Contract No:

This document was prepared in conjunction with work accomplished under Contract No. DE-AC09-08SR22470 with the U.S. Department of Energy (DOE) Office of Environmental Management (EM).

Disclaimer:

This work was prepared under an agreement with and funded by the U.S. Government. Neither the U. S. Government or its employees, nor any of its contractors, subcontractors or their employees, makes any express or implied:

1) warranty or assumes any legal liability for the accuracy, completeness, or for the use or results of such use of any information, product, or process disclosed; or
2) representation that such use or results of such use would not infringe privately owned rights; or
3) endorsement or recommendation of any specifically identified commercial product, process, or service.

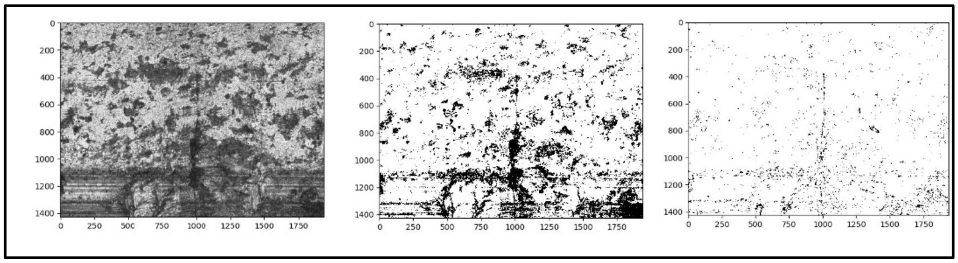
Any views and opinions of authors expressed in this work do not necessarily state or reflect those of the United States Government, or its contractors, or subcontractors.

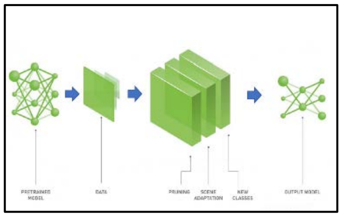
Title: Process Image Analysis using Big Data, Machine Learning, and Computer Vision

Within the DOE complex, 3013 canisters are used to store Pu-bearing waste. Wall penetration due to corrosion, specifically stress corrosion cracking, is considered to be the most likely cause of failure over the 50-year lifetime of the canisters.

This project has 2 objectives: The first is the development of machine learning algorithms to identify the presence of corrosion from a very large set of images generated by laser confocal microscope scanning of 3013 canister used to store Pu oxides. This portion of the LDRD constitutes image analysis of a metal surface for the presence of corrosion. The image processing algorithms developed for this project will provide a suitable basis for analysis of other types of corrosion data, which can be produced in vast quantities using modern devices.

The second component of the LDRD consists of the development of machine learning algorithms that obtain molecular mechanics force-fields from ab-initio Density Functional Theory (DFT) calculations for corrosive attack by chlorides on 304L or 316L stainless steel. The goal of this latter component is to determine means for mitigating corrosion on a fundamental level, including coatings, welding methods, metal composition, etc. Force-field modeling is necessary for this endeavor because the incipience and progression of corrosion is the governed by molecular structures, grain boundaries, dislocations and surface structures represented by large numbers of atoms. Although DFT calculations are extremely adept at describing molecular scale processes, they are computationally expensive making them ill-suited for calculations of more than several hundred atoms, especially for the repeated applications inherent in material design. Fortunately, force-field methods provide an avenue for viable molecular-scale calculations involving the numbers of atoms involved in corrosion processes. The accuracy of the force-field calculations depends strongly on the accuracy of the force-field model, which is extremely difficult and time-consuming to derive from either data or ab-initio calculations. The use of machine learning algorithms has the potential to make the calculation of force-fields much more efficient but has not been explored significantly for corrosion processes. If shown to be viable, the technique would have wide -ranging impact on design of corrosion resistant materials and on the mitigation of corrosion in existing process systems.





Original image→Gaussian blur→threshold→erosion+dilation to find crack. ML for derivation of FF from DFT

Awards and Recognition

Presented LDRD results at the 2019 AIChE spring meeting. Corresponding paper was published in the Proceedings: 2019 AIChE Spring Meeting and 15th Global Congress on Process Safety.

Intellectual Property Review

This report has been reviewed by SRNL Legal Counsel for intellectual property considerations and is approved to be publicly published in its current form.

