, and Matt Doolin. 2022. *Exploring New Ways to Classify Industries for Energy Analysis and Modeling*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-6A20-82957. https://www.nrel.gov/docs/fy23osti/82957.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-6A20-82957
October 2022

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Advanced Manufacturing Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Acknowledgments

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Advanced Manufacturing Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

The authors thank Anna Voss at the Association for Iron & Steel Technology for critical comments and insight into steel processes. We also thank Eric Stuart at the Steel Manufacturers Association for insight into electric arc furnace processes. We thank R. Neal Elliott at the American Council for an Energy-Efficient Economy for insights and resources on industry classification. We appreciate thoughtful reviews and comments from Alberta Carpenter, Elaine Hale, Jeff Logan, and Dan Bilello from NREL, Ookie Ma at the Department of Energy, and technical editing from Mike Meshek. Errors and omissions are the sole responsibility of the report authors.

DISCLAIMER:

Any views expressed here are those of the authors and not those of the U.S. Census Bureau. The U.S. Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this information product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 2173 (CBDRB-FY22-P2173-R9653/R9654).

List of Acronyms and Abbreviations

BF	blast furnace
BOF	basic oxygen furnace
BTU	British thermal units
CCUS	carbon capture, utilization and storage
CE	Common Era
CO_2	carbon dioxide
DRI	direct reduction of iron
EAF	electric arc furnace
EC	electricity use
ED	electricity demand
EIA	U.S. Energy Information Administration
EPA	U.S. Environmental Protection Agency
FReSMe	From Residual Steel Gases to Methanol
GHG	greenhouse gas
GHGRP	Greenhouse Gas Reporting Program
H_2DRI	hydrogen direct reduction
HBI	hot briquetted iron
HYBRIT	hydrogen breakthrough ironmaking technology
IAC	Industrial Assessment Center
KDE	kernel density estimate
MECS	Manufacturing Energy Consumption Survey
MMBtu	Million British thermal units
NAICS	North American Industry Classification System
NP	nondeterministic polynomial time
PAUP	Phylogenetic Analysis Using Paup
SIC	Standard Industrial Classification
SIDERWIN	development of new methodologieS for InDustrial CO_2-freE steel pRoduction by electroWINning

Executive Summary

Combustion, other emitting processes and fossil energy use outside the power sector have become urgent concerns given the United States' commitment to achieving net-zero greenhouse gas emissions by 2050. Industry is an important end user of energy and relies on fossil fuels used directly for process heating and as feedstocks for a diverse range of applications. Fuel and energy use by industry is heterogeneous, meaning even a single product group can vary broadly in its production routes and associated energy use. In the United States, the North American Industry Classification System (NAICS) serves as the standard for statistical data collection and reporting. In turn, data based on NAICS are the foundation of most United States energy modeling. Thus, the effectiveness of NAICS at representing energy use is a limiting condition for current expansive planning to improve energy efficiency and alternatives to fossil fuels in industry. Facility-level data could be used to build more detail into heterogeneous sectors and thus supplement data from Bureau of the Census and U.S Energy Information Administration reporting at NAICS code levels but are scarce. This work explores alternative classification schemes for industry based on energy use characteristics and validates an approach to estimate facility-level energy use from publicly available greenhouse gas emissions data from the U.S. Environmental Protection Agency (EPA). The approaches in this study can facilitate understanding of current, as well as possible future, energy demand.

First, current approaches to the construction of industrial taxonomies are summarized along with their usefulness for industrial energy modeling. Unsupervised machine learning techniques are then used to detect clusters in data reported from the U.S. Department of Energy's Industrial Assessment Center program. Clusters of Industrial Assessment Center data show similar levels of correlation between energy use and explanatory variables as three-digit NAICS codes. Interestingly, the clusters each include a large cross section of NAICS codes, which lends additional support to the idea that NAICS may not be particularly suited for correlation between energy use and the variables studied. Fewer clusters are needed for the same level of correlation as shown in NAICS codes. Initial assessment shows a reasonable level of separation using support vector machines with higher than 80% accuracy, so machine learning approaches may be promising for further analysis. The IAC data is focused on smaller and medium-sized facilities and is biased toward higher energy users for a given facility type.

Cladistics, an approach for classification developed in biology, is adapted to energy and process characteristics of industries. Cladistics applied to industrial systems seeks to understand the progression of organizations and technology as a type of evolution, wherein traits are inherited from previous systems but evolve due to the emergence of inventions and variations and a selection process driven by adaptation to pressures and favorable outcomes. A cladogram is presented for evolutionary directions in the iron and steel sector. Cladograms are a promising tool for constructing scenarios and summarizing directions of sectoral innovation.

The cladogram of iron and steel is based on the drivers of energy use in the sector. Phylogenetic inference is similar to machine learning approaches as it is based on a machine-led search of the solution space, therefore avoiding some of the subjectivity of other classification systems. Our prototype approach for constructing an industry cladogram is based on process characteristics according to the innovation framework derived from Schumpeter to capture evolution in a given sector. The resulting cladogram represents a snapshot in time based on detailed study of process

characteristics. This work could be an important tool for the design of scenarios for more detailed modeling. Cladograms reveal groupings of emerging or dominant processes and their implications in a way that may be helpful for policymakers and entrepreneurs, allowing them to see the larger picture, other good ideas, or competitors. Constructing a cladogram could be a good first step to analysis of many industries (e.g. nitrogenous fertilizer production, ethyl alcohol manufacturing), to understand their heterogeneity, emerging trends, and coherent groupings of related innovations.

Finally, validation is performed for facility-level energy estimates from the EPA Greenhouse Gas Reporting Program. Facility-level data availability continues to be a major challenge for industrial modeling. The method outlined by (McMillan et al. 2016; McMillan and Ruth 2019) allows estimating of facility level energy use based on mandatory greenhouse gas reporting. The validation provided here is an important step for further use of this data for industrial energy modeling.

Table of Contents

List of Figures

List of Tables

1 Grouping Industries

Industry is the largest end-user of primary energy in the United States (EIA 2021) and it depends primarily on fossil fuels combusted for process heating and onsite power generation and also used as feedstocks for a diverse range of applications. The renewable energy transition thus far has relied primarily on the addition of renewable capacity to the power generation sector; the use of renewables in industrial processes has grown modestly in comparison. Modeling industrial fuel and energy use is complicated, as even a single product group can vary broadly in its production routes and associated energy usage (Greening, Boyd, and Roop 2007). This heterogeneity contributes to less exploration of energy demand models for industry than other end-use sectors (Verwiebe et al. 2021). Classification schemes focused on how energy is used could aid development of such models as industries adapt to pressures to decarbonize. In turn, improved models of industrial energy and fuel use could better quantify potential savings from new technologies and aid in understanding trends and obstacles.

The industrial classification schemes that serve as the basis for data collection and reporting are at the foundation of energy systems models. Their effectiveness at representing energy use is a limiting condition for plans to improve energy efficiency and improve fuel and energy use in industry. Yet the classification schemes that divide the industrial sectors are not focused on energy use or environmental concerns. Instead, they are designed around broader categories of production activities that group similar types of products together and maintain consistency in production statistics across national borders. The challenge of relying on a classification scheme that is not based on energy use is compounded by limited data availability at the facility level, which provides a second important limitation for further development of energy systems models. Facility-level data could help describe heterogeneous sectors in more detail.

This work explores alternative classification schemes for industry based on energy use characteristics. We look at the types of facility-level data available and study them for patterns that could be used to find correlations in energy use. We take a more bottom-up approach in a detailed look at the steel industry—an industry with important impacts not only on greenhouse gas emissions, but also on basic human needs for infrastructure. Right now, the two dominant pathways to iron and steel production, electric arc furnaces (mini-mills) and blast furnaces/basic oxygen furnaces, are grouped together in existing classification schemes. The processes and their associated energy use profiles are quite different, however. As carbon-neutral and electricity-based production processes become introduced for the highest fuel-consuming sectors, the same dichotomy is likely to occur for other industries, with all activities that produce the same products falling under a single NAICS code despite very different energy use characteristics. We use a classification approach based on process characteristics to trace evolution within the steel industry and show different groupings that may have similar energy use profiles. Finally, we seek to make more facility-level fuel use estimates available via the validation of combustion energy use estimation methods that use publicly available data from the United States EPA Greenhouse Gas Reporting Program—see McMillan et al. (2016); McMillan and Ruth (2019)—using confidential microdata at the Georgetown University Census Research Data Center. While each research piece is distinct, together they provide an introduction into modeling techniques and data that can be used to improve energy modeling for industry.

The next sections provide brief background on approaches traditionally used to classify industry. Several approaches to taxonomies and classification are surveyed with respect to their usefulness for industrial energy modeling. The design of the research described in this report is given in Figure 1 (page 2). We use unsupervised machine learning techniques to cluster facility-level data (Wachs and McMillan 2021) while exploring a new way of using data to derive industrial classification systems. This clustering approach, while different from the North American Industry Classification System (NAICS), is likewise a phenetic[1] approach, that is, based on observed similarities and/or measures of similarity.

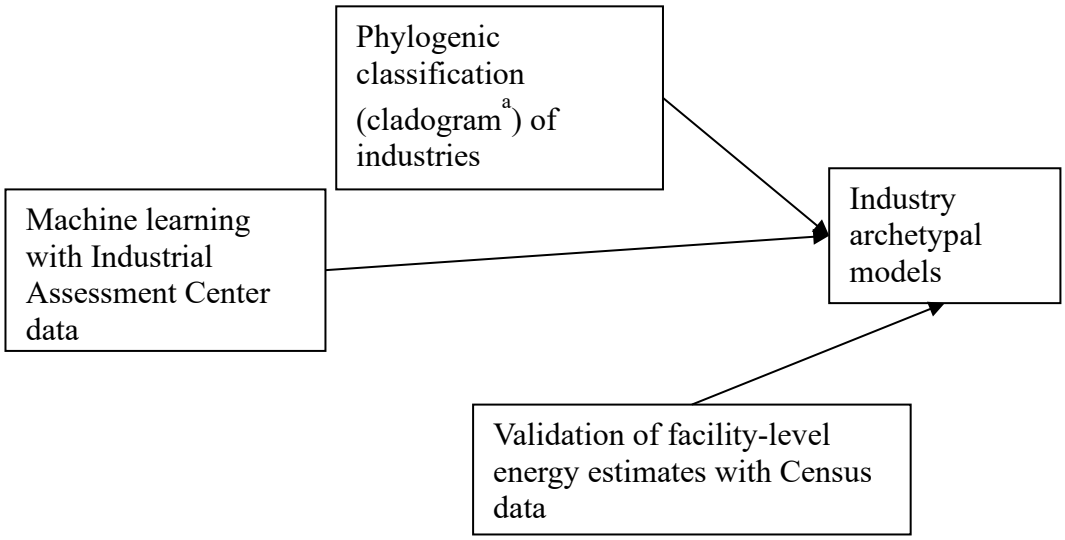


Figure 1. Approaching the definition of industry archetypal models: Methods followed in this work The validation work is foundational to future work and analysis in this area.

[a] A cladogram is a tree-like, hierarchical diagram that shows the shared evolutionary history of groups.

Decarbonization can be thought of as an evolutionary pressure that will drive new adaptations in industrial systems, such as substituting new processes or fuels (e.g., blue or green ammonia, and power-to-X processes) for traditional fuel-based processes. An industrial taxonomy—and resulting archetypal mass and energy models based on phylogeny[2], or evolutionary development—may prove more useful for capturing innovation, diffusion, and behavior under decarbonization policies. To test that approach, we applied an evolutionary method of classification (cladistics) to the iron and steel sector. We are unaware of a cladistic classification of industries based on their use of energy and materials.[3]

[1] Phenetic refers to the characteristics of an organism (biology) or entity (more broadly), regardless of the perceived relationship between the species sharing the characteristics. Characteristics can be shown by appearance or form (morphological), or may be other types of similarities that can be observed. Some of the characteristics we used in phenetic approaches here include number of employees, total sales and square footage.
[2] Phylogeny is a term used primarily in biology, described by Encyclopedia Britannica as: "the history of the evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms" (Gittleman 2016). Phylogenies represent hypotheses, as they generally rely on incomplete evidence.
[3] Baldwin's classification of industrial ecosystems (James Scott Baldwin 2008; James S. Baldwin 2011) is similar, as it does include physical characteristics.

1.1 Economic Classification

Industrial classification systems of national and international scope have historically been maintained to measure economic activity. The first Standard Industrial Classification (SIC) for the United States was issued in 1938; it was subsumed into the NAICS in 1997. The United Nations issued the first International Standard Industrial Classification of All Economic Activities in 1949. These systems follow a hierarchical approach, grouping similar activities together according to a mix of production domain and materials. They have been adopted widely for data collection related to energy and the environment, for example in the U.S. Energy Information Administration's Manufacturing Energy Consumption Survey (MECS) and the U.S. Environmental Protection Agency's Greenhouse Gas Reporting Program (GHGRP), which uses a distinct classification method[4] but also records the NAICS codes.

NAICS incorporates a supply-side approach, meaning an attempt is made to group similar production activities and processes in industries (Executive Office of the President Office of Management and Budget 2022).[5] Activities are based in part on product characteristics (e.g. metals; plastics), but also on the type of production process (e.g. agriculture; manufacturing; service). Note that the activity may not reflect process details, for example wet and dry corn milling for ethyl alcohol production are grouped together; wet corn milling for food production has a separate six-digit NAICS code.[6] Each physical facility is assigned a NAICS code according to the primary activity conducted there; for example, the headquarters of a mining company would be placed in the management code. This allows similar employment types to be grouped together. In some cases, multiple activities are performed in a single site. NAICS designations are then made usually based on the activity with the highest proportion of facility costs or the final activity in a vertically integrated process. Steel mills are an exception to this, with all facilities that produce steel grouped together despite downstream processing. Sometimes proxies such as employment or revenue rather than costs are used to make the NAICS designation.

---

[4] The GHGRP reporting follows a different assignment scheme than NAICS. A single facility may report emissions that fall into multiple categories, called subparts. For example, a single facility classified by NAICS into 325998, 'All Other Miscellaneous Chemical Product and Preparation Manufacturing,' reported emissions corresponding to subparts C, L, N and OO: Stationary combustion, Fluorinated GHG Production, Glass Production and Non-CO_2 Industrial Gas Supply. Since the program is only relevant for large emitters of greenhouse gases, it does not have to present a comprehensive classification scheme. Some industries grouped together in NAICS are more finely defined, such as NAICS 331492 Secondary Smelting, Refining, and Alloying of Nonferrous Metal (except Copper and Aluminum), whose process emissions are represented by subparts R Lead Production, T: Magnesium Production, GG: Zinc Production. Fossil fuel production and distribution receives particular attention, with distinctions between onshore and offshore production, storage, processing, transmission and other stages separated out. Hydrogen production and geologic sequestration of carbon dioxide also have their own subparts. Still, manufacturing besides petroleum refining is represented by just 24 subparts for process emissions.

[5] Supply-side means that the producers and their activities are the defining feature for the classification system. Product-based codes such as the International Harmonized System and the North American Product Classification System are considered demand-side approaches, since they define codes based on the characteristics that the end-user. Most of the time, the supply-side approach still groups similar products together, but it is possible for the same product to derive from multiple NAICS codes if the production activities are different. For example, carbon dioxide is a co-product of several industries. Air products are made on-site at diverse manufacturing plants.

[6] Electrification of industry (i.e., production processes based on electricity instead of fuel combustion) have not generally been separated into different NAICS codes, as seen in the case of steel mills. Unlike steel, however, aluminum production is split into two NAICS codes: one for primary production from alumina (an electrolytic process) and one for secondary production from scrap or dross (a thermal process).

