UHCSDB (UltraHigh Carbon Steel micrograph DataBase): tools for exploring large heterogeneous microstructure datasets

Brian L. DeCost, Matthew D. Hecht, Toby Francis,
Bryan A. Webler, Yoosuf N. Picard, Elizabeth A. Holm

May 12, 2017

preprint accepted for publication in IMMI (DOI: 10.1007/s40192-017-0097-0})

Abstract

We present a new microstructure dataset consisting of ultrahigh carbon steel (UHCS) micrographs taken over a range of length scales under systematically varied heat treatments. Using the UHCS dataset as a case study, we develop a set of visualization tools for interacting with and exploring large microstructure and metadata datasets. Based on generic microstructure representations adapted from the field of computer vision, these tools enable image-based microstructure retrieval, as well as spatial maps of both microstructure and related metadata, such as processing conditions or properties measurements. We provide the microstructure image data, processing metadata, and source code for these microstructure exploration tools. The UHCS dataset is intended as a community resource for development and evaluation of microstructure data science techniques, and for creation of microstructure data science teaching modules.

1 Introduction

We introduce a microstructure dataset[1] focusing on complex, hierarchical structures found in a single ultrahigh carbon steel (UHCS) alloy under a range of heat treatments performed by Hecht et al. [2, 3]. In a concurrent report, we use this dataset to evaluate several microstructure representations based on contemporary computer vision research, and discuss application of both supervised and unsupervised machine learning methods to yield insight into microstructure–properties relationships [4].

This document describes the contents of the UHCS dataset [1] in detail, and outlines the data visualization tools we developed for exploring microstructure datasets with processing and/or properties metadata. We reflect on our experience using a simple SQL database to manage microstructure and processing metadata instead of choosing one of the emerging materials data standards. We also present a responsive web application with microstructure-query and metadata visualization tools, currently accessible online at `http://uhcsdb.materials.cmu.edu` [5].

The UHCS microstructure and metadata dataset can be used by the materials community to define benchmark microstructure data science tasks, such as microstructure classification, microstructure clustering, and developing data-driven microstructure models for processing-structure-properties relationships. The dataset can also support the development and evaluation of new data-driven microstructure representations that address changes in physical scale, magnification, and sample orientation. Finally, the microstructure data visualization tools we present can be reused and extended to enable exploratory analysis of large microstructure datasets.

2 Materials and Methods

The Ultrahigh Carbon Steel (UHCS) dataset consists of scanning electron micrographs (SEM) collected to study the effects of a series of heat treatments on the microstructure features responsible for the hardness and toughness of UHCS [2, 3]. Specimens came from a commercial roll-mill casting with a nominal carbon content of 2%, with detailed composition shown in Table 1. The specimens were annealed at temperatures ranging from 700 $^{\circ}C$ to 1100 $^{\circ}C$ for durations ranging from 5 minutes to 85 hours before quenching, as shown in Tables 2, 3 and 4. The `quantity` column indicates the number micrographs in the dataset processed at the corresponding condition. Most of the microstructures in this dataset are the result of annealing at 970 $^{\circ}C$ and at 800 $^{\circ}C$, between the eutectoid temperature of 723 $^{\circ}C$ and the melting temperature of around 1150 $^{\circ}C$ for ultra high carbon steels. Specimens were water quenched or cooled more slowly in air or in the furnace, as indicated in Table 2. Most specimens for which annealing metadata are unavailable are in the as-cast pearlitic microstructure state.

Figure 1 shows one example micrograph each for the primary microconstituents found in the UHCSDB. These microconstituents include (a) pearlite typical of the as-cast material, (b) the proeutectoid cementite network characteristic of ultra-high carbon steels, (c) spheroidite, (d) pearlite containing

Table 1: Nominal as-cast composition of the present UHCS alloy.

C	Si	Mn	Cr	Ni	Mo
2.02	0.65	0.72	3.86	1.45	0.33

Table 2: Listing of quench methods.

quench method	description	quantity
AR	air cooled	20
FC	furnace cooled	73
Q	quench	489
650 1H	650 ° C for 1 hour	16
N/A	mostly as-cast	363
total		**961**

spheroidite, (e) Widmanstätten cementite, and (f) martensite. Heat treated material was polished with 0.04 μm alumina and etched in 4-5% Nital for 30 seconds before image acquisition in a Philips XL-30 SEM at an accelerating voltage of 20kV. Most of these micrographs were collected using secondary electron imaging, though some backscattered electron images appear in the dataset.