SRNL Legal Signature

______________________________ _______________

Signature **Date**

LDRD External Report Summary

Title: Process Image Analysis using Big Data, Machine Learning, and Computer Vision

Project Team: Bruce Hardy (PI), Anna d'Entremont, Michael Martinez-Rodriguez, Brenda Garcia-Diaz, Lindsay Roy,

Subcontractor: Jason Bakos (USC), Taylor Clingenpeel (USC), Phil Moore (USC), Ben Torkian (USC), R Doran (USC),AJ Medford (GT)

Thrust Area: SEM

Project Start Date: October 1, 2018
Project End Date: September 30, 2020

The development of algorithms for machine learning and data analysis for the 3013 MIS corrosion surveillance program is a collaborative effort by SRNL, USC and GT. For corrosion detection, LCM image data is extracted from large binary files, with software written to convert the data to physical attributes (i.e. height, color and grayscale values; all as functions of a location in a plane projection). The user interface for the software permits selective downloading of binary data and interrogation of attributes. User input thresholds are used to flag attributes of interest. Machine learning algorithms, developed for this application, are used to determine whether the features are the result of corrosion.

To address the fundamental mechanisms of corrosion, machine learning algorithms are being developed to derive interatomic potential force-fields from ab-initio DFT calculations. The goal is to apply molecular modeling on a large enough scale to guide the design of resistant materials.

FY2019 Objectives

- Develop machine learning methods, based on computer vision, to analyze imaging data for corrosion
- Develop machine learning methods for molecular modeling of corrosion processes
- Identify 3013 data sets, and numerical methods, suitable for near-term development
- Determine preliminary set of attributes for training supervised ML algorithms
- Assemble training sets, train and test ML algorithms
- Classify features by size, quantity, density, and location
- Utilize computer vision to reduce amount of data needing manual analysis
- Identify additional data sets within SRNL that can be analyzed using the methodologies developed as part of this project
- Begin development of ML methodology for obtaining adaptive force-fields from ab-initio molecular models (MM) for corrosion (added to originally approved scope)

Introduction

Halides contained in Pu-bearing material have been found and produce corrosion in the Inner Can Closure Weld Region (ICCWR for the 3013 canister system used throughout the DOE complex. Inspections using a LCM produce immense amounts of image data: approximately 6000 images per can, having 786,432 pixels per image, with 8 layers of data for each pixel. There is currently a 5-year backlog of images, with approximately 5 canisters/year, that must be evaluated. Simplistic computer-aided image analysis can flag parameters, such as pit depth and cracking to guide manual examinations for corrosion. However, while this approach greatly improves the efficiency of the examination process compared to unaided manual screening, it is still excessively time consuming. A more sophisticated approach is to assess the data using machine learning algorithms to identify corrosion without manual intervention. As a complement to corrosion detection, molecular level analyses can yield a fundamental understanding of corrosion occurring in the ICCWR and guide the design of corrosion resistant materials, welding processes, and coatings. These 2 efforts comprise the research in this LDRD.

This 2-year LDRD project has 2 concurrent objectives: The first is the development of machine learning algorithms to identify the presence of corrosion from a very large set of images generated by LCM scanning of 3013 canisters used to store Pu oxides. This portion of the LDRD constitutes image analysis of a metal surface for the presence of corrosion. The image processing algorithms developed for this project will provide a suitable basis for analysis of other types of corrosion data, which can be produced in vast quantities using modern devices.

The second component of the LDRD consists of the development of machine learning algorithms that obtain molecular mechanics force-fields from ab-initio Density Functional Theory (DFT) calculations for corrosive attack by chlorides on 304L or 316L stainless steel. The goal of this latter component is to determine means for mitigating corrosion on a fundamental level, including coatings, welding methods, metal composition, etc. Force-field modeling is necessary for this endeavor because the incipience and progression of corrosion is governed by molecular structures, grain boundaries, dislocations and surface structures represented by large numbers of atoms. Although DFT calculations are extremely adept at describing molecular scale processes, they are computationally expensive making them ill-suited for calculations of more than several hundred atoms, especially for the repeated applications inherent in material design. Fortunately, force-field methods provide an avenue for viable molecular-scale calculations involving the numbers of atoms involved in corrosion processes. The accuracy of the force-field calculations depends strongly on the accuracy of the force-field model, which is extremely difficult and time-consuming to derive from either data or ab-initio calculations. The use of machine learning algorithms has the potential to make the calculation of force-fields much more efficient. However, the application of machine learning to force-field derivation has not been explored significantly for corrosion processes and, if shown viable, would have wide-ranging impact on design of corrosion-resistant materials and on the mitigation of corrosion in existing process systems.