1.2 Biological Approaches to Classification

In biology, systematics, or taxonomy, is the study of organisms and their relationship with each other, usually in terms of classification and nomenclature. Species are delineated most notably in terms of their ability to interbreed, but there are always some vague designations (Simons 2013). This vagueness refers to species that are closely connected; perhaps they are capable of interbreeding but seldom do, for example. Biological classification was originally performed on the basis of observed characteristics and their similarities (phenetics). Over time, this has been replaced by classification on the basis of genealogy (Simons 2013). Biological classification efforts can be placed into four schools (McCarthy 1995):

1. **Essentialism:** Species have an essential characteristic that does not change over time.
2. **Nominalism:** All individuals are different; classification is an artefact of human thought.
3. **Numerical Taxonomy/Phenetics:** Statistical techniques can be used to identify similarities between organisms and to form groups.
4. **Cladistics/Phylogenetics:** Evolutionary relationships can be used to define clades, or groups of organisms with a common ancestor.

Classification is distinct from identification, which places an observed individual into a previously identified group (Simons 2013). A taxon can refer to any grouping in a classification, whether species, genus, or a more inclusive grouping, e.g. carnivores (Simons 2013). Categories are groupings of taxa that inhabit the same level of the classification structure, such as all species or all genera (Simons 2013).

Phylogenetic classification, which is now typically associated with cladistics, involves taxonomic classification based on inherited traits and evolution. Therefore, species that closely resemble each other might be quite distant on the phylogenetic tree if their evolutionary paths were distinct. Cladistics is currently the dominant method for species classification in biology. It was introduced by Hennig in 1950 but became well known by the moniker phylogenetic systematics after his publication in English (Hennig 1965).

The concept of applying biological classification approaches to manufacturing systems was first explored by McCarthy (1995), as models based on appropriate, accurate and general classification schemes could improve understanding of manufacturing, reduce time and costs for modeling, and improve the applicability of model-based solutions (Baldwin 2011). McCarthy proposed five essential principles for manufacturing classification (McCarthy 1995):

1. The classification must be based on key characteristics.
2. A general classification is more important than a specialized classification for understanding and predicting system behavior.
3. The classification should be parsimonious, with differentiations between systems made with the fewest number of taxa (a unit of classification).
4. Taxa are arranged hierarchically.
5. Classification should not be specific to a time period but should allow systematic analysis of past and future systems.

McCarthy determined that existing manufacturing classifications based on operational characteristics, operational objectives, or operational flow structure lacked objective bases for analyzing taxa (McCarthy 1995). Avoiding subjective classification is a primary advantage of cladistic classification and is a feature that distinguishes the approach from other taxonomic systems (Baldwin, Rose-Anderssen, and Ridgway 2012). Inspired by McCarthy's predication, this work looks at two biological methods, phenetic and phylogenetic, to classify industries.

1.2.1 Phenetic Classification

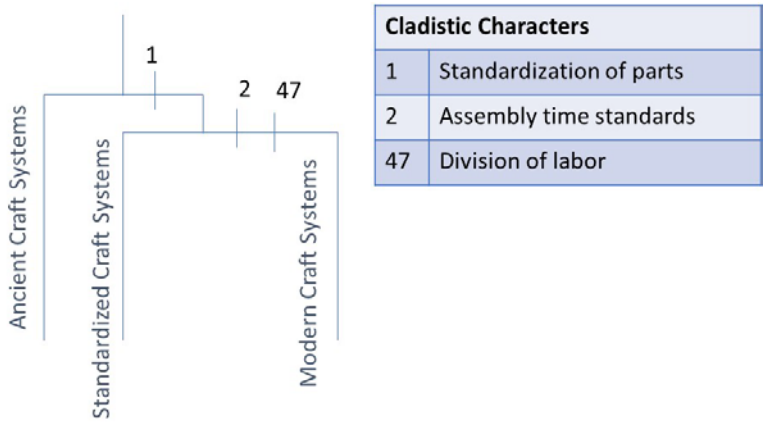
Phenetics is based on observed features and similarities: it groups items together more closely when they seem similar. Before cladistics became widespread, phenetic classification schemes traditionally underpinned taxonomic systems. Phenetic approaches can be useful when large amounts of data on characteristics are available. In this work, machine learning approaches are used to perform clustering based on data on the relationship between energy use and facility size, number of employees, plant area, and hours of operation. These variables are examples of qualities that can be observed and recorded. Such approaches can aid in understanding relationships or correlations in smaller portions of a data set, or clusters, even if these clusters do not share the same production activity. Recently, an energy end-use taxonomy was developed to assist with deploying energy-efficiency measures (Kanchiralla et al. 2020). Although their work is hierarchical in nature, Kanchiralla et al. (2020) take a phenetic approach and characterize energy use based on similarities (e.g., creating separate groupings of support processes and production processes). Phenetic classification is typically used first to identify groups, which are then validated using cladistics (McCarthy et al. 1997). Given that phenetic classifications are much more common when describing industries, we use them, including NAICS, to initially guide our phylogenetic classification.

1.2.2 Phylogenetic Classification (Cladistics)

Although cladistics creates classification schemes, more broadly it is an attempt to understand systems that evolve, have ancestors, show speciation, and are subject to natural selection (McCarthy et al. 1997). Grouping organisms on the basis of their evolution provides insight into the speed, adaptations, and morphology surrounding the changes. Still, each cladogram represents a hypothesis about evolutionary history. In biology, cladograms are generally inferred from DNA sequences (Warnow 2017). Cladistics have also been adopted more broadly, including by linguistics (Nakhleh et al. 2005), archaeology (O'Brien, Darwent, and Lyman 2001) and industry (McCarthy 1995), among other disciplines. McCarthy et al. (1997) produced the first phylogenetic classification of industrial systems. It focused on automotive assembly facilities. Cladistics has also been applied in studies of manufacturing to aerospace supply chains (Rose-Anderssen et al. 2009), the hand tool industry (Leseure 2000), general manufacturing systems (James Scott Baldwin et al. 2013), industrial ecosystems (James Scott Baldwin 2008), and product repair industries (Raza, Ahmad, and Khan 2018).

A portion of the automotive manufacturing cladogram is shown in Figure 2. Each branch of the cladogram is numbered with the characters that are significant in species evolution. For example, the standardization of parts is shown as Character 1 (McCarthy et al. 1997), the single character that separates ancient craft systems from all subsequent craft systems. Farther down the evolutionary line, McCarthy et al. identify agile automation and parallel processing as the characters that separate agile producers from lean producers. This cladogram, as well as the other

industrial cladograms in the literature, is primarily driven by characters focused on organizational management. That is only one of the drivers of innovation, which are described more thoroughly in the next section.



Cladistic Characters	
1	Standardization of parts
2	Assembly time standards
47	Division of labor

Figure 2. Example of industrial cladogram depicting the initial evolution of automotive assembly facilities

Reproduced from data in McCarthy et al. 1997

1.3 Evolution in Industry

Though the cladograms of industries are helpful, so far, they have primarily been compiled on the basis of management techniques that seek to maximize production and cost efficiency. Therefore, they are somewhat removed from the physical nature of products, including energy use. Although they have been useful in understanding the human behavior behind the evolution of industries, technological change and innovation within particular industries is frequently tied to the products themselves.

Still, cladistics is a flexible technique. Its use for archeological phylogenies (O'Brien, Darwent, and Lyman 2001) does rely on physical characteristics of artifacts. This is somewhat closer to the approach taken in this study, which focuses on physical characteristics of products and processes. Schumpeter's classic paradigm perceives innovations in products or production processes, or in markets or supply, or in management or organization, as the key drivers of economic development (Schumpeter and Elliott 2012). Specifically, Schumpeter describes new products, new production methods, new markets, new supply sources, and new organizations (Schumpeter and Elliott 2012). Thus, technological breakthroughs can offer opportunities to firms that can use them to create new products or who adopt them into their processes to make them more efficient. These breakthroughs can also offer new markets for processes that now require different ingredients, or allow existing processes to switch suppliers.

Table 1 lists examples of drivers of progress in the renewable transition space. Products act as drivers of progress, either as results of a more climate-friendly manufacturing process or as enablers of a more climate-friendly manufacturing process in other sectors. Production methods examined in the climate change space tend to be targeted toward (1) adjustment of a process to allow flexibility or use of alternative fuels, (2) carbon capture or other emissions savings, and (3) circular economy, including recycling and remanufacturing. Markets may be new consumers for existing products, for example wind turbine manufacturing as a market for steel products, or market segments such as firms that are specifically interested in purchasing lower-carbon intermediates; or, they may be able to benefit from changes in the final product that result when a traditional production method is altered (e.g., hot briquetted iron [HBI] is a final product from direct reduction of iron that differs from pig iron, the product from blast furnace production).

Supply sources offer opportunities as well; for example, in some energy-intensive industries, the supply of renewables may provide impetus for innovation, particularly considering the different risk profiles of renewables and conventional fuels. Organizational structures such as vertical integration can offer opportunities as well, for example by diversifying risk profiles in the asset portfolio.

Table 1. Drivers of Innovation Applied to Renewable Energy Transition

Drivers of Progress	Examples
Products	Green chemistry products, carbon credits, renewable tax credits, upcycled products, green fuels, biobased products
Production methods	CCUS, recycling/remanufacturing, electrification, power-to-X, use of alternative fuel or feedstock
Markets	Green power market associations, market for green products, market for lower-carbon products (can be driven by certifications and incentives)
Supply sources	Electricity from renewables, green hydrogen, alternative fuels
Organization structures	Vertical integration that allows different risk profile due to contracts with renewable generation

Schumpeter's framework is agnostic as to external forces leading to change, placing the system boundary around the firms in a particular sector. Still, in the case of technological changes related to the energy transition, firms and even industries do not act totally independently. Instead, they are part of a "technological regime," which also encompasses other actors, practices, institutions and infrastructures (Rip and Kemp 1998). Smith, Stirling, and Berkhout (2005) examined regime changes in the context of three factors: (1) the description of the change as being oriented toward a particular problem or goal, (2) resource availability[7] for the incumbent regime, and (3) coordination of responses (adaptive capacity). They devised a conceptual framework classifying regime changes into those with high external coordination of the adaptive response and external resources for adaptation ("purposive transitions"), high external coordination of adaptive response and internal resources for adaptation ("endogenous

[7] Resource availability or "environmental munificence" refers to the level of economic interest in the area; research funds, investment funds, access to capital, high prices for produced goods, good salaries versus a more resource-constrained environment.

renewal"), low coordination of adaptive response and internal resources for adaptation ("reorientation of trajectories"), and low coordination of adaptive response with external resources for adaptation ("emergent transformation"). Smith, Stirling and Berkhout classified the renewable energy transition into the first category, of a purposive transition (2005).

Climate change is the overriding concern motivating the renewable energy transition, which means levels of greenhouse gas (GHG) avoidance and time frames are subject to a high level of coordination, with pressure for adaptation coming outside individual regimes such as the energy regime or metals production regime. Therefore, unlike a technological change such as the adoption of combined-cycle gas turbines for power production, which came from technological changes developed as a response to many different challenges and were not coordinated within the sector—and so was not possible to predict—some characteristics of the changes taking place because of the renewable energy transition may be predictable, even if their exact form may be unknown.

Cladistics may provide a method of understanding the relationship of industrial development to the evolutionary pressures applied by the renewable energy transition. Specifically, cladistics can track how industries have begun to innovate and change in this period. The cladogram provides a visual map of the emerging evolution at a snapshot in time. It is easily possible to see where most branching is occurring and what types of processes are emerging. Comparisons of processes and their impacts in different branches of the cladogram can provide insight into how much GHG emissions could be curbed, which synergies with other industries may be likely, and which types of solutions are being actively developed (e.g., whether green fuels, electrification or carbon capture, utilization and storage [CCUS] is most prominent).

As mentioned previously, much of the work in cladistics for industrial classification has been focused on a single category: organizational structures. We choose a case study of the iron and steel industry to explore and demonstrate the applicability of cladistics to industrial classification in the renewable transition space. Rather than focusing on organizational structure, we broaden the classification approach to encompass all five drivers of innovation.

1.4 Data Needs for Classification

Any novel classification approach needs to be supported by data, particularly at the facility level so that conclusions are not confined to the current classification scheme. Until recently, the only publicly available source of U.S. facility-level energy data was the database of the Industrial Assessment Centers (IACs) sponsored by the U.S. Department of Energy ("Industrial Assessment Centers" n.d.). The IAC is a program whereby small and medium firms can apply to receive energy and waste audits and recommendations for energy savings, production issue diagnoses, information on opportunities for smart manufacturing, and cybersecurity enhancements. The centers are hosted at universities, providing an opportunity for students to gain practical experience. The facilities that participate must release anonymized information about their energy and water use as well as their location and industrial classification. The IAC program has worked on close to 20,000 facilities since its beginning in 1976.

Approximately 8,000 U.S. facilities are required to report annual GHG emissions under the GHG reporting rule (*Mandatory Greenhouse Gas Reporting* 2009). The rule applies to facilities that emit at least 25,000 t/year of CO_2 equivalent emissions of greenhouse gases, including carbon

dioxide, methane, nitrous oxide, hydrofluorocarbons, perfluorocarbons, sulfur hexafluoride, and select fluorinated gases such as nitrogen trifluoride and hydrofluorinated ethers. The EPA classifies the facilities into sectors that include power plants, mining and waste, and sectors that mostly include manufacturing, such as miscellaneous combustion (which also includes service industries), refineries, chemicals, metals, minerals, pulp and paper, and electronics manufacturing. A technique has been established to use emissions data reported under the GHGRP to estimate fuel use by facility (McMillan et al. 2016; McMillan and Ruth 2019). Still, the approach has not been validated, which has limited its use.

The MECS is currently performed every 4 years to record detailed information about manufacturing energy use, including consumption, costs, cogeneration, use of waste heat, and fuel-switching ability ("2014 Manufacturing Energy Consumption Survey Methodology and Data Quality: Survey Design, Implementation, and Estimates" n.d.). The data are processed and released to provide information on geographic and sectoral patterns while maintaining the confidentiality of the businesses surveyed. The facility-level data are not publicly available, but researchers with Sworn Special Status can apply to access it for approved projects inside secure U.S. Census Bureau research data centers, which we did for this work.

2 Statistical Classification

In this section, we examine available data on facility level energy use to find groups with particularly high correlations between energy use and other, observable characteristics. Ideally, such a grouping could represent an "archetype" to be studied further and used as a basis for more detailed energy models. The data set used is described in Section 2.1. Statistical approaches are used to understand the data and measure correlations based on existing classification schemes (Section 2.2). Then, in Section 2.3 unsupervised machine learning approaches are used to detect groupings from the data itself.[8] Because classification approaches represent a black-box type technique, where it is not immediately clear how the differentiation is made, an additional step of applying support vector machines, which quantify how easy it is to separate the resulting clusters, was also used in Section 2.4. The results of this work are explored in Section 2.5.

2.1 Description of the Data Set

We initially attempted to identify an evidence-based industrial classification by examining the IAC database using both unsupervised and supervised machine learning techniques to see whether meaningful clusters could be detected based on energy use; see also (Wachs and McMillan 2021). When these experiments were run, the IAC data set included 19,435 records reflecting operations as far back as the 1970s. A total of 5,027 records for 2010–2021 were used for this analysis. To give a sense of the relative numbers of industrial facilities participating in the IAC program, between 2010 and 2020 the IAC covered approximately 0.10–0.15% of manufacturing establishments each year (Wachs and McMillan 2021). The Bureau of Labor Statistics show 12,231,000 employees in manufacturing in December 2020. The IAC data set included 63,923 employees in 2020, or 0.5%.

Although the manual for the IAC database states that facilities should have fewer than 500 employees (Muller 2011), 239 facilities reported having a higher number. The largest number of employees for a facility in the data set was 6,500, and the average number was 177 employees. And the manual states that the energy bills should be between $100,000/annum and $2,500,000/annum, but the data set includes energy bill values ranging from $85 to more than $88,000,000. The facilities are also not evenly divided between manufacturing NAICS codes. More than half are from the following five three-digit codes: 311 (Food Processing, ~12%), 325 (Chemical Manufacturing, ~7%), 326 (Plastics and Rubber Products Manufacturing, ~10%), 332 (Fabricated Metal Manufacturing, ~13%), and 333 (Machinery Manufacturing, ~8%). The portion of facilities in Code 221 (Utilities) has been increasing in recent years.

