

April 2020

Volume 11 Issue 2

The background features a dark blue gradient with glowing yellow and white circuit-like patterns. Faint mathematical formulas are scattered throughout, including  $\frac{dN(t)}{dt}$ ,  $N(t) = N(0)e^{-\lambda t}$ ,  $\text{return}(\text{double})(U);$ ,  $(\text{double}) \text{fib}(n)/n;$ ,  $(\text{int}) n/f;$ ,  $(\text{int}) n-1;$ , and  $+$ .

# JOCSE

**Journal Of Computational Science Education**

**Promoting the Use of  
Computational Science  
Through Education**

**ISSN 2153-4136 (online)**



# JOCSE

Journal Of Computational Science Education

---

**Editor:** Steven Gordon  
**Associate Editors:** Thomas Hacker, Holly Hirst, David Joiner,  
Ashok Krishnamurthy, Robert Panoff,  
Helen Piontkivska, Susan Ragan, Shawn Sendlinger,  
D.E. Stevenson, Mayya Tokman, Theresa Windus

---

**CSERD Project Manager:** Jennifer Houchins **Managing Editor:** Jennifer Houchins. **Web Development:** Jennifer Houchins, Aaron Weeden, Joel Coldren. **Graphics:** Stephen Behun, Heather Marvin.

The Journal Of Computational Science Education (JOCSE), ISSN 2153-4136, is published quarterly in online form, with more frequent releases if submission quantity warrants, and is a supported publication of the Shodor Education Foundation Incorporated. Materials accepted by JOCSE will be hosted on the JOCSE website, and will be catalogued by the Computational Science Education Reference Desk (CSERD) for inclusion in the National Science Digital Library (NSDL).

**Subscription:** JOCSE is a freely available online peer-reviewed publication which can be accessed at <http://jocse.org>.

Copyright ©JOCSE 2020 by the Journal Of Computational Science Education, a supported publication of the Shodor Education Foundation Incorporated.



# Contents

Introduction to Volume 11 Issue 2 <i>Steven I. Gordon, Editor</i>	1
Using Molecular Visualization as a Tool for Culturally Competent and Culturally Relevant Teaching: A Guided-Inquiry Biochemistry Activity <i>Pumtiwitt McCarthy, Richard Williams, Cleo Hughes-Darden, Roni Ellington, Paminas Mayaka, Monica Jackson, and Asamoah Nkwanta</i>	2
The State of Undergraduate Computational Science Programs <i>Steven I. Gordon and Katharine Cahill</i>	7
Development of a Molecular Model for Understanding the Polymer-metal Interface in Solid State Pumps <i>Jaime D. Guevara, Matthew L. Jones, Peter Müllner, and Eric Jankowski</i>	12
Using Blue Waters to Assess Tornadic Outbreak Forecast Capability by Lead Time <i>Caroline MacDonald and Andrew Mercer</i>	23
Improvement of the Evolutionary Algorithm on the Atomic Simulation Environment Through Intuitive Starting Population Creation and Clustering <i>Nicholas Kellas and Michael N. Groves</i>	29



# Introduction to Volume 11 Issue 2

Steven I. Gordon

Editor

The Ohio State University

Columbus, OH

gordon.1@osu.edu

## FOREWORD

This issue begins with an article by McCarthy et al. describing the implementation of a molecular visualization model that was incorporated into two biochemistry courses. This inquiry-based activity explored the molecular basis and cultural relevance of sickle cell anemia. The article describes the activity and then provides an analysis of its impacts on student engagement and cultural awareness.

The article by Gordon and Cahill describes a recent survey of undergraduate programs in computational science. The results indicate that such programs face several challenges including student recruitment and faculty participation. They discuss the challenges that the programs face and some possible short- and long-term strategies that might address the challenges.

We also have three student articles in this issue. Those articles summarize the results of student internship experiences and the impacts of those experiences on the students' academic career.

Guevara et al. describe the development of a molecular model of relating to the interaction of polymer metal interfaces in micropumps. Their model simulates the sealant material polydimethylsiloxane and characterize its behavior with a model Ni-Md-Ga surface.

McDonald and Mercer used the Blue Waters supercomputer to assess tornadic outbreak forecast capability by lead time. They tested several parameter sets to ascertain which provided the best lead time forecasts of potential tornadic events.

Finally, Kellas and Groves tested several approaches to using an evolutionary algorithm to simulate lowest energy conformation molecule of a given stoichiometry. They compare the results using a clustering algorithm and intuitive population creation.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright ©JOCSE, a supported publication of the Shodor Education Foundation Inc.