Approach

First Component – Image Analysis

Corrosion is strongly, but not exclusively, associated with surface pitting and cracking, coloration, along with shapes and patterns of surface features. Conversely, not all pits and surface lesions are the result of corrosion: some are artifacts of fabrication, impact, scoring or other non-corrosion events. Corrosion is identified via the combined properties of pit depth, area, edge contour, color and clustering. Software was developed to extract these features from large binary files generated by the LCM. The individual images, which collectively span the ICCWR are stitched together and corrected to eliminate the effect of curvature on measurement of the local height. Various methods are applied to the data to best relate it to presence of corrosion. Mathematical operations invoked for computer vision and image interpretation include, but are not limited to: labeling, gradient methods, statistical characterization, correlations and filters[1,2]. The processed data is then input to ML algorithms; labeled data for training, and afterwards data for evaluation by the trained ML algorithm. The process is shown schematically in Figure 1.

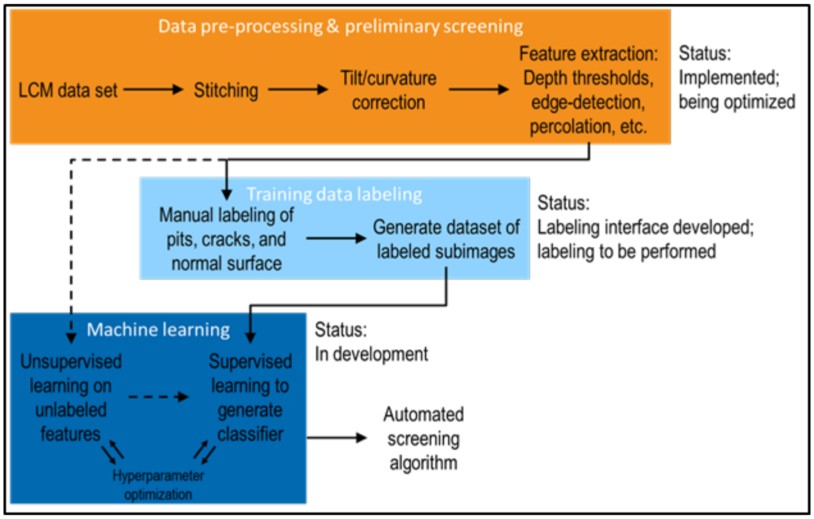

Data pre-processing & preliminary screening
LCM data set
Stitching
Tilt/curvature correction
Feature extraction: Depth thresholds, edge-detection, percolation, etc.
Status: Implemented; being optimized
Training data labeling
Manual labeling of pits, cracks, and normal surface
Generate dataset of labeled subimages
Status: Labeling interface developed; labeling to be performed
Machine learning
Unsupervised learning on unlabeled features
Supervised learning to generate classifier
Hyperparameter optimization
Status: In development
Automated screening algorithm

Figure 1. *Data processing and implementation of machine learning*

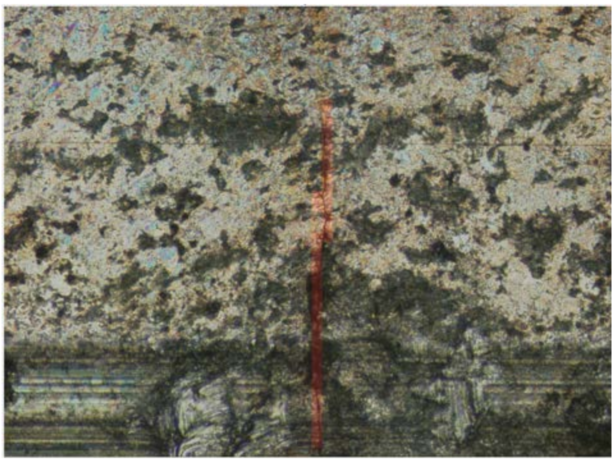
Second Component – ML based FF Derivation from DFT Calculations