The descriptive data for each facility includes NAICS or SIC codes, plant area, sales, hours, production (mass or volume), state, product or products, and number of employees. Twelve types of energy use are recorded in monetary and energy units. Water consumption and disposal, and liquid, solid and gas disposal are also recorded in monetary and physical units. Still, most fields are sparsely populated. Therefore, we summed all fuel use types in their energy units to make a total fuel use variable. Electricity usage and demand were well populated. Production in mass

[8] Bzdok, Altman and Krzywinski summarized the difference between statistics and machine learning thus: "Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.(Bzdok, Altman, and Krzywinski 2018)"

and volumetric units were not comparable, so were not considered. The products field, which is a qualitative field, was not used. The final set of variables used for our analysis included electricity use, electricity demand, total combustion fuel use, production hours, number of employees, plant area, and sales.

Preprocessing was performed on the data set. Duplicate records were removed. Then, we checked for NAICS codes outside the range denoting manufacturing (two-digit prefixes between 31-33) and examined the data for these facilities manually. Three facilities with NAICS codes above 339 were reclassified based on this inspection, and all others except facilities in NAICS code 562 (waste disposal) were excluded from the analysis. Facilities with NAICS codes below 311 were included (representing sectors 11, 21, 22 and 23); most facilities with these codes were wastewater treatment or water distribution with some in the categories of construction, pipes, and agriculture. The annual average U.S. Bureau of Labor Statistics' Producer Price Index was used to adjust all sales figures to 2019 figures, according to the following equation:

$$S_A = S \times \frac{PPI_{2019}}{PPI_{Annual}}$$

where S_A is adjusted sales (used in the analysis), and S is the total sales figure from the data.

Because a high level of skew was found in the data set, transformations were applied (a normal distribution is assumed for clustering methods); a log transform was used on electricity demand and consumption and the Box-Cox transformation was used on all other variables except hours (log transform was not sufficient to remove skew in most of the variables). The StandardScaler function (Pedregosa et al. 2011) was applied to transform each variable to give a mean of zero and a variance of one. Outliers were detected and values more than 2.5 times the interquartile range were removed to avoid their effects dominating the clustering. The final data set included 4,248 observations.

The analysis then proceeded as follows:

1) **Initial Preprocessing of Data:** Duplicates removed. NAICS codes outside 11, 21, 22, 23, 31–33, and 562 reclassified or removed. Box-Cox and log transformations applied. StandardScaler function applied. Outliers detected and removed.

2) **Supervised Approach:** Using NAICS codes as clusters, Pearson correlation between variable pairs run and visualized. This is described in Section 2.2.

3) **Unsupervised Approach:** Ratios of energy to predictor variables calculated. K-means clustering performed for a range of k (number of clusters) values. Hierarchical clustering run, visualized, and analyzed. This is described in Section 2.3.

4) **Application of Support Vector Machines to 20 Hierarchical Clusters:** This is described in Section 2.4.

2.2 Supervised Learning Approach: Pearson Correlation between NAICS and Energy Use Variables

In the first step, we measured the Pearson correlation between the three energy use variables and the predictor variables: production hours, number of employees, plant area and sales, within each NAICS code. This was done to create a baseline to which the unsupervised clustering approaches could be compared. The heat maps in Figure 3 provide a visual clue about how the different variables are correlated according to NAICS groupings. First, it is evident that many of the variables are sparsely populated. No strong correlations seem evident for any of the fuel streams or material waste streams. Electricity consumption and electricity demand rarely show strong correlations with any of the variables, although the number of employees and plant area do show a correlation in some NAICS codes.

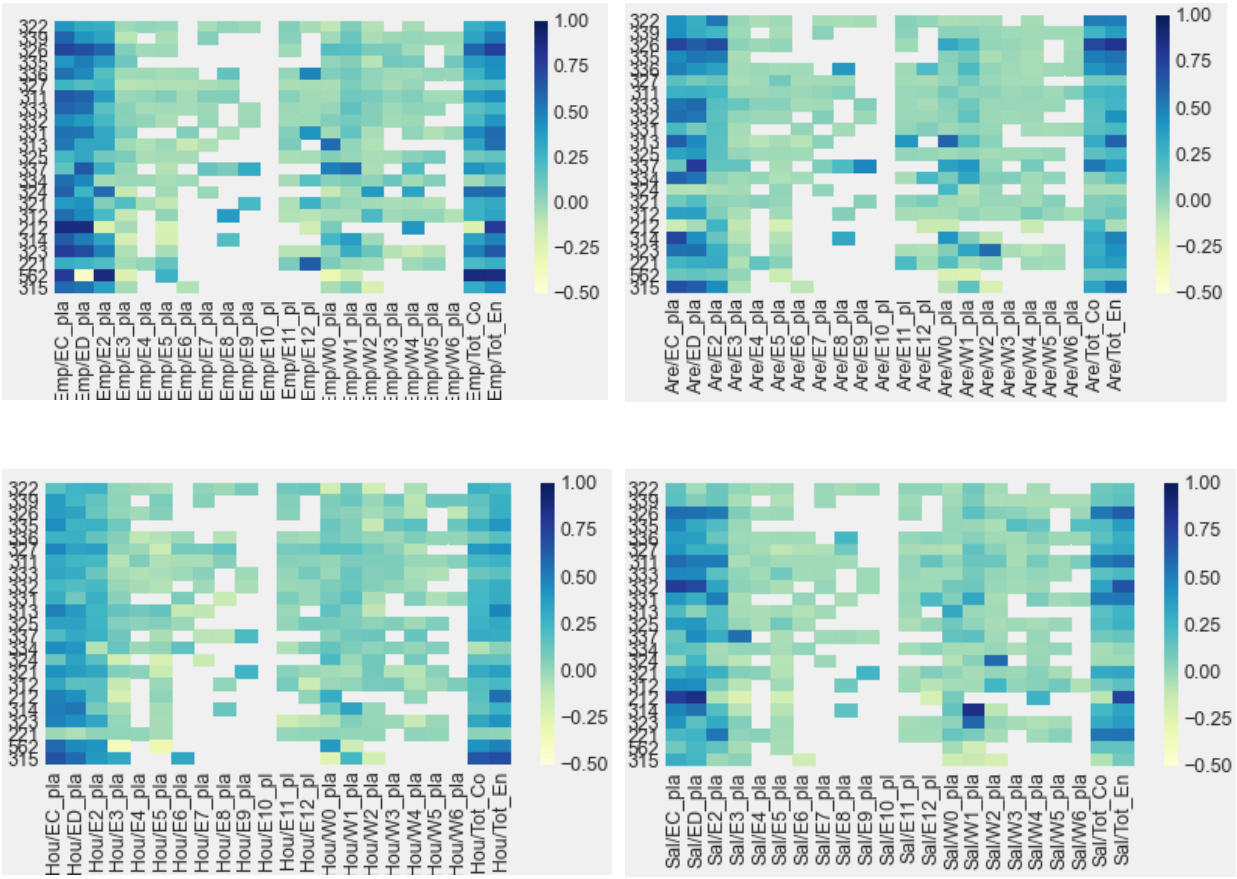


Figure 3. Pearson correlations between number of employees (Emp), plant area (Are), plant hours (Hou), and annual sales (Sal) and each of the energy consumption variables is shown for every three-digit NAICS code

Blank spaces indicate insufficient observations (<20) for correlation calculation.

EC is electricity use (in energy units). ED is electricity demand (in power units). E2 is natural gas, E3 is coal, and E4-E12 are other fuel types. W0-W6 represent waste streams. Tot_Co is the sum of all fuel consumption and Tot_En is the total energy including fuel and electricity.

2.3 Unsupervised Learning: Clustering

We calculated ratios of each predictor to each response pair, and we used these as the data set for unsupervised clustering. Unsupervised learning allows patterns and groupings to be found from the data set without making assumptions about independent and dependent variables. Clustering, a search for these groupings, is the most common unsupervised approach, and hierarchical and k-means are two prevalent clustering techniques. Hierarchical clustering begins with the assumption that every observation represents a single cluster. In each step, the distance between every pair of clusters is computed and the lowest score below a particular criterion results in those clusters joining to form a new cluster. Once all the clusters have been joined into one, the algorithm stops. For this work, the scipy Python library was used with the Ward distance metric (Virtanen et al. 2020). The resulting dendrogram is shown in Figure 4. At the base of the dendrogram, each observation is represented by a single node. The vertical distance in a dendrogram shows how closely related the clusters shown are. At any point on the y-axis, a horizontal cut shows how many clusters are similarly closely related. This dendrogram seems to show a natural categorization of the data into three main groups, which are differently colored in Figure 4. Clear distinctions are still visible in almost 50 clusters.

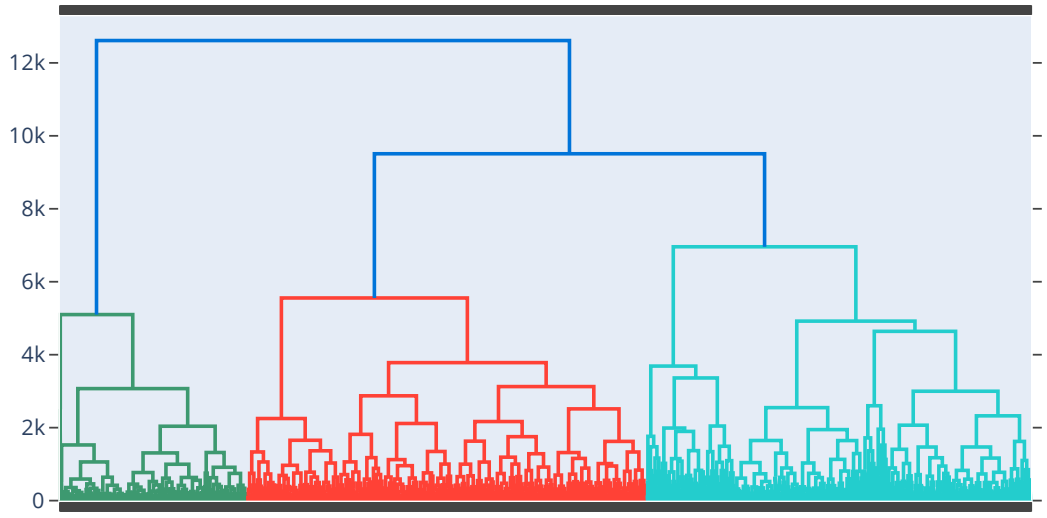


Figure 4. Dendrogram of hierarchical clusters formed on the IAC data set

The second clustering approach is the k-means algorithm. The k stands for the user-defined number of clusters. In the first step, k centroids are randomly assigned over the state space. Each observation is then assigned to the nearest centroid. Centroids are recalculated again on the basis of the assigned datapoints. Once the centroid locations do not change, the algorithm is stopped. For this work, 25 repetitions were completed with a random seed of 42. The algorithm was run for 1–100 clusters. The scikit-learn Python package was used (Pedregosa et al. 2011). Because results were better for hierarchical clustering, k-means results are not discussed here.

2.4 Supervised Approach: Support Vector Machines

Once clusters have been designated, it is necessary to delineate their boundaries in the state space so that new observations can be assigned to the correct cluster. This is done by classification. We used support vector machines for this step. Building on the concept of a maximal margin classifier, which is a linear approach of classification between two classes via a hyperplane that

leaves the maximal margin between the classes, support vector classifiers efficiently separate clusters allowing a soft margin, which can be trespassed when linear separation is not possible. Support vector machines extend this capability to cases where linear separation is less effective than other types of functions, allowing a kernel function to be used (James et al. 2021).

The scikit-learn Python package extends the capability from two classes to multiple classes via a one-versus-rest implementation[9] (Pedregosa et al. 2011). In addition, because support vector machines require a kernel function, as well as parameters C (how soft the boundaries are, the number corresponds to the number of observations that can be misclassified), and gamma (in the case of the radial basis function), to be determined, they are found using an exhaustive search via cross validation. While support vector machines are deterministic once the parameters have been set, the division of the data into training and test sets is stochastic, so a random seed of 101 was used.

2.5 What Facilities are in a Cluster?

When allowing machine learning classification of facilities according to characteristics of their energy use, it was unclear to what degree the results would mimic the facilities' original NAICS classification. So, we conducted additional analysis of the resulting clusters to understand their size, relationships to predictor variables, and industry compositions. Figure 5 shows the correlations between predictor variables when a cutoff of 20 clusters was selected. Correlations appear to be high, with all clusters except 16 and 19 showing a moderately high correlation between at least one pair of energy use and predictor variable. This is a smaller cutoff than the 26 three-digit NAICS codes present in the data. Thus, groupings could be made using this approach that demonstrate moderate to high correlations between electricity and other characteristics using a resolution even rougher than the three-digit NAICS codes. Different clusters show different relationships, with some showing a higher correlation between sales figures and electricity, with others showing a higher correlation between number of employees or plant area and electricity. So, these groupings could allow modeling electricity use based on the given characteristics.

[9] Support vector machines are a binary classification system, so can divide between just two classes. In order to extend their functionality to larger systems, support vectors can be sought that separate between every two clusters (one versus one), or between each single cluster and the rest of the observations (one versus rest), or the problem can be formulated in a way that allows all the vectors to be constructed such as via directed acyclic graph methods.

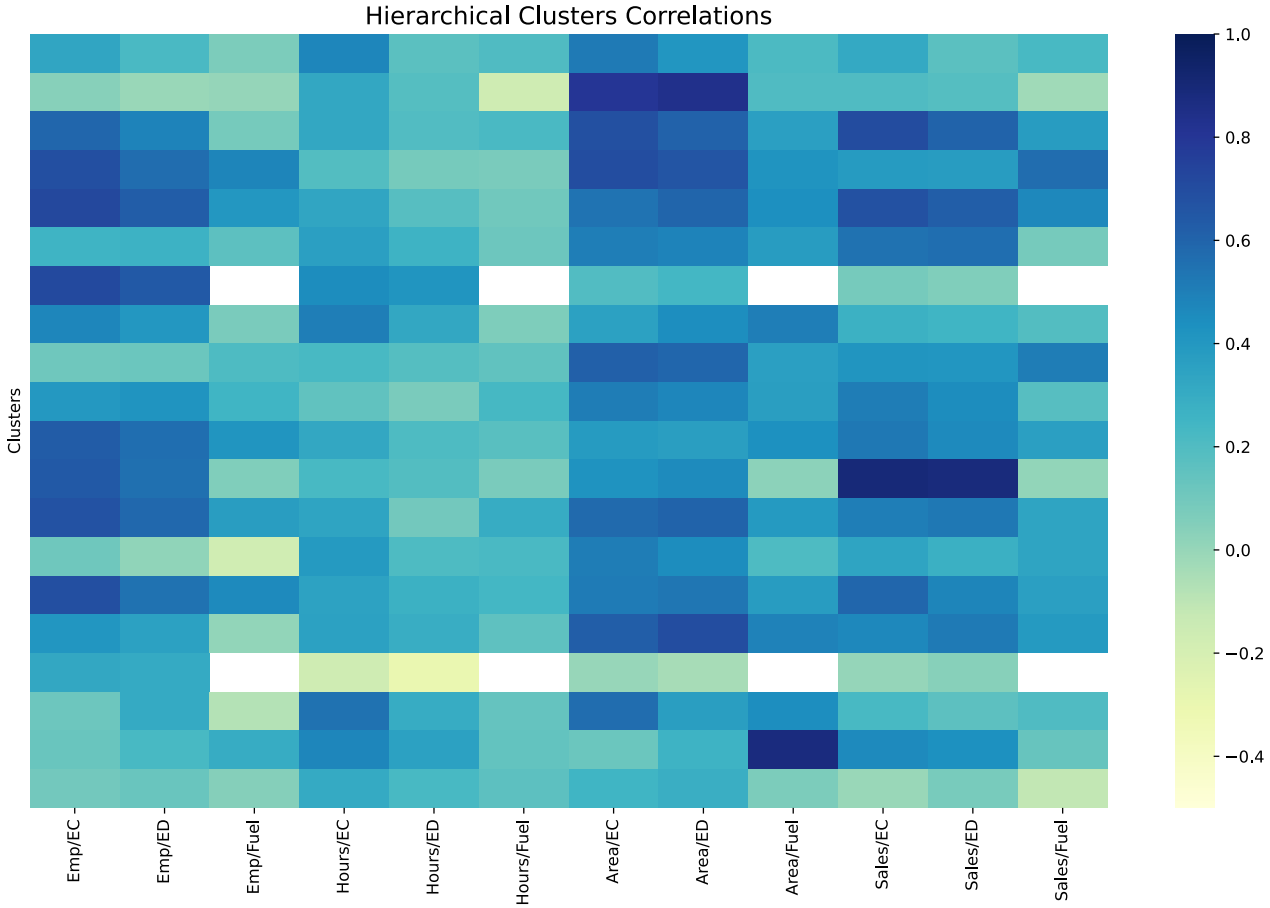


Figure 5. Pearson correlation between pairs of variables visualized on a heat map for 20 hierarchical clusters

Dark blue indicates highest correlation while the yellow color family indicates negative correlation. All but two clusters show moderate to high correlation between at least one variable pair.