The microscope used to collect these micrographs did not export any imaging metadata in a machine-readable format. Because the same SEM happened to be used for each image, the human-readable metadata accompanying the micron bar at the bottom of each image was laid out consistently (refer to Figure 1(a)–(f)). As a result, we were able to use a semi-automated

Table 3: Annealing temperatures in $^{\circ}C$.

annealing temperature	quantity
700	11
750	4
800	149
900	60
970	344
1000	14
1100	16
N/A	363
total	**961**

Table 4: Annealing times in minutes (M) and hours (H).

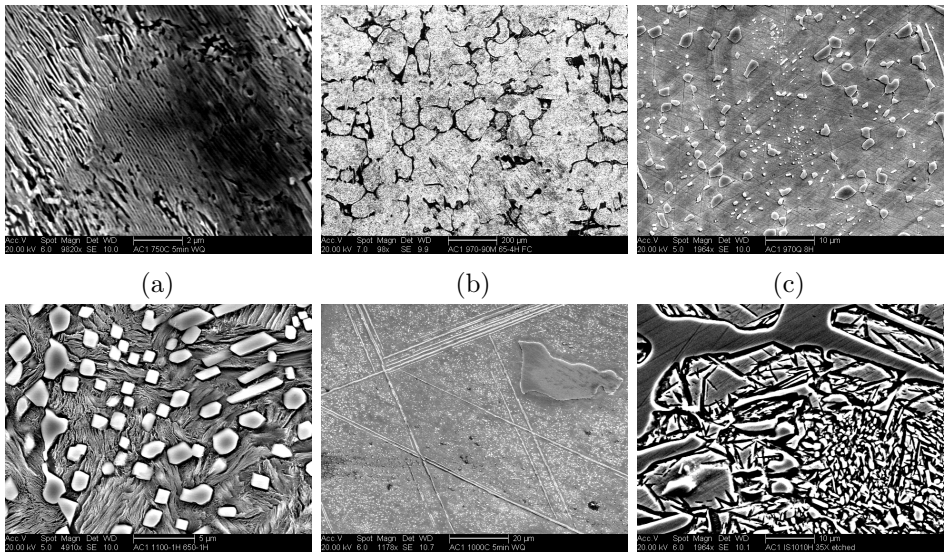
annealing time	quantity
5M	89
1H	16
90M	173
3H	100
8H	69
24H	115
48H	5
85H	31
N/A	363
total	**961**

approach to recover some imaging metadata, including the magnification, imaging mode, and most importantly the physical scale of each image in microns per pixel. In this dataset, the micron bars consistently have the highest aspect ratio out of any of the white elements on the black metadata panel, making it trivial to obtain their length in pixels. We extracted the textual metadata using `tesseract` OCR [6], an open source optical character recognition system. Because each text metadata field has a consistent bounding box, we can crop the corresponding image patch from the metadata panel and pass it to `tesseract` to obtain each metadata field as a string of characters.

Due to the wide variety of formats used for micron bars, this sort of automated metadata recovery is not possible in general for microstructure datasets where the imaging metadata was not preserved. Additionally, reliably using OCR to extract image metadata requires substantial tuning and review. One common error we encountered was substitution of pm or um for scale bar units shown in μm; in this instance, the number of unique results is low enough to manually identify and programmatically correct each type of erroneous reading. These factors highlight the need for ubiquitous and standardized storage of imaging metadata at the point of collection.

3 Database structure.

Internally, we use a SQLite database to manage the microstructure metadata and link it to raw image files and numerical microstructure representations. Raw images are stored as plain `png` and `tif` files, and numerical microstructure



(a) (b) (c)

(d) (e) (f)

Figure 1: Primary microstructure constituents in the UHCS dataset: (a) pearlite, (b) proeutectoid cementite network microstructure, (c) spheroidized cementite, (d) pearlite containing spheroidized cementite, (e) Widmanstätten cementite, and (f) martensite and/or bainite.