© 2020 Journal of Computational Science Education

# Using Molecular Visualization as a Tool for Culturally Competent and Culturally Relevant Teaching: A Guided-Inquiry Biochemistry Activity

Pumtiwitt McCarthy  
Department of Chemistry  
Morgan State University  
Baltimore, MD  
Pumtiwitt.McCarthy@morgan.edu

Richard Williams  
Department of Chemistry  
Morgan State University  
Baltimore, MD  
rjwilms3@gmail.com

Cleo Hughes-Darden et al.<sup>1</sup>  
Department of Biology  
Morgan State University  
Baltimore, MD  
Cleo.HughesDarden@morgan.edu

## ABSTRACT

The central dogma is a key foundational concept in biochemistry. The idea that DNA mutations cause change at the protein level can be abstract for students. To provide a real-world example of the effect of mutation on protein function, a molecular visualization module was developed and incorporated into two biochemistry courses. This inquiry-based activity explored the molecular basis and cultural relevance of sickle cell anemia. Hemoglobin structural changes from the disease were examined. Participants used free tools including NCBI, RCSB PDB, LALIGN and Swiss PDB DeepView protein visualization software from EXPASY. This module was an active, engaging exercise which exposed students to protein visualization and increased cultural awareness.

## Keywords

Biochemistry, Computer-Based Learning, Inquiry-Based/Discovery Learning, Internet/Web-Based Learning, Proteins/Peptides, Underrepresented Minorities, Culturally-Relevant Pedagogy

## 1. INTRODUCTION

Effective teaching of biochemistry often uses computational tools for students to visualize and analyze macromolecular structure [1, 2]. Numerous science education articles describe addition of bioinformatics to the chemistry/biochemistry curriculum [3]. Many exercises use free stand-alone or web-based tools [4] making incorporation of bioinformatics achievable. Students used bioinformatics to investigate multiple topics such as drug design [5, 6, 7, 8, 9], visualization of protein structure [10, 11, 12, 13, 14], homology modeling [15, 16, 17], genomics/proteomics [18, 19] and genetic diseases [20]. The work described here is distinct because students use bioinformatics to investigate a genetic disease and explores the racial disproportionality of this disease. The goal of this activity was to provide students an engaging experience in protein visualization in a culturally relevant way.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright ©JOCSE, a supported publication of the Shodor Education Foundation Inc.

© 2020 Journal of Computational Science Education  
DOI: <https://doi.org/10.22369/issn.2153-4136/11/2/1>

The term culturally relevant or culturally responsive pedagogy was first described over twenty years ago [21, 22]. In short, this describes a teaching approach that embraces students' different cultural perspectives to further the learning experience. One aspect of this pedagogy is cultural competence. Culturally competent pedagogy allows a teacher to impart knowledge that allows students' to appreciate their own culture and learn about the culture of others [23]. Sickle cell anemia afflicts African-Americans disproportionately. This activity provides an opportunity for a diverse set of students to research the epidemiology of the disease thereby increasing cultural awareness [24].

Genetic diseases arise from changes in DNA which leads to changes at the protein level [25]. Sickle cell anemia is caused by a point mutation in the gene message encoding hemoglobin [26]. The most prevalent mutation is a change of glutamic acid to valine in the beta subunit. Hydrophobic "sticky" patches are produced that associate with the beta subunit of other hemoglobin molecules [27]. One result of this aggregation is a decrease in iron binding and fibril formation leading to sickle-shaped red blood cells [28]. The DNA, mRNA and amino acid sequence of both types of hemoglobin are known. This activity uses free bioinformatics software and tools (**Table 1**) to increase student engagement with computational analysis and cultural awareness.

## 2. ACTIVITY INFORMATION

### 2.1 Software Used

Freely available software and tools were used in this activity. These are described in **Table 1**.

### 2.2 Learning Goals

There are five student learning goals each with specific outcomes. **Learning Goal 1.** Students will understand the relationship between DNA, RNA and protein sequence. Outcomes from Goal 1. If given an RNA sequence, students should be able to: a) derive the original strands of DNA; b) manually translate into amino acid sequence; use bioinformatic tools to translate into amino acid sequence. **Learning Goal 2.** Students will understand the effects of amino acid change on protein structure and function. Outcomes from Goal 2. Students should be able to: a) compare two amino acid sequences; b) classify and predict the severity of amino acid substitutions based on the chemical properties of amino acids; c) to highlight and label amino acids at both termini of a protein. **Learning Goal 3.** Students will understand how to use molecular visualization software to display and modify protein structures.