Individual clusters are not dominated by any particular NAICS code, whether at the six or three-digit level. This can be seen in Figure 6, which shows clusters colored by percentage of their total members from a particular NAICS code. The clusters are fairly robust, as support vector machines can distinguish between them with 83%–88% accuracy. Every cluster contains at least five corresponding light yellow squares, including the clusters with lowest membership: 16, 18 and 19, which are all very low in fuel consumption and fairly high in sales. Cluster 16 has very low operating hours and spans the plant area distribution with high variance. Likewise, NAICS codes which contribute substantially to a single cluster, such as 221, 311, 326 and 332, contribute at a similar level to many clusters. NAICS codes that are present only in 1–3 clusters are only found in 1–3 observations. Taken together, this indicates NAICS codes do not distinguish between characteristics that predict cluster differentiation. Nevertheless, in some cases, firms in specific NAICS codes fall heavily into certain clusters. At the three-digit level, just those with fewer than three entries are found to be above 30% in any particular cluster (with the exception of NAICS 562 Waste Management and Remediation Services in the hierarchical clustering, which has a total of eight entries).

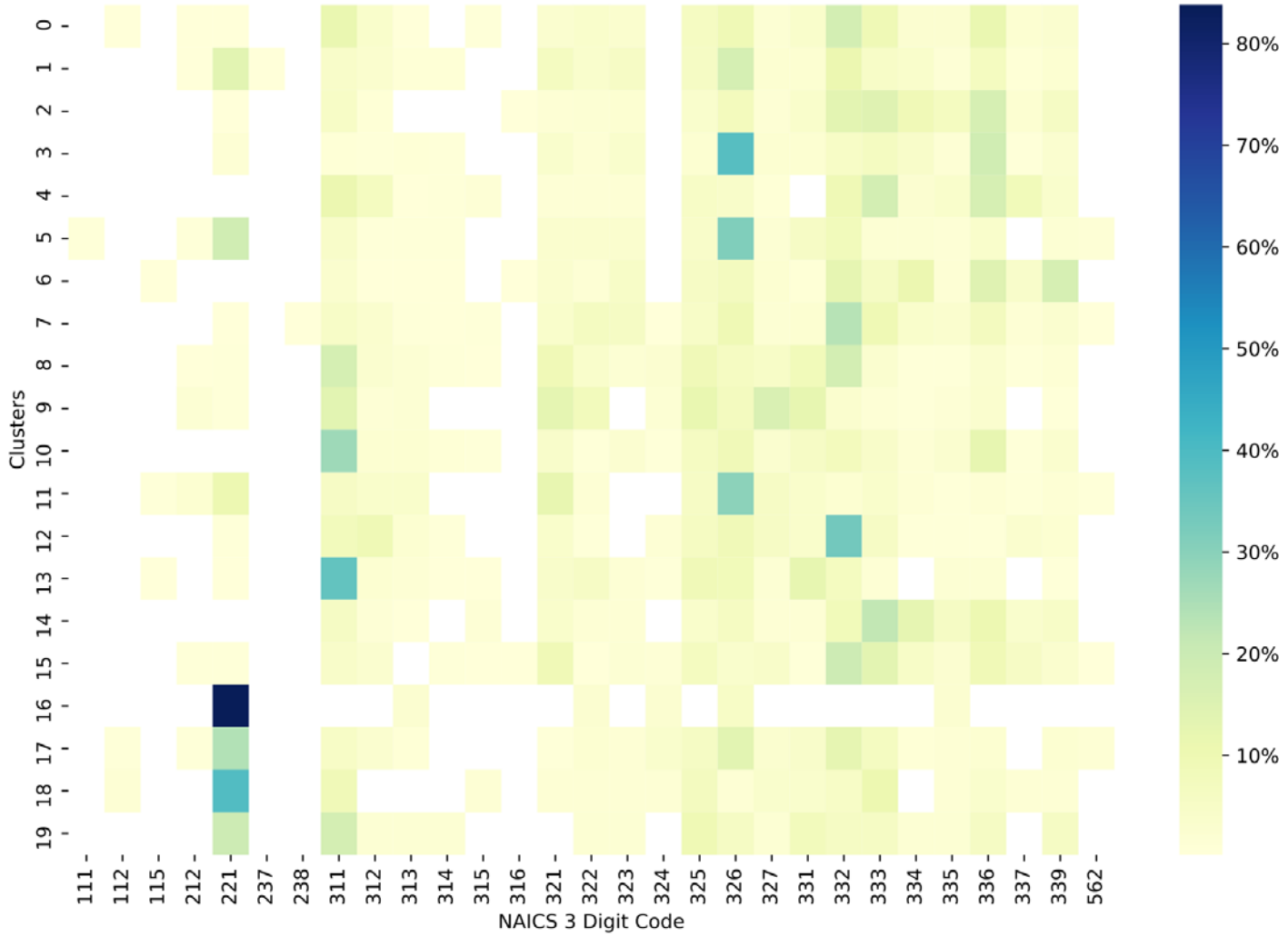


Figure 6. Percentage of total members of cluster (y-axis) pertaining to single three-digit NAICS code (x-axis)

Figure 7 shows the number of observations in each cluster. The number of facilities in each cluster found by the hierarchical method ranges from 37 to 454, with a median of 199.

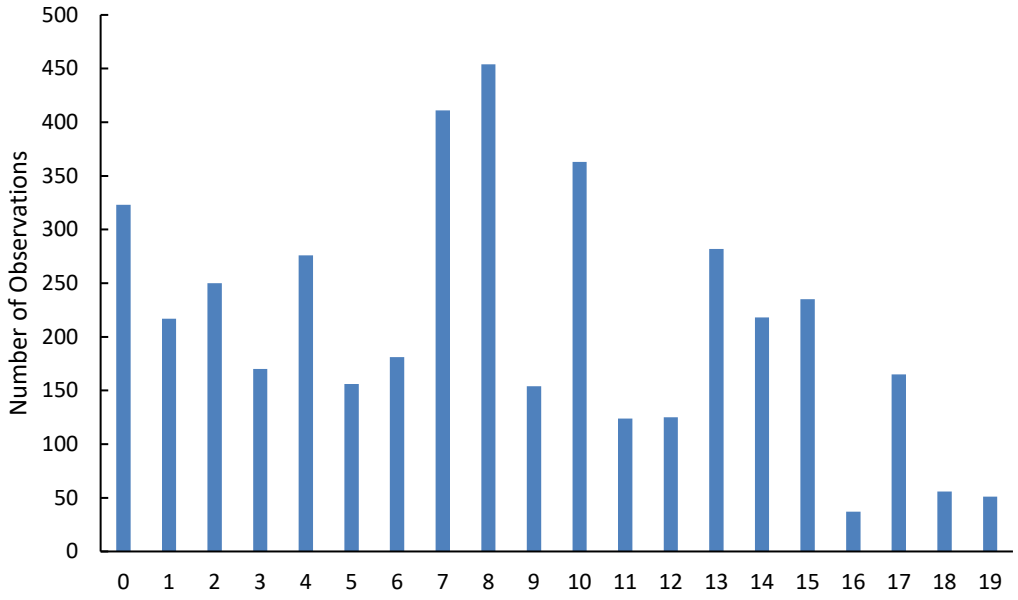


Figure 7. Number of facilities in each cluster

Section 3 focuses on a more mechanistic view of understanding these correlations. Though the phenetic classification approach via machine learning shows promise in better identifying correlations between energy use and facility characteristics, it cannot provide insight on process pathways and causes. The characteristics that are observable are artifacts of the manufacturing processes used, and a change in the process may change both the predictor and energy use variables. For a better understanding of processes as well as the changes occurring at the industry level, a complementary approach is needed.

3 Cladistics for Industry

Industrial cladistics is an example of phylogenetic classification, which seeks to trace the evolution of systems, specifically focused on the branching points (i.e., where different characteristics emerge) (Rose-Anderssen 2014). Thus, it involves denoting manufacturing "species," which is a somewhat subjective process, and tracing their evolution via branching points. This is done by labeling characters (traits) and finding the most parsimonious route of adaptation. Though cladistics has been applied broadly to manufacturing, most of the work on industrial classification relates primarily to organizational systematics. In this work, the conceptualization is expanded to include industry more generally.

A method for constructing manufacturing cladograms was suggested by McCarthy et al. (1997) and amended by McCarthy and Ridgway (2000) and Rose-Anderssen (2014). We followed a similar set of steps, given below:

1. Define the problem (i.e., why the cladogram is being constructed and what purpose it should serve).
2. Identify the "clade" of interest (i.e., the group of manufacturing systems to study).
3. Identify the "taxa." A taxon is a grouping such as a species or family that forms a distinct unit. For manufacturing systems, this can be an idealized type, or could reflect particular facilities. The study author must define the taxa.
4. Identify characters, the characteristics of the manufacturing process; these are assumed to be variables with evolutionary significance and are used for classification.
5. Code the characters: construct a character matrix including all the taxa identified in step 2 as well as all characters from step 3. Assign scores for each taxon for each of the characters (e.g., a binary score, 0 means character is absent, 1 means it is present.)
6. Construct a conceptual cladogram (discussed in detail in Section 3.1).
7. Construct a "factual" cladogram (i.e., validate the conceptual cladogram through site visits and interviews).
8. Establish a nomenclature.

This approach is applied to construct a cladogram for the iron and steel industry in Section 3.2. The resulting cladogram is presented and discussed in Section 3.3.

3.1 Constructing a Cladogram

Phylogenetic trees, also called cladograms, must be inferred after taxa and characters have been defined. A phylogenetic tree consists of any network with connections between all the taxa that uses all the characters. Many trees are possible, so different theoretical approaches are used for tree construction and selection. Maximum parsimony is one such approach. It assumes that the most optimal tree is the one with the minimum character changes. Parsimony is therefore easy to understand, and was one of the first and most widespread principles for phylogenetic tree construction (Felsenstein 2004). This is the approach we used, for its simplicity and the availability of software tools that could allow us to perform the analysis. Multiple other approaches are possible (see Felsenstein 2004; Warnow 2017; Schrago, Aguiar, and Mello 2018

for full discussion). Any tree, or cladogram, is ultimately a hypothesis. The question of which tree is most accurate (that is the tree derived with fewest evolutionary steps, versus trees derived or chosen based on some of the other available criteria) has implications in terms of philosophy as well as science, statistics, and in this case, industry. We cannot usually assess directly how evolution occurs, although in the case of industry we may have enough sources to perform some validation. Validation is a suggested step in the method devised by McCarthy et al. (1997), but does not seem to have been implemented. We could not perform validation due to the limited scope of this work, but it is an important exercise to undertake as industrial cladogram methodologies develop.

Maximum parsimony is a version of the Hamming Distance Steiner Tree Problem (Warnow 2017). It is one of the classic NP-complete[10] problems in combinatorics (Karp 1972), requiring an exhaustive search of the solution space; thus, the optimization version is NP hard. The number of tree possibilities begins to increase rapidly once more than two taxa have been identified. Methods that include exhaustive search can be tractable with fewer than 11 taxa, but larger numbers of taxa are typically approached by branch and bound algorithms and heuristic approaches for larger numbers (David L. Swofford and Jack Sullivan 2003). For this work, Phylogenetic Analysis Using Parsimony (PAUP*) software (Swofford 2003) was used with the maximum parsimony approach.[11]

3.2 Iron and Steel Cladogram

Iron and steel are closely related materials. Iron is the most abundant element on Earth by mass (Frey and Reed 2012), but like many metals it reacts easily with oxygen. Therefore, naturally occurring iron is usually found in an oxidized state as hematite (Fe_2O_3), magnetite (Fe_3O_4) or other ores. To obtain metallic iron, the oxygen must be removed by a reduction process. Due to its properties of strength and robustness and its ubiquity, iron has been used extensively for the fabrication of tools and weaponry for at least 4,000 years (Gale 1990). Since the discovery of iron in Middle Eastern cultures, the use of iron made its way throughout the world. As its production became widespread, it was refined in bloomeries, small-scale production facilities consisting of a clay oven powered by charcoal to heat iron ore, which was manipulated via bellows and tongs into wrought iron. The bloomery technology was remarkably consistent, undergoing little essential change during an approximately 3,000-year period, until roughly 1500

[10] NP is nondeterministic polynomial time.

[11] Cladograms are usually prepared in biology, and while they can employ morphological (descriptive, structural) characters such as those used in this work, molecular evidence such as DNA or amino acids are frequently used. Most software is designed to handle molecular evidence, and fewer are available for morphological data. When cladistics was first applied to industrial systems in the late 1990's, available software frequently included morphological characters and utilized the parsimony approach. Subsequent developments have focused more on support for molecular evidence and some of the newer approaches such as maximum likelihood model. The open source implementations MacClade and Mesquite no longer support phylogenetic tree inference. The maximum likelihood model has been extended to cover discrete morphological data, but it has not been implemented in Biopython, although it is available in MrBayes (Schrago, Aguiar, and Mello 2018). The Willi Hennig Society supports an implementation of the parsimony approach that applies to discrete and continuous alphanumeric characters in a free software called tnt, which stands for Tree Analysis Using New Technology. DendroPy allows use of the Phylogenetic Analysis Using Paup (PAUP*) module, which includes treatment of morphological characters, but the version of PAUP that includes an interface that is no longer supported on Mac (Mac OS X 10.15+).

CE, with the advent of the blast furnace in Europe. The blast furnace is still used today, although at a larger scale than first employed, and powered by coke rather than charcoal.

In combination with a small amount of carbon and also with other alloying elements, iron's properties can be even more useful. The most famous and important of these alloys is steel, an alloy of iron and carbon that can include other elements as well (stainless steel includes approximately 18% chromium by mass). Mass production of steel became feasible in the second half of the 19th century with the Bessemer process, which was supplanted by open hearth processing, but is now generally done in a basic oxygen furnace (BOF). Currently, iron and steel production is dominated by China, which makes more than half the world's steel ("2021 World Steel in Figures" 2021).

In the dominant pathway to produce steel today, iron ore is mined, reduced with coke in a blast furnace (BF) into pig iron, blasted with oxygen to remove impurities including carbon to form steel in a BOF, alloyed in a ladle to meet product specifications, cast into intermediate products, and then fabricated into final products. The electric arc furnace (EAF) presents a smaller-scale, more modular alternative, which primarily uses steel scrap (scrap usually forms 100% of the feed in the United States, although it is possible to use ore-based metallic iron in the feed as well) directly to form new steel. In the United States, scrap-based steel production in the EAF is currently the dominant pathway for steel production (Tuck 2021). Still, since most steel products consumed in the United States derive from imports, the domestic use of steel is far from circular (Cooper et al. 2020).