The objective of second part of the LDRD project is to advance the fundamental understanding of corrosion by developing novel methods to simulate the complex chemistry and physics through coupling quantum mechanical and empirical force field methods. In corrosion science, sophisticated multiscale models beginning at the *ab-initio* level provide mechanistic insight into metal-environment interactions resulting in general corrosion, intergranular corrosion, and pitting corrosion[3]. Specifically, this research focuses on designing machine-learning algorithms to develop and train adaptive force fields for the study and prediction of corrosion behavior, specifically the metal-environment interface[4]. Accurately calculating the parameters for a robust force field, however, is a much more complicated than a simple regression fit, requiring more sophisticated data analytics[5-7]. Further, the functional form of standard force fields is often insufficient to capture the complex physics of a reactive interface, especially in the case of the complex electronic structure of magnetic metal oxides. While atomistic modeling techniques are well suited to study the chemical reactions occurring at the interface between a material and its environment, modeling corrosion is computationally slow because the models must be constructed to resolve both relevant reaction mechanisms and mass transport processes. By developing machine learning methods to obtain quantitative structure-activity relationships considering both molecular and bulk boundary conditions, researchers will be able to significantly advance knowledge by exploring more combinatorial spaces and nonlinear processes which are difficult using traditional approaches.

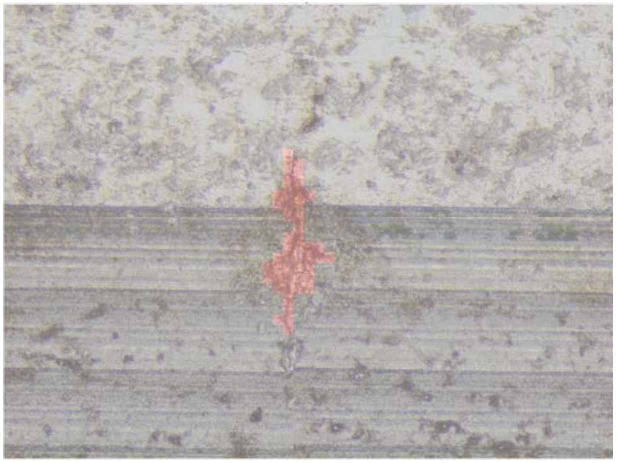
Results/Discussion

LCM image data taken for the MIS program was reviewed to obtain samples containing cracks, pits and other features characteristic of corrosion. The low incidence of corrosion in the actual ICCWR samples made it necessary to obtain more data for proper training of the ML algorithms that were developed for this program. Corrosion data from the 3013 canisters was therefore augmented with LCM images of coupons that were exposed to boiling MgCl. To provide an efficient means for handling large amounts of binary image data a GUI was developed to serve as an interface with the data files, manipulate and group images, label features for training the ML algorithm, group features with used defined thresholds, correct for sample tilt and curvature, and stitch images. The last of the GUI capabilities is extremely important as features of significance, especially cracks, can extend across many images or lie at the interface between them. Moreover, it was found that larger views, consisting of a composite of many images, are necessary for ML crack detection.