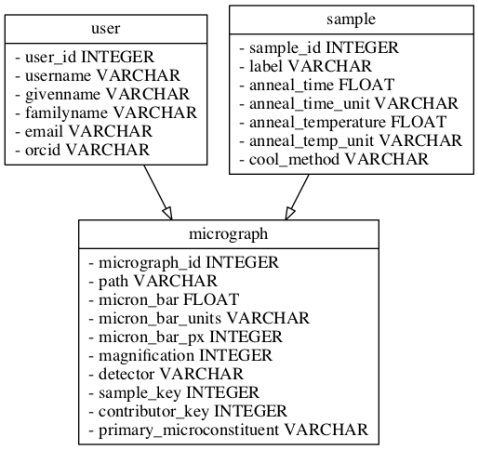


Figure 2: SQLite schema for UHCSDB.

representations are stored in a simple HDF5 format.

The structure of the UHCS dataset and its focus on microstructures and their processing and imaging metadata suggests a simple division into three tables, as illustrated in Figure 2. The `User` table contains identifying information for individual researchers (e.g. ORCID) and the `Sample` table contains metadata specific to a particular physical material specimen (e.g. processing conditions relevant to all measurements taken from that specimen, such as annealing time). The `Micrograph` table contains metadata describing individual micrographs (e.g. a filesystem path to the image file; a primary microconstituent label) and SQL relationships linking each micrograph back to the corresponding `Sample` and `User` records. Finally, an additional `Collection` table (not shown) can aggregate micrographs and samples at the level of individual publications, research groups, or projects. Collections might be associated with a DOI to facilitate citability, as well as to link individual records to the definitive specification of their processing and measurement history.

The relational structure of the metadata (where multiple micrographs share the same processing metadata) make this organization a clear choice over common text-based (comma-separated value, json) and binary (HDF5) formats for tabular data by reducing the complexity of code written to query, manipulate, and update the data. The binary data and UHCSDB web application URLs associated with the micrograph records in the SQLite database are organized using the integer primary keys for the `Micrograph` table (i.e. `Micrograph.micrograph_id`).

For example, the `Micrograph.path` field stores a relative filesystem path to the corresponding raw `png` or `tif` format microstructure image file (`micrograph1219.tif` for a micrograph with a primary key of `1219`). Similarly, the URL `http://uhcsdb.materials.cmu.edu/visual_query/1219` requests microstructure-based search results for micrograph `1219`.

Additional binary data associated with each micrograph (e.g. reduced-dimensionality microstructure representations) are stored in HDF5 format indexed with the integer primary key from the corresponding record in the `Micrograph` table. Microstructure representations from each method described in [4] are stored in separate HDF5 files (one HDF5 file for each method). Each vectorial microstructure representation is stored in a dataset named with the corresponding primary key in the `Micrograph` table: the feature vector for micrograph `1219` is stored in the HDF5 dataset `/1219`. The reduced-dimensionality microstructure representations are stored in a similar format, except that microstructure representations for each dimensionality reduction technique are organized into HDF5 groups, so that the t-SNE map point for micrograph `1219` is stored in the HDF5 dataset `/t-SNE/1219`. Parameters specific to each dimensionality reduction algorithm are stored as attributes to the top-level HDF5 group, including the implementation we used (e.g. `sklearn.decomposition.PCA` for the scikit-learn implementation of principal component analysis) to simplify reproducibility.

The main advantages of a SQL database over the more specialized emerging materials data formats are the simplicity, stability, and ubiquity of plain SQL, and the surrounding ecosystem of related supporting libraries available in many popular programming languages. Specifically, tools like the python libraries `sqlalchemy` and `pandas` allow users unfamiliar with database systems to interact with the data by writing plain python code, instead of learning a new database system or query language (e.g. SQL). For the binary data, HDF5 carries the advantage of accessibility to multiple programming languages compared with e.g. `matlab` or `numpy` binary formats, and offers more structure and performance compared with plain text formats. These factors simplify the process of loading microstructure image data and processing metadata for use with analytic and exploratory tools.