The carbon intensity of the two steelmaking processes is quite different, with estimates of national average intensities for the BF-BOF process, which uses coal directly, ranging between >1.5 t CO_2/t crude steel and just below 3 t CO_2/t crude steel (Hasanbeigi 2022), with many estimates being close to the global average of 1.81–1.89 t CO_2/t crude steel cast (worldsteel Association n.d.). The carbon intensity of the EAF, when it avoids the ironmaking step, is estimated to be much lower, from around 0.3 t CO_2/t crude steel to 0.9 t CO_2/t crude steel, with emissions depending largely on the composition of the power grid (Hasanbeigi 2022), although oxygen is typically injected into the furnace to take advantage of chemical energy, which also produces emissions.

There has been increased investment and interest in green technologies in the iron and steelmaking sector in recent years. New technologies under development cut energy use, replace the raw materials required, and decrease direct emissions and indirect emissions from the value chain. For example, in the European Union, the *Clean Steel Partnership Roadmap* (ESTEP 2020), which includes plans to fund the steel sector with €2 billion between 2021 and 2027 to reduce GHG emissions via the development and proving of new technologies, received approval for €0.7 billion of public funds and €1 billion of private funding (Eurofer 2021). In the United States, the private sector has provided some funding in this direction, with Boston Metal raising more than $60 million of venture capital (Forster 2021). Though this is much less than the public investment Europe is making in the sector, European steel production is more carbon-intensive

than U.S. steel production[12]. And concerns about climate change have already led California to publish carbon caps for four steel materials (hot-rolled sections, hollow structural sections, plate, and rebar, or concrete reinforcing steel), which went into place in July 2022 ("Buy Clean California Act" n.d.).

Dominant and emerging pathways to steel production are shown in Figure 8 (page 22), which can be read from left to right. The first steps in steel production, shown at the far-left side of the figure, entail raw material extraction and beneficiation. Raw materials shown are scrap, lime, coal, iron ore, natural gas, and water. All these except scrap and water involve a related extraction process. Emerging processes have been developed to require less processing. For example, HIsarna and Boston Metal avoid preprocessing of iron ore into agglomerates, as well as the use of coke (although coal is still required) for the production of pig iron (like blast furnace or direct reduction steps) (TATA Steel 2020). In the case of hydrogen direct reduction H_2DRI, as exemplified in the hydrogen breakthrough ironmaking technology (HYBRIT) process (SSAB n.d.), hydrogen is used as the reducing agent, and, as the technology is designed to be fossil free, is prepared via the electrolysis of water.

For ironmaking, the BF runs on coal (coke) for the energy-intensive and high-CO_2-emitting reduction of iron ore to produce pig iron. Much research on alternative ironmaking techniques has been taking place due to this stage's responsibility for the largest share of CO_2 emissions. The eventual buildup of copper in scrap and limitations on scrap availability make some production from ore necessary assuming demand for steel remains constant or increases. Five additional ironmaking processes are shown in Figure 8. As the price of natural gas in the United States has decreased, the direct reduction of iron (DRI) using natural gas (Midrex and other processes) has begun to gain traction, roughly doubling from ~8% in 2016 to ~16% in 2020 (USGS 2022). DRI avoids the use of coke, and the hot briquetted iron (HBI) or sponge iron produced can be used in either a vertically integrated mill or an EAF.

Interest has been growing in the use of hydrogen gas (H_2) produced by electrolysis of water using renewables as the sole reducing agent for iron ore (H_2DRI) (SSAB n.d.; Bailera et al. 2021). Note that it is also possible to use a combination of electrolytic hydrogen and natural gas, but this method is not included in our cladogram development. Electrolytic reduction of iron is being developed in at least two distinct processes:

- Boston Metal, is a modular, high-temperature process for molten electrolysis that can use iron ore fines ("Boston Metal" n.d.; Boston Metal 2019).
- The development of new methodologieS for InDustrial CO_2-freE steel pRoduction by electroWINning (Siderwin), is an electrolytic process for recovery of iron in alkaline solution at lower temperatures (~110 °C) followed by the use of induction heating for the production of steel or an EAF that does not require lime inputs (Tecnalia n.d.).

In addition to these two processes, Parkinson et al. (2017) suggest a molten salt electrolysis process that could be coupled with chemicals production (e.g., ethylene, benzene) for CO_2-free

[12] While Hasanbeigi does not provide numerical estimates, interpolation from Figure 7 (Hasanbeigi 2022) shows a carbon intensity of roughly 1.25 t CO_2/t crude steel for the EU-27 countries versus slightly lower than 1 t CO_2/t crude steel for US steel production

ironmaking[13] in a process referred to here as power-to-iron/organic chemical production. Smelting reduction is an alternative to DRI, which operates at a higher temperature but needs less refining of inputs. The only process included that uses smelting reduction is the SuSteel process, in which hydrogen plasma serves as the reducing agent, allowing ironmaking with no CO_2 emissions (Seftejani et al. 2020; K1-Met GmbH 2022).

Just two steelmaking processes are included in our cladogram: the EAF and BOF. Most emerging ironmaking technologies have been designed to work with one of these two existing processes, although sometimes with minor modifications that are not discussed here (note that the power-to-iron/organic chemical production process does not mention a particular steelmaking process). Produced steel is then cast. Nucor has invested in a process called Castrip that casts EAF steel directly into thin sheets, but all the other processes involve the preliminary casting stage before moving to steel products fabrication.

Other processes included in some of the pathways are carbon capture, utilization and storage; methanol; and other chemicals. The From Residual Steel Gases to Methanol (fReSMe) project constructed a pilot plant in Sweden that allows the production of methanol suitable for use as shipping fuel from residual blast furnace gases ("From Residual Steel Gases to Methanol" 2021), which have undergone the water gas shift reaction to have higher CO_2 and H_2 content (Bonalumi et al. 2018). The Carbon2Chem project in Germany also uses coke oven gases and blast furnace gases to produce base chemicals for fertilizer, plastic, and fuel production via an industry coordination approach, with the aim of eliminating all CO_2 emissions from iron and steelmaking (Deerberg, Oles, and Schlögl 2018), although because fuels such as methanol would be produced, their combustion would still produce emissions. The eForFuel project, which seeks to make fuels from CO_2 via biotechnology and electricity, uses blast furnace gases as an important feedstock (Bar-Even, Keller, and Rettenmaier 2021).

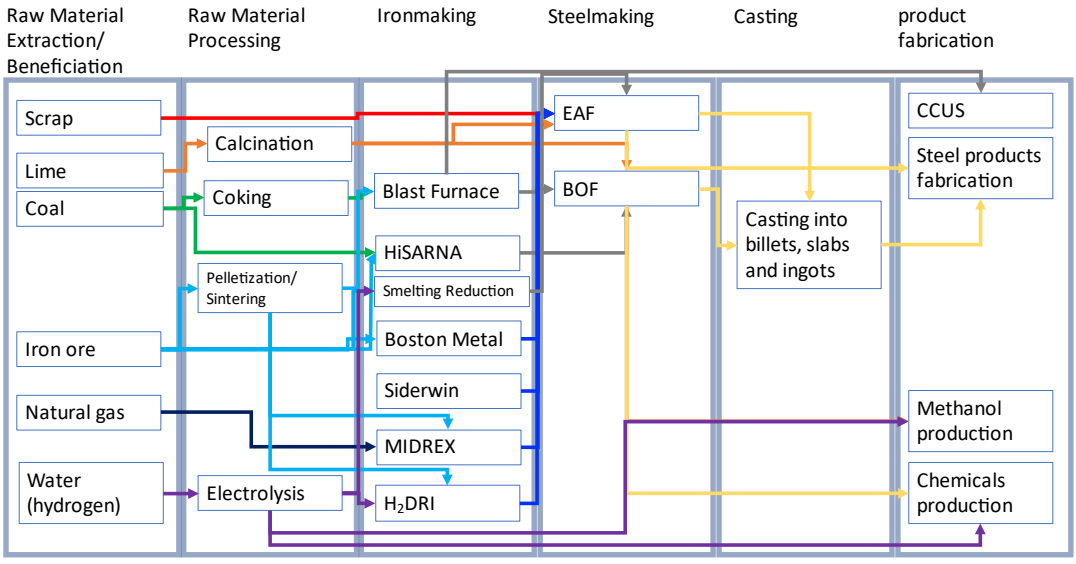


Figure 8. Process pathways for steel production

[13] Natural gas is used as a feedstock for chemical production in this process.

The definition of taxa can be ambiguous even in biology. For industry, various units could be considered taxa, for example:

- Individual industrial plants
- Individual production units in a single plant
- Types of production processes
- Idealized production processes
- Companies
- Products
- Industries.

We used the following idealized production processes as the taxa used for the cladogram construction:

- BF - BOF
- BF - BOF - CCUS
- Scrap - EAF
- DRI & BF - BOF
- DRI & Scrap - EAF
- SIDERWIN
- H2DRI
- Boston Metal
- HIsarna
- Nucor Castrip
- SuSteel (Power to Hydrogen)
- Power-to-iron/organic chemical production
- power to methanol - Carbon2Chem
- power to methanol - FReSMe
- eForFuel

3.2.1 Derivation of the Character Matrix

External pressure and coordinated resources toward combating climate change have assisted in the development of new ironmaking processes and steelmaking improvements that may eventually replace current processes, even though the particular process mixture is not yet clear.

Such external pressure—from either a broad energy transition or a more narrow push in a particular state to avoid climate change—is acting on many industrial sectors and providing impetus for innovation. Drivers of innovation (Schumpeter and Elliott 2012) as applied to the case of the renewable energy transition in the iron and steel sector are shown in Table 2.

Table 2. Drivers of Innovation in the Iron and Steel Sector

These are used to derive potential characters for phylogenetic classification.

Category of Drivers	Examples
Products	• HBI • Zinc (coproduct in HYBRIT) • Coproduct methanol and other chemicals • Byproduct gases to methane
Production methods	• High-temperature BF to BOF • EAF • CCUS (reduce CO_2 emissions) • increasing proportion of EAF & scrap reuse • Scrap in BOF • H_2DRI (e.g., HYBRIT) • Electrowinning at low temperatures (SIDERWIN) • Molten electrolysis • Avoidance of preprocessing • Allowance of use of lower-grade coal—not coke (HIsarna) • DRI with natural gas • DRI with coal
Markets	• California standards • Wind turbines • Markets for lower-carbon products
Supply sources	• Electricity • Electricity from renewables • Green hydrogen • Coal (without need for coke production) • Different grades of iron ore • Natural gas (rather than coal)
Organization structures	• Collaboration to spread cost of risk among a variety of stakeholders including steel companies and association groups, research groups • Larger mills • Mini-mills

The categories in Table 2 are used to derive characters, the technical term describing evolutionary characteristics useful for deriving the cladogram. Thus, we have five character "types" in our analysis. When the energy transition is viewed as a driving force for evolutionary change, some of the character types are more closely studied than others. In this case, processes and supply sources seem to be key drivers. This is because the processes typically used for steel production involve direct combustion of coke, a product derived from coal, which involves mining and energy intensive preprocessing, as well as large CO_2 emissions. Therefore, any processes that reduce mining and preprocessing, or can substitute for the direct combustion (particularly via power-to-X, thereby switching supply characters as well), will help reduce CO_2 emissions. Since across

categories, some of the examples turn out to be duplicative, the final list of characters differs from these examples (for example EAF and mini-mills are often used interchangeably, so only one could be used in the final matrix).[14]

The characters and taxa together define a character matrix, with characters as rows and taxa as columns. Because idealized processes were used as taxa, evaluating process-based characters was straightforward. Characters related to products, however, were less clear-cut. For example, most continuous cast products (slabs, billets and ingots) are not specific to any particular process.

Differences in products are not easily tied to the process used to make steel. Though it is true most rebar comes from EAFs, the product mix from BF-BOF versus EAF is not constant or exclusive. Historically, after EAFs entered the market, BF-BOF firms began to specialize in higher-value products (Christensen 2011). EAFs, however, continued to work to improve quality so that they could also produce those products (Christensen 2011). Therefore, product differentiation is complex. Likewise, for market-based characters, data tying production processes to different types of markets (e.g., machinery used for renewables or fossil energy production) is scarce.

Character values can be ambiguous for other reasons. For example, no process in operation is carbon-neutral, and even the planned processes may not be carbon-neutral if Scope 3 emissions are included. Therefore, there is some subjectivity. The carbon-neutral character is used to show whether a process aspires to or is compatible with full carbon neutrality. In the case of EAF, for example, if the grid is decarbonized, the process could be almost entirely carbon-neutral. In EAFs, chemical reactions do occur, including reduction reactions with carbon. Though we assume they are negligible, this is not always the case, as the reactions provide a very convenient source of heat, so frequently the furnaces are injected with oxygen to take advantage of the chemical energy. Likewise, we assume all DRI—except in the case of an experimental setup such as the H_2DRI processes—requires natural gas. For processes that include DRI, we assume natural gas is a major part of the process, comprising a large part of the feed. Therefore, an iron product is created before the steel product in the processes that specifically mention DRI.

BFs, BOFs, and EAFs operate at very high temperatures. Smelting reduction and molten electrolysis (e.g., Boston Metal) also operate at very high temperatures, with fully liquefied metal. Direct reduction operates at a lower-temperature threshold as the metal is not fully liquefied. Electrolysis in alkaline environments requires a temperature of 110°C, which is much lower than other processes. The high-temperature character refers to the ironmaking process. For example, in the case of power-to-iron/organic chemical production, the character value is 0 because a lower-temperature process could be used. All characters used along with the values and key are listed in Appendices A–E. The character matrix with scores for each taxon and character is shown in Appendix F.

3.3 Iron and Steel Cladogram: Results and Discussion

When we ran the heuristic version of maximum parsimony, we found two alternative trees with an identical minimum score. The final cladogram, shown in Figure 9, represents a strict

[14] Character descriptions and final character/taxa matrix are provided in the appendices.

consensus tree, where only groupings found in both trees are included. When the trees do not all agree on a division, more than two processes branch at a single point; this case is called a polytomy. Polytomies can also occur if no clear hierarchy is found at the division point. At the top of the cladogram, the BF-BOF is a reference point that is used to root the tree. Other processes are shown as diverging from this root process.