<sup>1</sup> Additional authors under "Author List"



Outcomes from Goal 3. Students should be able to: a) highlight and label amino acids that interact with a ligand during binding; b) to report the racial incidence of sickle cell anemia. **Learning Goal 4.** Students will understand the cultural relevance of sickle cell anemia. Outcomes from Goal 4. Students should be able to: report the global incidence of sickle cell anemia. Summative assessments were used for each outcome with specific problems within the assignment.

**Table 1. Software and tools used in activity [29, 30, 31]**

Name	Provided by	Address	Function
NCBI	NIH National Library of Medicine	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>	-Provides access to biomedical and genomic data
RCSB PDB	The Research Collaboratory for Structural Bioinformatics	<a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a>	-Repository of protein structural data
LALIGN	ExPASy SIB Bioinformatics Resource Portal	<a href="http://embnet.vital-it.ch/software/LALIGN_form.html">http://embnet.vital-it.ch/software/LALIGN_form.html</a>	-Aligns two sequences to determine matching segments
DeepView	ExPASy SIB Bioinformatics Resource Portal	<a href="http://spdbv.vital-it.ch/refs.html">http://spdbv.vital-it.ch/refs.html</a>	-Protein visualization software
Translate	ExPASy SIB Bioinformatics Resource Portal	<a href="https://web.expasy.org/translate/">https://web.expasy.org/translate/</a>	-Translates nucleotide sequence to protein sequence
Google Docs	Google	<a href="http://docs.google.com">http://docs.google.com</a>	-Web-based word processing tool

### 2.3 Tenets of Culturally Sensitive Teaching Adopted

This activity was developed as one part of a wider inter-department goal of increasing culturally sensitive teaching in STEM courses [32]. A cohort of faculty from the departments of Mathematics, Computer Science, Chemistry, Biology, Physics, Industrial Engineering and Information Sciences and Systems developed five tenets for culturally-sensitive pedagogy in our courses. These were: 1) to incorporate physical and hands on activities in instructional practice, 2) to incorporate physical and hands on activities in instructional activities, 3) to become conscious of biased judgements about students based on limited perceptions of them and be willing to change these perceptions, 4) to have student apprenticeships that foster empowered learning communities and 5) to use students' lived experiences as content for course content and activities.

### 2.4 Activity Information

#### 2.4.1 Participants and Materials

Students in CHEM 202 Biochemistry for Health Majors are second-year undergraduate Medical Technology majors and upper-division undergraduate Nutrition majors. Students in CHEM 304 Biochemistry are upper-division level Chemistry and Biology majors. These students anticipate entering graduate or medical school after graduation. Activities were performed during consecutive laboratory periods. CHEM 202 has 2 hr 50 min allotted for lab each week while CHEM 304 has 3 hr 50 min allotted.

Ten laptop computers were provided by the Chemistry Department. Groups of 2-4 students performed the activity. One computer was provided per group.

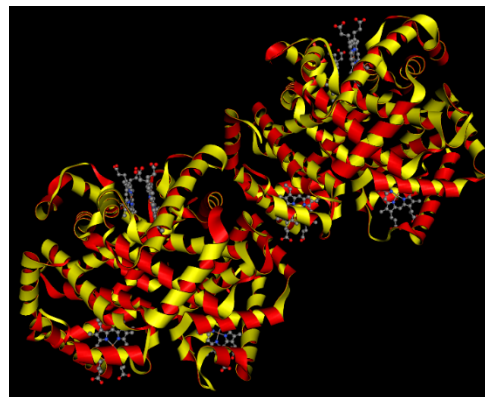
#### 2.4.2 Pre-laboratory Activities

Students were assigned a pre-lab activity about the Protein Data Bank (PDB) and hemoglobin structure and function. A YouTube video [32a] about the PDB was made available on the online course management system. An article from the Molecule of the Month on the PDB website [33] about hemoglobin and sickle-cell anemia was assigned and students answered instructor-supplied questions.

#### 2.4.3 Overview of Laboratory Activity

Modules were performed over two course laboratory periods. In the first week, students learned crystal structure basics and how to use computer software to visualize them. Targeted questions discuss cultural relevance and impact. The assignment called for students to download the crystal structure of normal hemoglobin from the PDB (PDB ID: 1hho [34]) and visualize its structure in Swiss DeepView. Next, they examined the secondary structure elements present in the protein. Students identified and labeled the N- and C-termini and the histidine residues that coordinate the heme group. The amino acid position that will be changed in the sickle cell variant was also identified and labeled by students. At the end of the exercise students are asked to answer questions related to rate of occurrence, racial incidence, symptoms and cure for the disease.