The most significant disadvantage of using a custom SQL schema is its inflexibility. The SQLite schema presented in this section was designed for expedience in organizing the data for the experiments presented in [4], and will not generalize to new microstructure datasets with different sets of processing and properties metadata. Moving forward, two general options are available: commit to one of the emerging materials data formats (e.g. Citrine's PIF [7, 8] or Materials Commons [9]), or iteratively adapt custom

organizations while mapping out the data and infrastructure requirements of microstructure data science applications. As the microstructure community converges on data standards and data infrastructure matures and stabilizes, well-documented custom data formats can readily be converted into standard formats and integrated into community repositories.

4 Tools for exploratory analysis of microstructure datasets

The ability to concisely describe, evaluate, and synthesize large bodies of microscopy work performed over an extended period of time in a collaborative environment is a challenge of long-term, large-scale materials research projects. Often, the institutional memory surrounding microstructure data strongly depends on the humans involved. Even where data is stored digitally, it is typically inaccessible for automated analysis, and it may be difficult for humans to locate and discover specific pieces of data[1]. The global image representations discussed in [4] support multiple novel tools of exploring microstructure datasets and scaling up collaborative research efforts by enabling new means of interacting with and exploring microstructural image datasets. High-dimensional nearest neighbor search can rank micrographs by some measure of microstructural similarity (Section 4.1). Dimensionality reduction algorithms can also be applied to display thumbnail images (Section 4.2) or processing/properties metadata on a spatial microstructure map (Section 4.3).

We built a simple microstructure-oriented responsive web application using open source tools, allowing users to interactively explore the UHCSDB metadata and microstructure dataset via our microstructure dataset exploration tools. Such a web application can easily be deployed locally on a personal machine or local network for internal use, or on the public internet via appropriate infrastructure (web server, hosting, domain name registration, etc.). The UHCSDB web application is currently available at `http://uhcsdb.materials.cmu.edu`. See Section 6 for details on accessing the full microstructure dataset along with source code for the data visualization tools.

4.1 Microstructure query tool

The primary interface of the UHCSDB web application is the microstructure query tool which, given a micrograph, conducts a nearest neighbor search

[1] Datasets organized in collections of slide decks are commonplace.

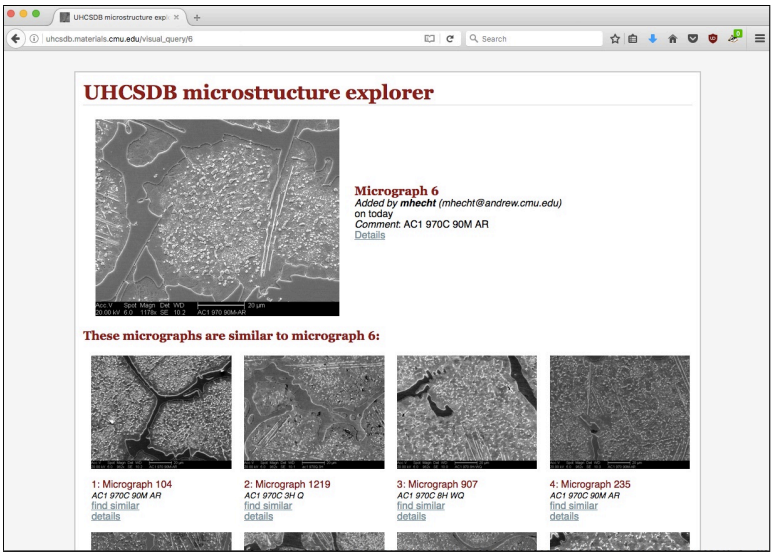


Figure 3: Screenshots from microstructure query tool.

for images in the dataset with similar microstructural content. The nearest neighbor search can operate on any suitable microstructure representation; here we use the multiscale convolutional neural network (CNN) representation described in [4]. CNNs compose multiple layers (as many as one hundred layers in some modern architectures) of convolution filters to extract highly abstract image representations useful for a variety of visual, spatial, and auditory tasks [10]. Our CNN representations are constructed from the internal activations of a 13-layer CNN trained to perform an object recognition task [4]. We combine activations from multiple scales in the input images to obtain representations that are more robust with respect to changes in magnification.