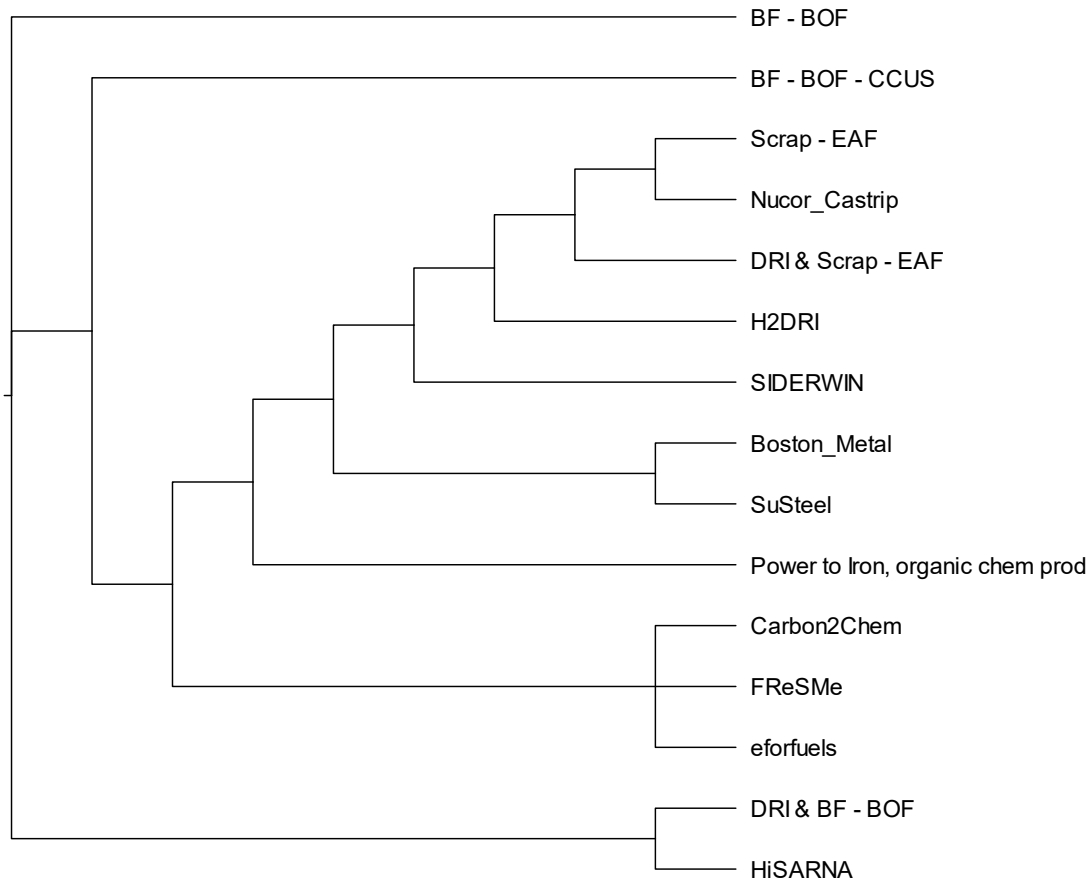


Figure 9. Cladogram of the iron and steel sector

As can be seen in the figure, distinct groupings of processes emerge. In the cladogram, areas where much development and innovation seems to be taking place appear as tightly branched processes. At the first division point, DRI mixed with BF-BOF and HIsarna are grouped together, diverging from the other processes. Next, BF-BOF with CCUS diverges from the rest of the group. The three processes focused on chemical production using blast furnace gases and coke oven gases form a polytomy in the next branching. The following branch shows the power-to-iron/organic chemical production. Then, Boston Metal and SuSteel are together in a branch, both employing electric technologies at high temperature. SIDERWIN, the other electrolytic process, is on a separate branch. The next three branches all include EAF based processes, first H_2DRI, then DRI & scrap to EAF, and finally Nucor Castrip with Scrap-EAF.

The cladogram can reflect the relative carbon intensity of the processes, with DRI & BF-BOF and HIsarna both representing processes with moderate potential for improvement, while other processes offer more potential for emissions decreases. The EAF-based processes are grouped together. EAF does not have the waste gas coproducts that allow production of chemicals.

The BF-BOF process with carbon capture makes its own stand-alone branch in the low-carbon section of the tree, with proximity to both the DRI & BF-BOF/HIsarna branch as well as the processes that utilize the waste gas to produce chemicals.

The cladogram highlights many different threads of innovation occurring in the iron and steel sector. The bottom of the cladogram shows two market-ready technologies, combining more DRI with the standard BF-BOF route and HIsarna, both of which offer modest improvements in terms of greenhouse gas emissions versus the dominant technology. The increase in DRI as a portion of market share in the United States is compatible with such a pathway.

Larger decreases in carbon intensity require moving towards the middle of the cladogram. There, several processes that make use of the carbon-containing waste streams are under development. Alternatively, four processes that rely on electrification of the process, via electrolytic production in the case of SIDERWIN and Boston Metal, or via plasma in the case of SuSteel, and finally electrolytically produced hydrogen in the case of H_2DRI. The other group of processes are focused on reduced use of ore based metallics, primarily recycling scrap via EAFs. The cladogram is a partial representation of the sector because there have been other innovations as well in terms of EAF design.

Therefore, the cladogram reveals a business-as-usual scenario, based on the BF-BOF and the EAFs. A scenario with minor reductions in greenhouse gas emissions would focus on the lowest branch, with increased implementation of DRI & BF – BOF as well as HIsarna, (the choice would likely be based on the location since most DRI plants use natural gas). A scenario focused on capturing carbon from fossil fuels used in reduction (the power-to-iron process does not use fossil fuels for reduction but rather as the feedstock for chemical production that is coupled with the electrolytic iron production in a single plant) for use or storage would look at the BF-BOF-CCUS as well as the closely related Carbon2Chem, FReSMe, and eforfuel as well as power-to-iron, organic chemical production. These technologies vary quite a bit in terms of their costs, feedstocks (power-to-iron avoids the need for coal) and associated products. Their emissions should also vary, but each has the potential to significantly decrease emissions, although the use of carbon may simply delay emissions in some cases.

A scenario based on electrification may consider the electrolytic reduction processes as well as the use of hydrogen plasma and the electrolytic production of hydrogen. Costs for these processes vary, and emissions decreases depend on the composition of the power generation resources. Finally, EAFs already have a much lower carbon intensity than traditional routes, and they continue to be studied for improvements. An increase in the proportion of steelmaking from EAFs may be considered along with the closely linked process pathways, such as innovations surrounding casting as well as an increased proportion of ore-based metallics as feedstock.

Therefore, examining the iron and steel cladogram for the closest relationships allows distinguishing six main groupings:

1. Traditional BF-BOF
2. “Modified” BF-BOF (changes in preprocessing, process changes, input changes)
    - DRI & BF – BOF

- HIsarna

3. Carbon capture, use and/or sequestration pathways (mostly still BF-BOF, chemicals produced include methanol, formic acid, BTX, methane and others; CO_2 could also be stored or used directly (e.g. carbonation, enhanced oil recovery)
    - Carbon2Chem
    - FReSMe
    - eForFuel
4. Electrolysis/carbon use
    - Power-to-iron/organic chemical production
5. Electrification
    A. High temperature electric process (electrolysis, plasma-based reduction, hydrogen reductant)
    - Boston Metal
    - SuSteel

    B. Low temperature electrolytic process
    - SIDERWIN

    C. Electrolytic hydrogen as reducing agent for DRI
    - H_2DRI
6. Scrap-based EAF routes
    - DRI & Scrap - EAF
    - Scrap – EAF
    - Nucor Castrip

Groupings 1-5 are focused on ore-based metallics, while grouping 6 relies primarily on scrap. Existing and increased demand for BF-BOF products may be met by movement or new entry by firms into any of groups 2-6. EAFs may be disrupted by these innovations that reduce the carbon intensity of steel production from ore-based metallics. Movement of BF-BOFs to group 2 represents opportunities for small improvements in CO_2 intensity, with an advantage that these technologies are already market ready. Cleveland Cliffs highlights increased use of HBI in its integrated mills in its 2021 sustainability report (Cleveland-Cliffs Inc. 2022). US Steel highlights capability for injecting natural gas in the BOF (U. S. Steel 2022).

Movement by existing BF-BOF producers towards many of the pathways in group 3 should also be possible; as group 3 and 4 technologies continue to rely on fossil fuels as a feedstock, some existing infrastructure outside the steel plant would still be usable. Group 3 and 4 technologies allow larger reductions in emissions intensity but are less developed than group 2. The pathways in group 5 are also at a low TRL and are quite different from existing processes, so may be less suitable for retrofitting but may be able to meet demand for BF-BOF products once they mature. They avoid the use of fossil fuels as feedstock but increase demand for electricity.

Movement to group 6, which represents the dominant steel production pathway in the United States, can reduce CO_2 intensity, but employs a substantially different feedstock and technology than the BF-BOF route. This means that the final product is also not a perfect replacement. Thus there is an upper limit for the amount of shift towards EAFs, which also reflects scrap availability or contamination of scrap resources (Cooper et al. 2020). Still, US Steel recently decided not to go forward with upgrades to its integrated mills at Mon Valley Works (Deaux 2022) and highlighted the acquisition of Big River Steel (a LEED certified plant with an EAF), the addition of an EAF at the Fairfield site (formerly the site of an integrated mill) as well as a "green" steel product called verdeX™ that consists of up to 90% recycled steel made at Big River Steel in its sustainability report (U. S. Steel 2022). The Big River Steel site uses high-quality scrap, so the products made are similar to those made in integrated mills. Cleveland-Cliffs acquired Ferrous Processing Technologies in 2021, giving them EAFs throughout the United States despite their primary focus on iron ore and ore-based metallics. The company mentions in their 2021 sustainability report that all their steel contains scrap (Cleveland-Cliffs Inc. 2022).

While cladograms are not yet a fully mature method for industrial energy modeling, this exercise demonstrates the information and insights that can be gleaned from the process of their construction. Still, to understand the energy use associated with industries, facility level data is also needed. This need is discussed in the following section.

4 Data for Classification

In this section, we validate an approach for estimating energy use from reported GHG emissions data (McMillan et al. 2016; McMillan and Ruth 2019), which helps overcome limitations of data availability for fuel use at the facility level, and for industry energy estimates in years when MECS is not available. In addition to analyses developed by the National Renewable Energy Laboratory, energy estimates developed using this approach have been incorporated into a data platform for state and local energy estimates by the U.S. Department of Energy ("SLOPE: State and Local Planning for Energy," n.d.) and for analysis of industrial heat pumps (Alstone et al. 2021). Section 4.1 describes the dataset used for validation. The first step in the validation is matching the facilities between the datasets, which is described in Section 4.2. Once there are sufficient matching records, the validation analysis is conducted as described in Section 4.3. The results of the validation are provided and discussed in Section 4.4.

4.1 Overview of Validation Data

In MECS, the Energy Information Administration conducts a national survey of manufacturing establishments with paid employees; firms with a single establishment are excluded ("2014 Manufacturing Energy Consumption Survey Methodology and Data Quality: Survey Design, Implementation, and Estimates" n.d.). For the 2014 MECS, sampling was done based on the 2012 Economic Census and modified by the 2013 business register to account for new establishments.[15] A total of 14,900 establishments were sampled, that is, sent a survey form to be completed. Though this represents a small portion of the approximately 170,000 manufacturing establishments found in the 2012 census, 23 industry groupings were always selected ("selected with certainty, including 322121 (Paper Mills Except Newsprint), 322130 (Paperboard Mills), 324110 (Petroleum Refineries) and 331110 (Iron and Steel Mills and Ferroalloy Manufacturing) ("2014 Manufacturing Energy Consumption Survey Methodology and Data Quality: Survey Design, Implementation, and Estimates" n.d.). Other facilities were selected based on the number of facilities in the corresponding NAICS code, level of fuel use in the prior MECS as well as the requirements by the U.S. Energy Information Administration.

Although the establishments may be sampled at a high or complete rate, there is some nonresponse. As large users of fuel are always sampled, it seemed logical that records from facilities reporting to GHGRP would typically match to records included in MECS. The U.S. Energy Information Administration (EIA) does not specify the threshold for large users of fuels, and not all fuel types are included. Nevertheless, published MECS data indicate natural gas (38%) and other (37%) represent most of the fuel use for all purposes, although coal (7), coke and breeze (2%), and hydrocarbon gas liquids (15%) make up more than 1% apiece. Census microdata are not publicly available. Researchers must pass through a vetting process to obtain Special Sworn Status to obtain access to Federal Statistical Research Data Centers (United States Census Bureau 2022). Any data that are released must be approved through a disclosure process (see United States Census Bureau's Center for Economic Studies (2009) for a more thorough description of the process).

[15] The 2018 MECS was not available at the time.

4.2 Matching

The first stage of the work involved matching data from the Business Register (United States Census Bureau n.d.) to MECS, and then from MECS to the GHGRP facilities. Matching MECS to GHGRP facilities is a type of unbalanced linear assignment problem. There can be at most a 1:1 matching; not all facilities are expected to have a match. There is a fuzzy component because the entries in the two data sets are not expected to have perfect alignment since the same name or address may be written differently, and NAICS categorization has known variation.

With more than 2,000 manufacturing facilities in the GHGRP system and 14,900 facilities in MECS, many matches are possible for each GHGRP facility. We computed the Cartesian product with the constraint that the state and ZIP code fields were equal. We then performed string-matching on facility names, addresses, and NAICS codes. The Jaro-Winkler distance in the jellyfish library for Python (Turk 2021) was computed for all possible matches between those pairs of items. The Jaro-Winkler distance gives higher priority to the earlier part of string values. To emphasize the first part of the strings even more, we truncated the strings for address and facility names to 10 digits. A threshold value of two was used for initial matching. Still, because facilities retain the same ID number even if the ownership changes, location and NAICS codes are more important. Thus, we also manually inspected the matches before checking them against known criteria (e.g., NAICS code and high fuel use) that would predict near-perfect inclusion in MECS.

Fuzzy matching was chosen because while some consistency is expected between the two data sets, perfect consistency is unlikely. Facilities in the GHGRP data set have records for 2010–2019. Iron and steel production was assigned the NAICS Code 331110, which is one of the codes MECS tries to sample completely. The GHGRP data set has many facilities in Codes 331111 and 331112, which were valid in 2007 but not in the 2012 system. Therefore, even NAICS codes, which seem to be the most straightforward of the data pieces, are not expected to be exact matches. Likewise, address and names can be written differently and can even be somewhat ambiguous for the person writing them; some examples are physical versus mailing addresses, a facility that is not very close to a street or which has multiple entrances, name changes in roads, a facility name versus a company name, abbreviations versus names or addresses written out, and a name of a parent company versus a subsidiary.

4.3 Analysis

For the analysis portion, we restricted the GHGRP data set to manufacturing facilities reporting in 2014. This reduced the data set to 2,253 facilities. Table 3 shows the aggregation of the fuel types estimated by the GHGRP procedure next to the fuel types reported via Form EIA-846. Conversion factors provided in Form EIA-846 were used to convert the physical quantities to energy units (MMBtu). As Table 4 indicates, a higher level of disaggregation is possible for MECS fuel types. Where there was disagreement on NAICS code, the MECS NAICS code was used for any analysis. Once the facilities were matched, the log difference of fuel use by facilities in each data set was taken. A translation was used, adding 1 BTU to both sides (to enable calculation of finite log values in all cases). Cement facilities (NAICS Code 327310) were excluded from the analysis, as the GHGRP-based fuel use estimation methodology for cement facilities has known issues (see McMillan and Narwade 2018). As this was a preliminary study, we restricted results to total fuel.

Table 3. Concordance of Fuel Types from the GHGRP Data Set and Form EIA-846

GHGRP Aggregation	MECS (Form EIA-846)
Biomass	• Wood harvested directly from trees • Pulp and black liquor • Agricultural waste • Wood residues and byproducts from mill processing • Wood and paper-related refuse
Blast furnace/coke oven gas	• Blast furnace gas • Coke oven gas
Coal	• Anthracite • Bituminous and subbituminous coal • Lignite
Coke & breeze	• Breeze • Coal coke
Diesel	• Diesel fuel oil
Liquefied petroleum gas and natural gas liquids	• Butane • Ethane • Propane • Mixtures of butane ethane and propane • Other liquefied petroleum gas and natural gas liquids, including butylene • Ethylene and propylene
Natural Gas	• Natural gas
Other	• Distillate fuel oil • Kerosene • Motor gasoline • Naphtha and heavier gas oils • Bitumen • Acetylene • Hydrogen • Steam • Industrial hot water
Petroleum Coke	• Marketable petroleum coke—unrefined or green • Marketable petroleum coke—calcined
Residual Fuel Oil	• Residual fuel oil
Waste Gas	• Waste and byproduct gases
Waste oils tars waste materials	• Waste oils and tars • Tire-derived fuel

4.4 Results and Discussion

Our results are summarized in Table 4, Table 5, and Figure 10. The tests were run on the facilities that we were able to match between the GHGRP and MECS records. Therefore, the total facilities designation corresponds to that grouping. Energy use estimates from the GHGRP-based method were within +/- 25% of MECS estimates in over 70% of matched facilities[16] (Table 4). Facilities within these bounds accounted for close to half of the energy use reported in matched MECS facilities. To understand the difference between the two data sets, we used the Wilcoxon-signed rank test. We choose it because the estimates correspond to the same facilities and the same year, so are paired rather than independent. It is a nonparametric test, so the distribution shape is not assumed. The hypotheses tested were that (1) the mean of the two samples were equal and (2) either MECS or GHGRP were higher. The results mean the higher MECS estimates are statistically significant, as shown in Table 5.