In the second week, students compare the 3-dimensional structure of sickle cell hemoglobin (PDB ID: 2hbs [35], **Figure 1**) to normal.



**Figure 1. Crystal Structure of Sickle Cell Hemoglobin (2hbs)**

They also engage with the central dogma by analyzing the DNA, RNA and amino acid sequence manually and with online bioinformatics software. As before, students use Swiss DeepView to label pertinent histidine residues. Students are also asked to discuss changes in the crystal structure between the normal and sickle hemoglobin. A portion of mRNA from wild type and sickle cell hemoglobin are given to students to translate into amino acid sequence manually. They also derive the original DNA sequence for both variants of hemoglobin. Next, the entire mRNA sequence of both types of hemoglobin is provided and online tool is used to translate sequences. LALIGN was used to compare similarity. All answers went into a Google document shared with group members and the instructor.

### 3. RESULTS AND DISCUSSION

#### 3.1 Impact of Activity on Student

##### Engagement and Cultural Awareness

This activity was developed over two summers and implemented in Spring 2016 for CHEM 202. It was used in both courses the following academic year. Over 70 students have been exposed to this activity over this time. This activity is student-centered with very little instructor interaction. The instructor offers a brief introduction to the activity and the students begin following the step-by-step instructions for each week. When questions arise, the instructor clarifies for students. They also have access to the internet during this time. This activity was successful in engaging students. During implementation of the module in CHEM 202 in Spring 2016, a team of evaluators observed the class and scored the activity on the level of student engagement and use of computational tools. The scale runs from 1 (lowest) to 7 (highest). Evaluators found that the module scored high on all aspects. Specifically, the module “incorporated engaging team-based, real world projects” (score = 7), “used computational tools for modeling and simulations” (score = 7), students “understood why they were using computational tools” (score = 6) and “student’s interest in the class” (score = 7).

Sickle cell anemia was chosen for its cultural relevance and because students were introduced to it earlier in the semester. Morgan State University is becoming increasingly diverse. Morgan State University is a Historically Black College or University (HBCU) located in Baltimore, MD. In 2016-2017, 22 participants identified themselves as African-American (64%), 6 as Caucasian (18%), 5 were international (15%), and 1 was multiracial (3%). Therefore, many cultural backgrounds learned about this disease and its incidence.

#### 3.2 Alignment with Developed Tenets of Culturally Sensitive Teaching

This exercise was developed keeping five main culturally-sensitive goals in mind. These were 1) to incorporate physical and hands on activities in instructional practice, 2) to incorporate more student-led discussions and teaching opportunities in class, 3) to becoming conscious of biased judgments about students based on limited perceptions of them and be willing to change these perceptions, 4) to have student apprenticeships that foster empowered learning communities and 5) to use students’ lived experiences as content for course content and activities. All five of these goals were achieved. This was primarily a student-led protein visualization project. There was a very brief introduction to the students by the instructor about what they will be doing, but the students discussed within their group and across groups how best to complete assignments (Tenets 1, 2, and 4). All students were treated in an equitable fashion and resources were made available to all students (Tenet 3). During the activity, one student was also able to offer a personal perspective as they had a family member that has the disease (Tenet 5).

#### 3.3 Activity Outcomes

At implementation of this activity, students have already learned about the different levels of protein structure and are starting to learn the details of the central dogma. Hemoglobin and sickle cell hemoglobin had been discussed. Relevant concepts such as quaternary structure, protein-ligand binding and allosterism had also been discussed. This activity was designed to integrate early course material with new material that they are learning.

Learning goals and outcomes are described in Section 2.2. In the Bloom’s taxonomy hierarchy [36], these integrated questions range from lower level “knowledge” to higher level “evaluation” and “synthesis”. One example of an integrated question asks students to look at the amino acid sequences of the proteins and label the N- and C-termini of each subunit and determine whether the structure is the oxygen-bound form. Students must recall their knowledge of amino acid codes, protein directionality, and hemoglobin subunits. An example of an “evaluation” type is when students are asked to translate both types of hemoglobin into amino acid sequence. During alignment of the two sequences students must determine which alignments are correct and the number of sequence positions where they differ. Overall, the module successfully introduced students to protein visualization, reinforced knowledge of the central dogma and increased cultural awareness.