Figure 3 shows how the UHCSDB web application displays nearest neighbor results with a brief summary of the available metadata for each micrograph. Clicking on any of the image search results will initiate a new microstructure query for similar micrographs to the selected search result. Users can surf links between similar microstructure images to explore the full dataset, much as a user of the internet might surf links between e.g. related Wikipedia articles.

4.2 Offline t-SNE image maps

High-dimensional microstructure representations enable us to train classification models to map microstructure inputs to the primary microstructure class labels, or to the annealing conditions [4]. However, humans have limited ability to interpret these high-dimensional vectors, so we turn to dimensionality reduction techniques [11] to gain intuitive insights into microstructure datasets by embedding the complex high-dimensional data into an easily explorable 2D (or 3D) space. We use these low-dimensional microstructure embeddings to create maps of microstructure image thumbnails, as shown in Figure 4, which help address the problems of microstructure dataset discoverability and summarization.

The dimensionality-reduction techniques commonly used in materials data science (e.g. principal component analysis) sometimes fail to adequately represent the structure of complex, noisy, and potentially nonlinear real-world data distributions, such as the CNN representations for the present UHCS micrographs. While we explored multiple dimensionality reduction techniques (see Section 4.3), we found that t-SNE (stochastic neighbor embedding) [12] consistently yields high quality data visualizations for the UHCS microstructure data. The t-SNE algorithm [12] yields a low-dimensional representation by using a stochastic optimization procedure that attempts to conserve the local neighborhood structure of the high-dimensional data, rather than the global structure as in PCA. The ability of t-SNE to reveal local structure in high-dimensional data comes at the cost of distorted depiction of larger distances. t-SNE heavily penalizes large low-dimensional distances between pairs of map points that have small high-dimensional distances, but effectively ignores pairs of points with large high-dimensional distances. Thus large distances between pairs of data points in the low-dimensional maps produced by t-SNE carry little significance. Despite this compromise, t-SNE often provides interesting and visually useful depictions of real-world high-dimensional datasets [12], often in settings where linear techniques break down.

The microstructure map in Figure 4 shows some of the key microstructures in the UHCS dataset; the inset scatter plot shows the full t-SNE map with the selected view indicated by the black frame. Colors indicate the primary microconstituent labels. The image map is best viewed electronically in its complete form, available in the supplemental materials along with maps for additional microstructure representations. This microstructure map was obtained by applying t-SNE to multiscale fifth block CNN features as described in detail as $mVGG_5$ features in [4].

The main focus of this microstructure map view is the initial pearlitic

structure. Starting from the bottom right of the map and moving upwards, the pearlite structure changes from high-magnification images of fine pearlite to lower-magnification views of coarser, more complex pearlite structures. On the left half of the map, there is a cluster of pearlitic microstructures containing spheroidized cementite, and at the top left corner of this view there are several micrographs with extensive Widmanstätten cementite. These microstructure maps are useful for summarizing large bodies of characterization work collected over a potentially long time frames in a clear and concise manner.

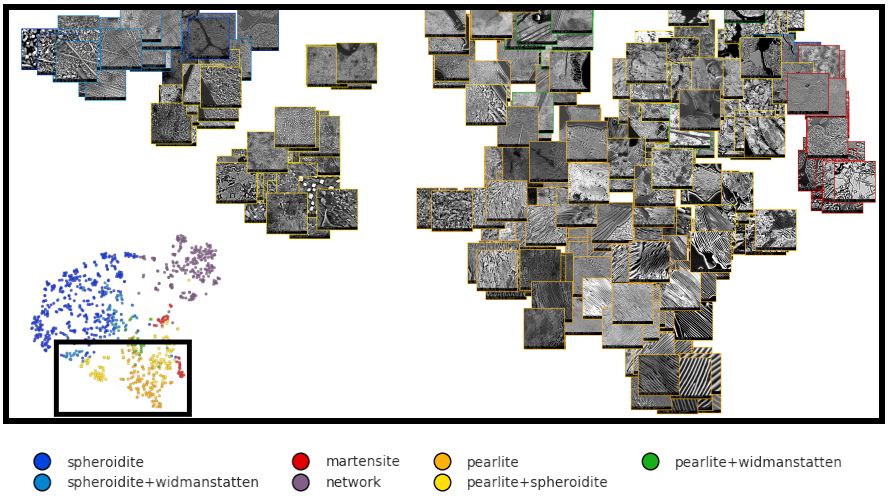


Figure 4: Section of a t-SNE map displaying UHCS micrographs annotated with primary microconstituent, laid out using multiscale CNN features from the fifth block of the VGG network (described in detail as $mVGG_5$ features in [4]).