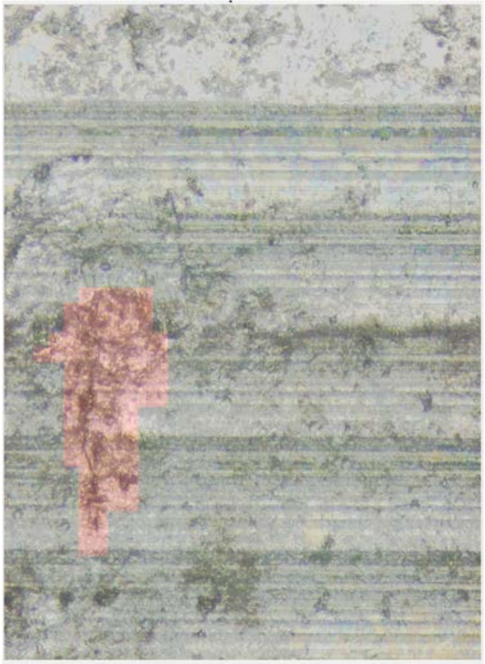
Cracks, pits and color patterns are all associated in various forms with corrosion. Pits can readily be detected using height data thresholds. Cracks, particularly "hairline" cracks do not always have a definitive height signature. Rather, crack identification is a combination of grayscale image intensity (pixel value) and height data. Initially, it was hoped that standard edge detection methods could be used with pixel values to extract crack edges. Methods considered included: erosion and dilation, blurring, Fourier and Gaussian filters, and gradient methods. Unfortunately, other surface features combined to create background noise that was similar in frequency to that associated with crack edges. To overcome this problem, DNN methods were developed and applied to identify cracks. Early in the development of this approach the training of the DNN algorithms suffered due to the small amount of crack data available. Training data was expanded by using synthetic images and will be further expanded using coupon data. After increasing the amount of labeled training data and adjusting the DNN algorithms, selection and recall for the crack identification was significantly improved. The most promising results were achieved by training classifiers using a training set consisting of four portions of scans that each contain one known defect. Each of these was decomposed into fixed-sized blocks, with each block that overlaps any part of a known defect region manually labeled as "defect" and all other blocks labeled as "normal". The blocks labeled as "defect" comprise only a small portion of the total number of blocks. Additional examples of blocks containing defects are generated using data augmentation techniques. The resulting labeled data is used to train a convolutional neural network classifier. As shown in Fig. 2, the scans surrounding four known crack-like defects, named "Bellatrix," "Acrux," "Cursa Major," and "Cursa Minor", serve as the training data. "Bellatrix" and "Acrux" are comprised of 2x2 data file mosaics, while "Cursa Major" and "Cursa Minor" are comprised of 2x1 mosaics. The defect areas are highlighted in red. Prior results indicate that classification performance improves when there are significantly more blocks labeled as "defect" than as "normal". For this, we apply an augmentation approach, in which we apply small 2D spatial offsets to all blocks labeled as "defect" and add the resulting blocks back into the data set (except for those generated blocks where the offset places it outside the image boundary or no longer covering part of the original defect region). This approach is illustrated in Fig. 3. A 5-channel convolutional neural network was used to classify blocks as "defect" or "normal". The input channels include the red, green, and blue colors from the LCM "peak" image, as well as the height data after correcting for curvature, and laser intensity. Feature data for pits cracks and color will be used in a higher level ML algorithm for association with corrosion.

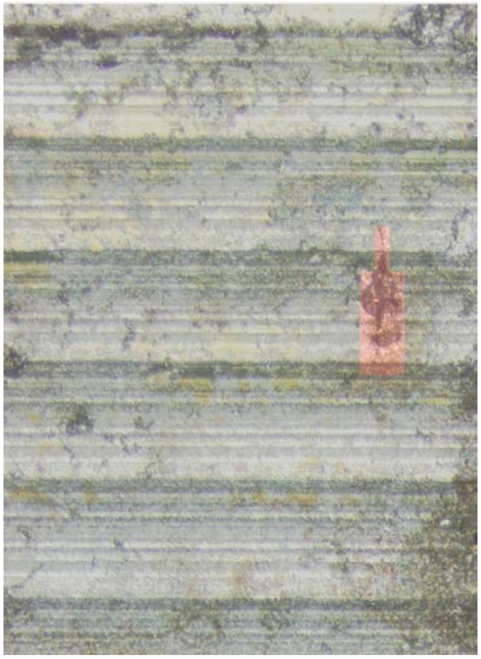


"Bellatrix": 2x2 stitch consisting of 1407x1930 pixels including mosaic overlap. Defect region highlighted in red, comprising 1.4% of total area.



"Acrux": 2x2 stitch consisting of 1407x1925 pixels including mosaic overlap. Defect region highlighted in red, comprising 1.7% of total area.





"Cursa Major": 2x1 stitch consisting of 1407x1024 pixels including mosaic overlap. Defect region highlighted in red, comprising 6.8% of total area.

"Cursa Minor": 1407x1024 pixels including mosaic overlap. Defect region highlighted in red, comprising 1.7% of total area.

Figure 2 Training samples.

Development of FF methods for MD simulation began with DFT and MD training of neural networks using Gaussian descriptors to construct symmetry invariant features into the NN. Benchmark calculations included 100 DFT configurations with a target accuracy energy of 1 meV/atom and MD simulations utilizing a canonical ensemble for 500 MD steps at 300K with 0.5 fs integration timestep. The training set included a pristine Fe 111 surface examining the water/Fe interface. The validation set included 5-6 water layers with the Fe surface. The results indicate that even this relatively large number of DFT calculations is insufficient to capture the wide range of environments that will occur in a corrosion process, and the challenging electronic structure of the magnetic metal/water interface makes generating substantially larger datasets intractable. For this reason, a new strategy of "transfer learning" is being pursued. In this approach, inexpensive MD methods (EAM, TIP3P) are utilized to generate large data set (10-100k images). These images are used to pre-train a deep neural network capable of reproducing the lower-level MD theory. Some intermediate layers of the network are then "frozen" by holding their weights constant, and the remaining layers will be re-trained to reproduce the DFT data. This approach is expected to drastically reduce the amount of DFT data needed to obtain accurate predictions, and also lead to force fields that are reliable even outside the training data regime.