Table 4. Matched Facilities Whose Fuel Use Estimates are Within +/- 25% Represent 49.6% of the MECS Fuel Reported for the Matched Facilities and 72.6% of the Total Number of Matched Facilities

Percentage Bound	% MECS Total Fuel	% Matched Facilities, all Fuels
25	49.60%	72.60%

Table 5. Results from the Wilcoxon-Signed-Rank Test Showing the MECS Estimates Trend Higher than GHGRP-Based Estimates

Hypothesis	Data	P-Value
Diff = 0	All fuels	3.329E-66
MECS > GHGRP	All fuels	1
GHGRP > MECS	All fuels	1.664E-66

The kernel density estimate (KDE) is also nonparametric. In KDE, a kernel function (in this case, the gaussian kernel), is applied at each data point and then the functions are summed. This is a probability density function, so its integral sums to 1. The KDE of the log difference between the fuel estimates according to MECS versus the GHGRP method is shown in Figure 10. The higher bias of the MECS values is shown, as is the right-skew of the log difference. This shows visually that while MECS estimates are generally higher than GHGRP estimates, in most cases the estimates are quite similar. This provides some justification for using GHGRP data to estimate fuel use but shows that there are more outlying cases with higher MECS than GHGRP values than cases with higher GHGRP values. Due to the limited nature of this study, we cannot determine why MECS fuel use is higher than the GHGRP estimates or why many cases fell outside of the bounds. We recommend further study to better understand the discrepancies.

[16] We cannot disclose the total percentage of facilities that were matched.

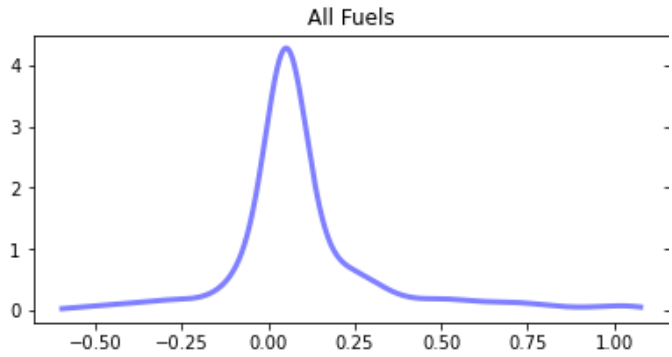


Figure 10. Kernel density estimation (KDE) of the log difference between the estimates of fuel use by MECS versus the GHGRP-based method (log(MECS fuel use) – log (GHGRP-based fuel use))

5 Conclusions

5.1 Implications for Energy Analysis and Modeling

In 2020, industry accounted for 36% of the total end uses for energy, more than any other sector (EIA 2021). Almost 80% of the energy use by industry involves direct use of fossil fuels, particularly via combustion processes (EIA 2018). For the coming decades, industrial decarbonization is a key goal (Jean-Pierre 2022; Rissman et al. 2020; Bataille et al. 2018). But modeling and projecting industrial energy use is challenging due to heterogeneity in recognized sectors (Greening, Boyd, and Roop 2007). Differences in energy use are not well-captured by NAICS, and hence energy use projections based on NAICS have high uncertainty.

Clustering IAC data showed that resulting taxonomies are more explanatory for energy use than NAICS. The clusters always include facilities from many NAICS codes. This work shows clusters can be found that show similar levels of correlation between energy use and variables such as number of employees, sales, and plant area (hours shows less correlation), as do three-digit NAICS codes. These clusters cut across a swath of NAICS codes, and they can be defined at a broader level (i.e., fewer categories or clusters are needed for the same level of correlation). The clusters can be distinguished at higher than 80% accuracy using support vector machines (using a training set of 90% and test set of 10% of the values).

Every cluster in our analysis is drawn from a range of NAICS codes, even at the three-digit level, lending additional support to the contention that NAICS does not seem to be a good indicator of how energy use correlates with the variables in the data set (number of employees, hours, facility area, and sales). Still, the IAC data has some weaknesses, as it focuses on smaller and medium-sized facilities and is biased toward higher energy users for a given facility type. These weaknesses could be overcome by a larger-scale study with more representative data. One weakness of the phenetic approach is its exclusive focus on energy use and other high-level statistics rather than process characteristics, so that causation and facility-level improvements may be difficult to infer. Qualitative approaches for analyzing IAC data did not show easily identifiable patterns between the plants connected in clusters, suggesting machine learning approaches may be useful for determining new classifications of industries that may improve industrial energy analysis and modeling. Our approach could be combined with more study of industries excluded in the data set to designate a group of industry archetypes.

In contrast, the cladogram of iron and steel is based on process components, which are key drivers of energy use. Though the cladogram contains subjective aspects, so is not as unsupervised as the clustering on IAC data, it still reveals patterns that may not be apparent. In that sense, phylogenetic inference is similar to machine learning approaches.

Our prototype approach for constructing an industry cladogram based on process characteristics according to the innovation framework derived from Schumpeter is flexible and captures evolutionary behavior. The resulting cladogram represents a snapshot in time based on detailed study of process characteristics. A key application for this work could be in designing scenarios for more detailed modeling, as different branches represent alternative pathways. Cladograms may be particularly useful for scenario-based modeling because they can be used to sense and compare groupings of emerging or dominant processes and their implications. For example, in

steel production, scenario modeling could be used to see the relationship of sectors that use the emitted carbon for chemical production such as CO_2, organic chemicals, and methanol versus processes that do not use carbon as a reductant and therefore do not produce chemicals or CO_2. The processes are arrayed together making it easy to explore them and understand tradeoffs. Constructing a cladogram could be a good first step to analysis of many industries (e.g. nitrogenous fertilizer production, ethyl alcohol manufacturing), to understand their heterogeneity, emerging trends, and coherent groupings of related innovations. Much attention is sometimes given to a small number of processes, but the cladogram helps show related threads of innovation in a way that may be helpful for policymakers and entrepreneurs, allowing them to see the larger picture, other good ideas, or competitors.

Because cladistics has not been widely applied to industrial systems, some methodological issues must still be worked out. In particular, because the approach is more widely used in biology, some software tools need further development to support best practices for cladograms that rely on morphological characters rather than molecular evidence. If expanded more generally to process technologies that impact a range of facilities, cladistics could uncover trends and groupings that allow wider analysis throughout industry. It could also be possible to apply this approach to capture a wider portion of industries and identify innovations shared across them, and cladistics could thus have wide applicability and potential for industrial modeling.

Facility-level data availability continues to be a major challenge for industrial modeling and for developing industrial modeling approaches that do not rely on NAICS classification. The method outlined by (McMillan et al. 2016; McMillan and Ruth 2019) could be a key step toward a solution, as it makes facility-level estimates available for industries that are major sources of GHG emissions. Until now, however, these energy estimates had not been validated, and the validation here provides an important step for further use of estimated facility-level energy use.

5.2 Additional Research

This work uncovered several promising avenues for future work. The validation of the method outlined by McMillan et al. (McMillan et al. 2016; McMillan and Ruth 2019) supports a broader use of the GHGRP data set to provide facility-level estimates, such as providing between-MECS-year estimates. Because MECS is currently conducted every 4 years, important annual changes in energy use may occur that are missed by the MECS survey schedule. The approach that has been validated here can be used to create time series of annual data that may be useful in analyzing trends in relationships between economic activity and energy use, for example. The validation results also raise questions about the areas that show a mismatch between the MECS and GHGRP data, and these questions could be explored and used to improve energy estimates based on the GHGRP and, potentially, MECS. Due to the limited scope of this project, we could not fully investigate the discrepancies.

The GHGRP-based energy estimates could also be used for additional exploration of taxonomies and thus provide a complement to the analysis of the smaller, less energy-intensive IAC facilities. The cladistics methodology is promising for showing the evolutionary groupings in industry. However, it needs further development and expansion before it can be used to look at additional process details (e.g., furnaces or machinery) that could help identify wider trends in energy use throughout industry. Perhaps it is possible to construct cladograms of types of furnace or boilers to see how market-based characters affect the patterns that emerge. For industries whose evolution is

of particular interest due to concerns about climate change, such as iron and steel, cement or fertilizers, cladograms could be developed as done in this work.

6 References

"2014 Manufacturing Energy Consumption Survey Methodology and Data Quality: Survey Design, Implementation, and Estimates." n.d. US Energy Information Administration. Accessed May 23, 2022. https://www.eia.gov/consumption/manufacturing/data/2014/index.php?view=methodology_2014.

"2021 World Steel in Figures." 2021. worldsteel Association. https://worldsteel.org/publications/bookshop/world-steel-in-figures-2021/.

Alstone, Peter, Evan Mills, Jerome Carman, and Alejandro Cervantes. 2021. "Toward Carbon-Free Hot Water and Industrial Heat with Efficient and Flexible Heat Pumps." Arcata, CA: Schatz Energy Research Center. http://schatzcenter.org/pubs/2021-heatpumps-R1.pdf.

Bailera, Manuel, Pilar Lisbona, Begoña Peña, and Luis M. Romeo. 2021. "A Review on CO2 Mitigation in the Iron and Steel Industry through Power to X Processes." *Journal of CO2 Utilization* 46 (April): 101456. https://doi.org/10.1016/j.jcou.2021.101456.

Baldwin, James, Christen Rose-Anderssen, and Keith Ridgway. 2012. "Evolving Manufacturing Systems: Hierarchical and Cladistic Classifications." In . Rotterdam. https://www.researchgate.net/profile/James-Baldwin-7/publication/248386882_Evolving_Manufacturing_Systems_Hierarchical_and_Cladistic_Classifications/links/0deec51de6cf3bcd17000000/Evolving-Manufacturing-Systems-Hierarchical-and-Cladistic-Classifications.pdf.

Baldwin, James S. 2011. "The Complexity of Industrial Ecosystems: Classification and Computational Modelling." In *The SAGE Handbook of Complexity and Management*, edited by Peter Allen, Steve Maguire, and Bill McKelvey, 299–316. London: SAGE.

Baldwin, James Scott. 2008. "Industrial Ecosystems: An Evolutionary Classification Scheme." *Progress in Industrial Ecology, An International Journal* 5 (4): 277. https://doi.org/10.1504/PIE.2008.021920.

Baldwin, James Scott, Christen Rose-Anderssen, Keith Ridgway, Fabian Boettinger, Marcus Michen, Kwabena Agyapong-Kodua, Ivan Brencsics, Istvan Nemeth, and Roland Krain. 2013. "The Evolution of Manufacturing SPECIES." *Procedia CIRP*, Forty Sixth CIRP Conference on Manufacturing Systems 2013, 7 (January): 187–92. https://doi.org/10.1016/j.procir.2013.05.032.

Bar-Even, Arren, Heiko Keller, and Nils Rettenmaier. 2021. "EForFuel Fuels from Electricity: De Novo Metabolic Conversion of Electrochemically Produced Formate into Hydrocarbons: Deliverable 5.1 Interim Report on Definitions, Settings and System Description." H2020-LCE-2017-RES-RIA. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5c466b64a&appId=PPGMS.

Bataille, Chris, Max Åhman, Karsten Neuhoff, Lars J. Nilsson, Manfred Fischedick, Stefan Lechtenböhmer, Baltazar Solano-Rodriquez, et al. 2018. "A Review of Technology and Policy Deep Decarbonization Pathway Options for Making Energy-Intensive Industry Production Consistent with the Paris Agreement." *Journal of Cleaner Production* 187 (June): 960–73. https://doi.org/10.1016/j.jclepro.2018.03.107.

Bonalumi, D, G Manzolini, J Vente, H.A.J. van Dijk, L. Hooey, and Emeric Sarron. 2018. "From Residual Steel Gases to Methanol: The FReSMe Project." In *Ghgt-14*. Melbourne,

Australia. https://re.public.polimi.it/retrieve/e0c31c0f-19e7-4599-e053-1705fe0aef77/Bonalumi-Fresme_ghgt14.pdf.
Boston Metal. 2019. "Boston Metal Raises $20M to Deliver Industrial-Scale Electrolysis Solution for High-Volume Ferroalloys Production." Boston Metal. January 10, 2019. https://www.bostonmetal.com/news/boston-metal-raises-20m-to-deliver-industrial-scale-electrolysis-solution-for-high-volume-ferroalloys-production/.
"———." n.d. Accessed June 7, 2022. https://www.bostonmetal.com.
"Buy Clean California Act." n.d. Department of General Services, State of California. Accessed June 22, 2022. https://www.dgs.ca.gov/PD/Resources/Page-Content/Procurement-Division-Resources-List-Folder/Buy-Clean-California-Act.
Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. 2018. "Statistics versus Machine Learning." *Nature Methods* 15 (4): 233–34. https://doi.org/10.1038/nmeth.4642.
Christensen, Clayton M. 2011. *The Innovator's Dilemma: The Revolutionary Book That Will Change the Way You Do Business ; [with a New Preface]*. 1. Harper Business paperback publ. New York, NY: Harper Business.
Cleveland-Cliffs Inc. 2022. "Sustainability Report 2021." Cleveland-Cliffs Inc. https://d1io3yog0oux5.cloudfront.net/_cd3f3130bd33f34db5d73099656ad78b/clevelandcliffs/db/1188/11273/file/CLF_Report_Sustainability_2021_SinglePages.pdf.
Cooper, Daniel R., Nicole A. Ryan, Kyle Syndergaard, and Yongxian Zhu. 2020. "The Potential for Material Circularity and Independence in the U.S. Steel Sector." *Journal of Industrial Ecology* 24 (4): 748–62. https://doi.org/10.1111/jiec.12971.
David L. Swofford and Jack Sullivan. 2003. "Phylogeny Inference Based on Parsimony and Other Methods Using Paup*." In *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny*, 160–206. http://www2.ib.unicamp.br/profs/sfreis/SistematicaMolecular/Aula09MetodoParcimonia/Leituras/ThePhylogeneticHandbookParcimonia.pdf.
Deaux, Joe. 2022. "U.S. Steel Bets on a New Technology--and the South--to Survive." *Bloomberg*, May 6, 2022. https://www.bloomberg.com/news/articles/2022-05-06/u-s-steel-jobs-shift-to-arkansas-away-from-pittsburgh.
Deerberg, Görge, Markus Oles, and Robert Schlögl. 2018. "The Project Carbon2Chem®." *Chemie Ingenieur Technik* 90 (10): 1365–68. https://doi.org/10.1002/cite.201800060.
Eurofer. 2021. "Launch of the Clean Steel Partnership Pabes the Way for Further Research and Deployment of Ground-Breaking Technology," June 23, 2021. https://www.eurofer.eu/news/launch-of-the-clean-steel-partnership-paves-the-way-for-further-research-and-deployment-of-ground-breaking-technology/.
European Steel Technology Platform (ESTEP). 2020. "Clean Steel Partnership Roadmap." https://www.estep.eu/assets/Uploads/200715-CSP-Roadmap.pdf.
Executive Office of the President Office of Management and Budget. 2022. "North Americal Industry Classification System: United States, 2022." https://www.census.gov/naics/reference_files_tools/2022_NAICS_Manual.pdf.
Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, Inc.
Forster, Hector. 2021. "BMW Invests in Zero-Emissions Steel Technology Firm Boston Metal." S&P Global Platts. *Eurometal* (blog). March 16, 2021. https://eurometal.net/bmw-invests-in-zero-emissions-steel-technology-firm-boston-metal/.
Frey, Perry A., and George H. Reed. 2012. "The Ubiquity of Iron." *ACS Chemical Biology* 7 (9): 1477–81. https://doi.org/10.1021/cb300323q.