The bottleneck in this process (for the relatively small UHCS dataset at least) is the dimensionality reduction step. Because we have precomputed the reduced-dimensionality representations these microstructure maps are easy to generate on demand, subject to the practical constraints of sufficient server-side resources or a client-side implementation.

4.3 Interactive metadata visualization tool

The data visualization tool[2] is a `Bokeh` application that displays scatter plots with reduced-dimensionality microstructure representations. A collec-

[2] `http://uhcsdb.materials.cmu.edu/visualize`

tion of drop-down web forms enable the user to choose between available microstructure representations, dimensionality reduction methods, and metadata annotations, as illustrated in Figure 5. The default view (Figure 5(a)) is a t-SNE map with marker colors indicating the primary microconstituents in each micrograph. The drop-down forms allow the user to select alternative metadata to display via the marker colors and relative sizes, including annealing time and temperature (shown in Figure 5(b)) and the magnification. The user can reproduce and explore all of the scatterplot figures in [4] and its supplemental materials, as well as scroll, zoom, and export individual plots (as shown in Figure 5(c), which focuses on the upper-right quadrant of Figure 5(b).).

As shown in Figure 5(d), placing the mouse cursor over a scatterplot marker will display a thumbnail image of the corresponding micrograph, along with some relevant processing metadata. Clicking on a marker will open an URL that triggers a microstructure query for the corresponding micrograph record. This view is not as complete as the microstructure maps presented in Section 4.2, but allows the user to explore relationships between microstructure and metadata in a more interactive manner.

Currently, the (precomputed) reduced-dimensionality methods available on the UHCSDB web application include principal component analysis (PCA) [13], t-SNE [12], multidimensional scaling [MDS, shown in Figure 5(d)] [14], locally linear embedding (LLE) [15], Isomap [16], and spectral embedding [17]. Presently, we precompute reduced-dimensionality representations for each microstructure representation, but in principle they can be computed on demand. Most of these dimensionality reduction methods require only a few seconds of computation for a dataset of this size and complexity, though t-SNE can require a few minutes, especially when computing multiple independent maps. Front-end (client-side) dimensionality reduction implementations may be useful for exploratory and collaborative deployments, compared with the precomputed dimensionality reduction workflow we employ presently.

5 Potential applications

The UHCS microstructures and metadata can serve the materials data science community as a source of benchmark tasks for evaluating and comparing microstructure representations. Though the dataset is small compared to many of the standard datasets used in the computer vision and robotics literature, it's size is likely representative of microstructure datasets currently