FY2019 Accomplishments

- Developed a pre-processing GUI using the Matlab© user interface
  - Easily used by investigators
  - Reads LCM binary data in native vk4 format
  - Stitches images, tested with 300 but can do many more
  - Corrected jump discontinuity between images
  - Corrected for tilt and curvature to get accurate local surface height data
  - Can simultaneously view multiple features corresponding to different LCM channels
  - Developed feature selection and labeling in the GUI
  - Reduced time to identify features having significant thresholds by factor of 10-18
- Developed methods for processing data for machine learning algorithms
  - Improved edge detection for cracks and pits
  - Developed preliminary machine learning algorithms for crack detection
  - Algorithms for generating labeled data for training sets
- Generating baseline metal lattice configurations for FF training set
  - Calculated Fe/H2O interfaces using DFT and MD
  - Developed preliminary machine learning algorithms for FF generation

Future Directions

- L-basin corrosion analysis.

- Proposals for more complete development of ML applications for FF derivation form DFT calculations, including experimental validation.
- Extension of image analysis to inclusions in articles produced by additive manufacturing.
- Application of AI methods to data analysis, particularly for corrosion and material degradation. This would also include analysis of analytical data (XPS, XRD, SEM, etc.).
- Development of surrogate molecular models having reduced complexity but retaining a high degree of accuracy. This is related to material design at a fundamental level and is a compliment to data analysis that is used to identify material degradation.
- Guided material synthesis based on empirical data with imposed physical constraints.
- Advanced process control, invoking reachability theory and fault tolerance. This aspect of AI, which takes advantage of the volume of data yielded by advanced sensor capability, would be particularly suitable for isotope separation, pit productions and waste processing.
- Test FF algorithm with validation set
- Incorporate iron/iron oxide defects with different water phases
- Replace water with halides for corrosion and reactivity

FY 2019 Publications/Presentations

1. Presented LDRD results at the 2019 AIChE spring meeting. Published associated paper in the Proceedings: 2019 AIChE Spring Meeting and 15th Global Congress on Process Safety.
2. Publication in progress on application of machine learning and edge detection methods in the image analysis for identification and interpretation of cracks, pits and other features related to corrosion.
3. Publication in progress for DFT baseline calculations for training supervised ML algorithm applied to derivation of FF's.

References

1 Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, Cambridge MA (2016)

2 Aurélien Géron. Hands-On Machine Learning with Scikit-Learn &Tensorflow. O'Reilly, Sebastapol, CA (10/12/2018).

3 Gunasegaram, D. R.; *et al.*, *Int. Mater. Rev.* **2014**, *59*, 84.

4 Ghosh, S.; Suryanarayana, P. *Comput. Phys. Commun.* **2017**, *216*, 109.

5 Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R. *J. Phys. Chem. A* **2009**, *113*, 6745.

6 Joyce, J. J.; *et al.*, *Mater. Res. Soc. Symp. Proc.* **2010**, *1264*, Z09-04.

7 Li, W.; Ando, Y. *Physical Chemistry Chemical Physics* **2018**, *20*, 30006.

Acronyms

AI	Artificial Intelligence
AIMD	Ab initio molecular dynamics
CS	Computer Science
DNN	Deep Neural Network
DOE	Department of Energy
DFT	Density Functional Theory

EAM	Embedded Atom Model – an interatomic potential that represents the energy between atoms
FF	Molecular mechanics Force-Field
GT	Georgia Institute of Technology, Atlanta, GA
ICCWR	Inner Can Closure Weld Region
LCM	Laser Confocal Microscope
MD	Molecular Dynamics
MIS	Material Identification and Surveillance program
ML	Machine Learning
MM	Molecular Models
NN	Neural Network
TIP3P	Transferrable Intermolecular Potential with 3 Points – a 3 site rigid water model
USC	University of South Carolina, Columbia, SC

Intellectual Property

None

Total Number of Post-Doctoral Researchers

1 Post-doctoral researcher, performed work at GT

Total Number of Student Researchers

2 Undergraduate students, performed work at USC
1 Graduate student, performed work at GT
1 MSIPP student, performed work at SRNL