"From Residual Steel Gases to Methanol." 2021. http://www.fresme.eu/index.php#TOP.
Gale, W. K. V. 1990. "Ferrous Metals." In *An Encyclopedia of the History of Technology*, 146–85. London: Routledge.
Gittleman, John. 2016. "Phylogeny." In *Encyclopedia Britannica*. https://www.britannica.com/science/phylogeny.
Greening, Lorna A., Gale Boyd, and Joseph M. Roop. 2007. "Modeling of Industrial Energy Consumption: An Introduction and Context." *Energy Economics* 29 (4): 599–608. https://doi.org/10.1016/j.eneco.2007.02.011.
Hasanbeigi, Ali. 2022. "Steel Climate Impact: An International Benchmarking of Energy and CO2 Intensities." Global Efficiency Intelligence. https://static1.squarespace.com/static/5877e86f9de4bb8bce72105c/t/624ebc5e1f5e2f3078c53a07/1649327229553/Steel+climate+impact-benchmarking+report+7April2022.pdf.
Hennig, Willi. 1965. "Phylogenetic Systematics." *Annual Review of Entomology* 10 (1): 97–116.
"Industrial Assessment Centers." n.d. US Department of Energy. Accessed May 23, 2022. https://iac.university.
James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. Second edition. Springer Texts in Statistics. New York NY: Springer. https://doi.org/10.1007/978-1-0716-1418-1.
Jean-Pierre, Karine. 2022. "Fact Sheet: Biden-Harris Administration Advances Cleaner Industrial Sector to Reduce Emissions and Reinvigorate American Manufacturing." White House. https://www.whitehouse.gov/briefing-room/statements-releases/2022/02/15/fact-sheet-biden-harris-administration-advances-cleaner-industrial-sector-to-reduce-emissions-and-reinvigorate-american-manufacturing/.
K1-Met GmbH. 2022. "Project SuSteel: Sustainable Steel Production Utilising Hydrogen." https://www.k1-met.com/en/non_comet/susteel.
Kanchiralla, Fayas Malik, Noor Jalo, Simon Johnsson, Patrik Thollander, and Maria Andersson. 2020. "Energy End-Use Categorization and Performance Indicators for Energy Management in the Engineering Industry." *Energies* 13 (2): 369. https://doi.org/10.3390/en13020369.
Karp, Richard M. 1972. "Reducibility among Combinatorial Problems." In *Complexity of Computer Computations*, edited by Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, 85–103. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-2001-2_9.
Leseure, Michel J. 2000. "Manufacturing Strategies in the Hand Tool Industry." *International Journal of Operations & Production Management* 20 (12): 1475–87. https://doi.org/10.1108/01443570010353112.
Mandatory Greenhouse Gas Reporting. 2009. *Code of Federal Regulations*. Vol. 40. http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title40/40cfr98_main_02.tpl.
McCarthy, Ian. 1995. "Manufacturing Classification: Lessons from Organizational Systematics and Biological Taxonomy." *Integrated Manufacturing Systems* 6 (6): 37–48. https://doi.org/10.1108/09576069510099365.
McCarthy, Ian, Michel Leseure, Keith Ridgway, and Nick Fieller. 1997. "Building a Manufacturing Cladogram." *International Journal of Technology Management* 13 (3): 269. https://doi.org/10.1504/IJTM.1997.001664.
McCarthy, Ian, and Keith Ridgway. 2000. "Cladistics: A Taxonomy for Manufacturing Organizations." *Integrated Manufacturing Systems*.

McMillan, Colin A., Richard Boardman, Michael McKellar, Piyush Sabharwall, Mark Ruth, and Shannon Bragg-Sitton. 2016. "Generation and Use of Thermal Energy in the U.S. Industrial Sector and Opportunities to Reduce Its Carbon Emissions." INL/EXT--16-39680, NREL/TP--6A50-66763, 1334495. https://doi.org/10.2172/1334495.

McMillan, Colin A., and Vinayak Narwade. 2018. "The Industry Energy Tool (IET): Documentation." NREL/TP-6A20-71990. Golden, CO: National Renewable Energy Lab (NREL). https://doi.org/10.2172/1484348.

McMillan, Colin A., and Mark Ruth. 2019. "Using Facility-Level Emissions Data to Estimate the Technical Potential of Alternative Thermal Sources to Meet Industrial Heat Demand." *Applied Energy* 239 (April): 1077–90. https://doi.org/10.1016/j.apenergy.2019.01.077.

McMillan, Colin, Richard Boardman, Michael McKellar, Piyush Sabharwall, Mark Ruth, and Shannon Bragg-Sitton. 2016. "Generation and Use of Thermal Energy in the U.S. Industrial Sector and Opportunities to Reduce Its Carbon Emissions." NREL/TP--6A50-66763, INL/EXT--16-39680, 1335587. https://doi.org/10.2172/1335587.

Muller, Michael. 2011. "IAC Assessment Database Manual." Version 10.2. US Department of Energy/Rutgers University. https://iac.university/technicalDocs/IAC_DatabaseManualv10.2.pdf.

Nakhleh, Luay, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. "A Comparison of Phylogenetic Reconstruction Methods on an Indo-European Dataset." *Transactions of the Philological Society* 103 (2): 171–92. https://doi.org/10.1111/j.1467-968X.2005.00149.x.

Naseri Seftejani, Masab, Johannes Schenk, Daniel Spreitzer, and Michael Andreas Zarl. 2020. "Slag Formation during Reduction of Iron Oxide Using Hydrogen Plasma Smelting Reduction." *Materials* 13 (4): 935. https://doi.org/10.3390/ma13040935.

O'Brien, Michael J., John Darwent, and R.Lee Lyman. 2001. "Cladistics Is Useful for Reconstructing Archaeological Phylogenies: Palaeoindian Points from the Southeastern United States." *Journal of Archaeological Science* 28 (10): 1115–36. https://doi.org/10.1006/jasc.2001.0681.

Parkinson, B., C. Greig, E. McFarland, and S. Smart. 2017. "Techno-Economic Analysis of a Process for CO2-Free Coproduction of Iron and Hydrocarbon Chemical Products." *Chemical Engineering Journal* 313 (April): 136–43. https://doi.org/10.1016/j.cej.2016.12.059.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, et al. 2011. "Scikit-Learn: Machine Learing in Python." *The Journal of Machine Learning Research* 12 (November): 2825–30.

Raza, Umair, Wasim Ahmad, and Atif Khan. 2018. "Transformation from Manufacturing Process Taxonomy to Repair Process Taxonomy: A Phenetic Approach." *Journal of Industrial Engineering International* 14 (2): 415–28. https://doi.org/10.1007/s40092-017-0232-8.

Rip, Arie, and René Kemp. 1998. "Technological Change." In *Human Choice and Climate Change: Vol. II, Resources and Technology*, 327–99. Battelle Press. https://kemp.unu-merit.nl/Rip%20and%20Kemp.pdf.

Rissman, Jeffrey, Chris Bataille, Eric Masanet, Nate Aden, William R. Morrow, Nan Zhou, Neal Elliott, et al. 2020. "Technologies and Policies to Decarbonize Global Industry: Review and Assessment of Mitigation Drivers through 2070." *Applied Energy* 266 (May): 114848. https://doi.org/10.1016/j.apenergy.2020.114848.

Rose-Anderssen, Christen. 2014. “Methodologies for Practical Applications of Linnaean and Cladistic Classification of Production Systems.” PhD, Sheffield, United Kingdom: University of Sheffield. https://etheses.whiterose.ac.uk/5967/.
Rose-Anderssen, Christen, James Baldwin, Keith Ridgway, Peter Allen, Liz Varga, and Mark Strathern. 2009. “A Cladistic Classification of Commercial Aerospace Supply Chain Evolution.” *Journal of Manufacturing Technology Management* 20 (2): 235–57. https://doi.org/10.1108/17410380910929646.
Schrago, Carlos G., Barbara O. Aguiar, and Beatriz Mello. 2018. “Comparative Evaluation of Maximum Parsimony and Bayesian Phylogenetic Reconstruction Using Empirical Morphological Data.” *Journal of Evolutionary Biology* 31 (10): 1477–84. https://doi.org/10.1111/jeb.13344.
Schumpeter, Joseph A., and John E. Elliott. 2012. *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*. Translated by Redvers Opie. Sixteenth printing; New material this edition copyright 1983, Original material copyright 1934. Social Science Classics Series. New Brunswick (U.S.A.) London (U.K.): Transaction Publishers.
Simons, Peter. 2013. “Vague Kinds and Biological Nominalism.” *Metaphysica* 14 (2): 275–82. https://doi.org/10.1007/s12133-013-0127-0.
“SLOPE: State and Local Planning for Energy.” n.d. National Renewable Energy Laboratory (NREL). https://maps.nrel.gov/slope.
Smith, Adrian, Andy Stirling, and Frans Berkhout. 2005. “The Governance of Sustainable Socio-Technical Transitions.” *Research Policy* 34 (10): 1491–1510. https://doi.org/10.1016/j.respol.2005.07.005.
SSAB. n.d. “HYBRIT: Fossil-Free Steel.” Accessed June 6, 2022. https://www.hybritdevelopment.se/en/.
Swofford, D.L. 2003. “PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).” Sinauer Associates, Sunderland, Massachusetts.
TATA Steel. 2020. “HISARNA: Building a Sustainable Steel Industry Factsheet.” https://www.tatasteeleurope.com/sites/default/files/tata-steel-europe-factsheet-hisarna.pdf.
Tecnalia. n.d. “Siderwin: Development of New MethodologieS for InDustrial CO2-FreE Steel PRoduction by ElectroWINning.” Accessed June 7, 2022. https://www.siderwin-spire.eu.
Tuck, Christopher Candice. 2021. “Iron and Steel.” Mineral Commodity Summaries. US Geological Survey. https://pubs.usgs.gov/periodicals/mcs2021/mcs2021-iron-steel.pdf.
Turk, James. 2021. “Jellyfish.” https://jamesturk.github.io/jellyfish/.
U. S. Steel. 2022. “2021 Sustainability Report.” U. S. Steel. https://www.ussteel.com/documents/40705/43725/USS_CSR21_Full_Report.pdf/b990344d-2d81-f112-2a8f-b1c584989345?t=1658343821405.
United States Census Bureau. 2022. “Federal Statistical Research Data Centers.” https://www.census.gov/about/adrm/fsrdc.html.
———. n.d. “Business Register.” Accessed September 2, 2022. https://www.census.gov/econ/overview/mu0600.html.
United States Census Bureau’s Center for Economic Studies. 2009. “The Researcher Handbook: U.S. Bureau of the Census Center for Economic Studies Research Data Center’s Handbook for Researchers.” https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/Researcher_Handbook_20091119.pdf.

US Energy Information Administration. 2018. “2018 Manufacturing Energy Consumption Survey (MECS).” https://www.eia.gov/consumption/manufacturing/.
———. 2021. “US Energy Consumption by Source and Sector, 2020.” https://www.eia.gov/energyexplained/us-energy-facts/images/consumption-by-source-and-sector.pdf.
USGS. 2022. “Advance Data Release of the 2020 Annual Tables: Iron and Steel.”
Verwiebe, Paul Anton, Stephan Seim, Simon Burges, Lennart Schulz, and Joachim Müller-Kirchenbauer. 2021. “Modeling Energy Demand—A Systematic Literature Review.” *Energies* 14 (23): 7859. https://doi.org/10.3390/en14237859.
Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72. https://doi.org/10.1038/s41592-019-0686-2.
Wachs, Liz, and Colin McMillan. 2021. “One of These Things IS Like the Other: Pursuing a New Taxonomy of Industry for Improved Energy System Modeling.” In . Virtual. https://www.nrel.gov/docs/fy21osti/80141.pdf.
Warnow, Tandy. 2017. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781316882313.
worldsteel Association. n.d. “Our Performance.” Accessed June 23, 2022. https://worldsteel.org/steel-topics/sustainability/sustainability-indicators/.

Appendix A. Product-Related Characters

Table A-1. Characters related to products

No.	Character	No.	States
1	Iron product	0 1	None Iron product
2	Coproduct	0 1 2 3	None Methanol Methanol and other chemicals Other chemicals

Appendix B. Process-Related Characters

Table B-1. Characters related to process attributes

No.	Character	No.	States
1	Ironmaking	0 1 2 3	None Pig iron Pig iron and sponge iron/DRI/HBI Sponge iron/DRI/HBI
2	Power to X	0 1	Not electrolytic process Electrolysis based process
3	Electrolytic iron production Process	0 1	No electrolysis Electrolytic iron reduction
4	Carbon reductant	0 1	Carbon used as reducing agent Carbon not used as reducing agent
5	Preprocessing of coal required	0 1 2	No Some Yes
6	Pelletization of iron ore required	0 1	No Yes
7	High temperature	0 1 2	Lowest range Medium range Highest range
8	Carbon-neutral	0 1	No Yes
9	Casting process	0 1	Casting Direct plate (no casting)
10	Facility size/modularity	0 1	Small/modular Large
11	Circularity	0 1	Does not contain CE component (extract → consume) Contains circular economy component

Appendix C. Markets

Table C-1. Characters related to markets

No.	Character	No.	State
1	Low-carbon market eligible	0	No
		1	Some
		2	Yes

Appendix D. Supply Sources

Table D-1. Characters related to supply sources

No.	Character	No.	State
1	Grid electricity for process	0 1	Not used Yes
2	Green hydrogen for process	0 1	Not used Yes
3	Coal as process feedstock	0 1 2	Not used Coal coke
4	Scrap	0 1 2	Not used <50% iron feed >50% iron feed
5	Iron ore	0 1 2	Not used Low grade High grade
6	Natural gas as process feedstock	0 1	Not used Yes

Appendix E. Organization Structures

Table E-1. Characters related to organizational structure

No.	Character	No	State
1	Collaboration	0 1	No Yes

Appendix F. Character Matrix

Table F-1. Character matrix used for cladogram construction

Characters/Taxa	BF - BOF	BF - BOF - CCUS	Scrap - EAF	DRI & BF - BOF	DRI & Scrap - EAF	SIDERWIN	H_2DRI	Boston Metal	Hlsarna	Nucor Castrip	SuSteel (Power to Hydrogen)	Power-to-iron/organic chemical production	power to methanol - Carbon2Chem	power to methanol - FReSMe	eForFuels
Iron product	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1
Coproduct	0	3	0	0	0	0	0	0	0	0	0	3	2	1	3
Ironmaking	1	1	0	2	3	1	3	1	1	0	1	1	1	1	1
Power to X	0	0	0	0	0	1	0	1	0	0	1	1	1	1	1
Electrolytic Iron Reduction Process	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0
Carbon used as reductant	0	0	1	0	1	1	1	1	0	1	1	1	0	0	0
Preprocessing of coal required	2	2	0	1	0	0	0	0	1	0	0	0	2	2	2
Pelletization of iron ore required	1	1	0	1	0	1	1	0	0	0	0	1	1	1	1
High temperature	2	2	0	2	0	0	1	2	2	2	2	2	2	2	2
Carbon-neutral potential	0	1	1	0	1	1	1	1	0	1	1	1	0	0	0
Casting Process	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
Facility size/modularity	1	1	0	1	0	0	0	0	1	0	0	1	1	1	1
Circularity	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Low-carbon market eligible	0	2	1	0	1	2	2	2	1	0	2	2	1	1	1
Grid electricity for process	0	0	1	0	1	1	0	1	0	1	1	1	1	1	1
Green hydrogen for process	0	0	0	0	0	0	1	0	0	0	1	0	1	1	1
Coal as process feedstock	2	2	0	0	0	0	0	0	1	0	0	0	2	2	2
Scrap	0	0	2	0	2	0	0	0	0	2	0	0	0	0	0
Iron ore	2	2	0	2	2	2	2	1	1	0	1	2	2	2	2

Characters/Taxa	BF - BOF	BF - BOF - CCUS	Scrap - EAF	DRI & BF - BOF	DRI & Scrap - EAF	SIDERWIN	H_2DRI	Boston Metal	Hlsarna	Nucor Castrip	SuSteel (Power to Hydrogen)	Power-to-iron/organic chemical production	power to methanol - Carbon2Chem	power to methanol - FReSMe	eForFuels
Natural gas as process feedstock	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0
Collaboration	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1