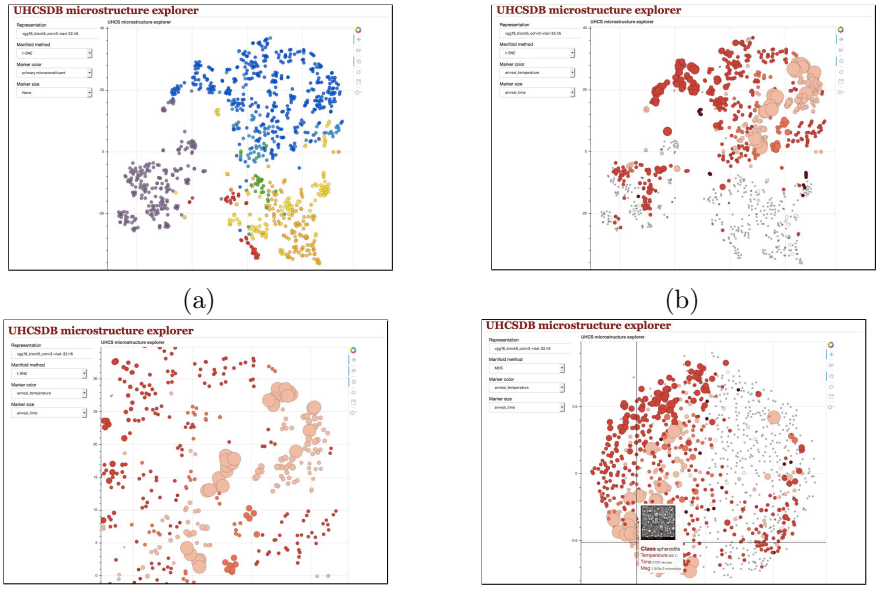


Figure 5: Screenshots from the interactive data visualization tool. (a) Default view: t-SNE map with marker colors indicating primary microconstituent. (b) t-SNE map with marker colors indicating annealing temperature and relative marker sizes indicating annealing time. (c) Zoomed-in view of the map in (b). (d) Multidimensional scaling map showing time and temperature metadata, illustrating the micrograph thumbnail tooltip.

collected by individual researchers.

Microstructural complexity makes the UHCS dataset an interesting challenge for microstructure segmentation and representation. As in many important microstructure systems, the relevant microstructure features (especially spheroidized cementite in this case) vary in physical length scale as well as in the relative length scale of the image reference frame (i.e. differing magnification). These aspects make the UHCS dataset a useful resource for researchers interested in developing microstructure representations that are invariant, equivariant, or covariant to both scale and rotation. Similarly, the UHCS dataset is a promising resource for materials data scientists to build teaching modules for microstructure informatics techniques.

We are also currently using the microstructure dataset to explore application of semantic segmentation techniques to complex microstructure systems, which could support and accelerate conventional microstructure-based research. For example, Hecht et al. [3] used a laborious semi-automated process to segment the spheroidized cementite particles (e.g. in Figure 1(c), Figure 1 in [3]) to enable their analysis of the cementite coarsening kinetics. They also manually traced branches of the proeutectoid cementite network (e.g. in Figure 1(b), Figure 3 in [3]) and the surrounding particle-free denuded regions to support their particle coarsening analysis. Automation of these kinds of microstructure analytics tasks could significantly lower the cost of gathering and analyzing statistically-meaningful quantities of microstructure data.

Finally, the web application and exploratory analysis tools presented in this manuscript can be repurposed and adapted for analysis of other microstructure datasets. These tools could help individual researcher groups scale up their analysis and interpretation of microstructure data internally. Additionally, combined with emerging data curation platforms such as Citrination[7], Materials Commons[9], and the NIST DSpace[18], these microstructure dataset visualization tools could impact the way researchers interact with the materials science literature, as is being done for numerical materials properties[7, 19, 20]. What if every materials characterization paper had an interactive microstructure and metadata supplementary publication, instead of merely including a select few 'representative' micrographs? Integration of microstructure-based search and visualization tools into the materials data infrastructure could significantly improve discoverability.

6 Archival data accessibility

The complete set of micrographs, metadata, and web application source code are available on the NIST repository materialsdata.nist.gov [1].

7 Software requirements

All of the software used in this project is available under permissive open source licenses, and is listed in Table 5, along with additional required python libraries in Table 6. Additional software and configuration may be necessary for public deployment of the interactive data exploration application, such as a web server (we currently use `nginx` and `gunicorn`).

Table 5: Software dependencies used in this project.

project name	use
`SQLite`	database management system
`python`	general purpose programming
`bh_tsne`	Reference t-SNE implementation [12]

Table 6: Python dependencies not included in the python standard library.

module name	use
`h5py`	HDF5 interface
`bokeh`	interactive data visualizations
`flask`	web framework
`keras`	neural network library [21]
`numpy`	numerical computing library
`pandas`	data frames library
`skimage`	image processing library [22]
`seaborn`	statistical plotting library
`sklearn`	machine learning library [23]
`matplotlib`	plotting library
`sqlalchemy`	SQLite interface

8 Conclusion

We present an ultrahigh carbon steel microstructure dataset and suite of microstructure visualization tools. The UHCS dataset is a promising commu-

nity resource for researchers interested in developing data-driven methods linking microstructure with processing/properties metadata. The dataset is also ideal for the creation of microstructure data science teaching resources to enable workforce development. We hope the microstructure and metadata visualization tools we present will be integrated into the burgeoning ecosystem of materials data repositories to increase the discoverability of microscopy datasets. Finally, these tools may help large collaborative projects scale up and speed up the microstructure collection, curation, and analysis components of their work.

Acknowledgements

We gratefully acknowledge funding for this work through National Science Foundation grants DMR-1307138 and DMR-1501830, and through the John and Claire Bertucci Foundation. Data visualization tool development by B.D., T.F., and E.H.; UHCS microscopy work by M.H., Y.P., and B.W. This work was supported in part by the Commonwealth of Pennsylvania Department of Community and Economic Development (DCED) Developed in PA program (D2PA), and by National Science Foundation grant CMMI-1436064. The as-cast and heat treated UHCS samples were provided by Miller Centrifugal Casting.

References

[1] Matthew D. Hecht, Brian L. DeCost, Toby Francis, Elizabeth A. Holm, Yoosuf N. Picard, and Bryan A. Webler. Ultrahigh carbon steel micrographs. `https://hdl.handle.net/11256/940`. [Cited on pages 1, 2, and 15.]

[2] Matthew D Hecht, Bryan A Webler, and Yoosuf N Picard. Digital image analysis to quantify carbide networks in ultrahigh carbon steels. *Materials Characterization*, 117:134–143, 2016. [Cited on pages 1 and 2.]

[3] Matthew D. Hecht, Yoosuf N. Picard, and Bryan A. Webler. Coarsening of inter- and intra-granular proeutectoid cementite in an initially pearlitic 2c-4cr ultrahigh carbon steel. *Metallurgical and Materials Transactions A*, 48(5):2320–2335, 2017. [Cited on pages 1, 2, and 14.]

[4] Brian L. DeCost, Toby Francis, and Elizabeth A. Holm. Exploring the microstructure manifold: image texture representations applied to

ultrahigh carbon steel microstructures. *Acta Materialia*, page In Press, 2017. [Cited on pages 1, 7, 8, 9, 10, 11, and 12.]

[5] Brian L. DeCost, Toby Francis, and Elizabeth A. Holm. Uhcsdb microstructure explorer. `http://uhcsdb.materials.cmu.edu`. Accessed 14 April 2017. [Cited on page 2.]

[6] Ray Smith. An overview of the tesseract ocr engine. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 629–633. IEEE, 2007. [Cited on page 4.]

[7] Jordan O'Mara, Bryce Meredig, and Kyle Michel. Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access. *JOM*, 68(8):2031–2034, 2016. [Cited on pages 7 and 14.]

[8] Kyle Michel and Bryce Meredig. Beyond bulk single crystals: A data format for all materials structure–property–processing relationships. *MRS Bulletin*, 41(8):617–623, 2016. [Cited on page 7.]

[9] Brian Puchala, Glenn Tarcea, Emmanuelle A Marquis, Margaret Hedstrom, HV Jagadish, and John E Allison. The materials commons: a collaboration platform and information repository for the global materials community. *JOM*, 68(8):2035–2044, 2016. [Cited on pages 7 and 14.]

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. [Cited on page 9.]

[11] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10:66–71, 2009. [Cited on page 10.]

[12] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. [Cited on pages 10, 12, and 15.]

[13] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002. [Cited on page 12.]

[14] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. [Cited on page 12.]

[15] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. [Cited on page 12.]

[16] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. [Cited on page 12.]

[17] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. [Cited on page 12.]

[18] NIST repositories. `https://materialsdata.nist.gov`. Accessed 14 April 2017. [Cited on page 14.]

[19] Joanne Hill, Gregory Mulholland, Kristin Persson, Ram Seshadri, Chris Wolverton, and Bryce Meredig. Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bulletin*, 41(05):399–409, 2016. [Cited on page 14.]

[20] Ram Seshadri and Taylor D Sparks. Perspective: Interactive material property databases through aggregation of literature data. *APL Materials*, 4(5):053206, 2016. [Cited on page 14.]

[21] François Chollet. Keras, 2015. [Cited on page 15.]

[22] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. [Cited on page 15.]

[23] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011. [Cited on page 15.]