### 4. CONCLUSIONS AND FUTURE WORK

Genetic diseases can serve as tractable examples of the central dogma’s importance. A mutation in sickle cell anemia leads to hemoglobin structural changes that students were able to investigate using free resources. This module explored the molecular basis and cultural relevance of sickle cell anemia to help increase cultural competency. Future versions of this module will better assess student engagement with protein visualization.

Student handouts with step-by-step instructions and laboratory questions, Student pre-lab activity; mRNA sequence of normal and sickle-cell hemoglobin chain A, and example student data are available from the authors by email request to [Pumtiwitt.McCarthy@morgan.edu](mailto:Pumtiwitt.McCarthy@morgan.edu).

### 5. AUTHOR LIST

Additional authors in order are: Roni Ellington<sup>2</sup>, Paminas Mayaka<sup>2</sup>, Monica Jackson<sup>3</sup> and Asamoah Nkwanta<sup>4</sup>

<sup>2</sup>Advanced Studies, Leadership, and Policy, Morgan State University, 1700 East Cold Spring Lane, Baltimore, MD 21251,

<sup>3</sup>Department of Mathematics and Statistics, American University, 4400 Massachusetts Avenue, NW, Washington, DC 20016,

<sup>4</sup>Department of Mathematics, Morgan State University.

### 6. ACKNOWLEDGMENTS

The authors were partially supported by the Association of American Colleges & Universities, Project Kaleidoscope (PKAL) through a Teaching to Increase Diversity and Equity in STEM (TIDES) grant. None of the authors have a financial, personal, or professional conflict of interest related to this work.

### 7. REFERENCES

- [1] Jaswal, S. S., O'Hara, P. B., Williamson, P. L. and Springer, A. L. Teaching structure: Student use of software tools for understanding macromolecular structure in an undergraduate biochemistry course. *Biochemistry and Molecular Biology Education*, 41, 5 (2013), 351-359.
- [2] Schonborn, K. J. and Anderson, T. R. The importance of visual literacy in the education of biochemists\*. *Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology*, 34, 2 (Mar 2006), 94-102.

- [3] Magana, A. J., Taleyarkhan, M., Alvarado, D. R., Kane, M., Springer, J. and Clase, K. A Survey of Scholarly Literature Describing the Field of Bioinformatics Education and Bioinformatics Educational Research. *CBE life sciences education*, 13, 4 (Winter 2014), 607-623.
- [4] Craig, P. A., Michel, L. V. and Bateman, R. C. A survey of educational uses of molecular visualization freeware. *Biochemistry and Molecular Biology Education*, 41, 3 (2013), 193-205.
- [5] Hayes, J. M. An Integrated Visualization and Basic Molecular Modeling Laboratory for First-Year Undergraduate Medicinal Chemistry. *Journal of Chemical Education*, 91, 6 (2014/06/10 2014), 919-923.
- [6] Peterson, R. R. and Cox, J. R. Integrating Computational Chemistry into a Project-Oriented Biochemistry Laboratory Experience: A New Twist on the Lysozyme Experiment. *Journal of Chemical Education*, 78, 11 (2001/11/01 2001), 1551.
- [7] Rodrigues, R. P., Andrade, S. F., Mantoani, S. P., Eifler-Lima, V. L., Silva, V. B. and Kawano, D. F. Using Free Computational Resources To Illustrate the Drug Design Process in an Undergraduate Medicinal Chemistry Course. *Journal of Chemical Education*, 92, 5 (2015/05/12 2015), 827-835.
- [8] Swope, N. K., Fryfogle, P. J. and Sivy, T. L. Detection of the cp4 epsps Gene in Maize Line NK603 and Comparison of Related Protein Structures: An Advanced Undergraduate Experiment. *Journal of Chemical Education*, 92, 7 (2015/07/14 2015), 1229-1232.
- [9] Tsai, C. S. Using Computer Applications and Online Resources To Teach and Learn Pharmaceutical Chemistry. *Journal of Chemical Education*, 84, 12 (2007/12/01 2007), 2019.
- [10] Cox, J. R. Teaching Noncovalent Interactions in the Biochemistry Curriculum through Molecular Visualization: The Search for pi Interactions. *Journal of Chemical Education*, 77, 11 (2000/11/01 2000), 1424.
- [11] Lowery, M. S. and Plesniak, L. A. Some Like It Cold: A Computer-Based Laboratory Introduction to Sequence and Tertiary Structure Comparison of Cold-Adapted Lactate Dehydrogenases Using Bioinformatics Tools. *Journal of Chemical Education*, 80, 11 (2003/11/01 2003), 1300.
- [12] McLaughlin, K. J. Understanding Structure: A Computer-Based Macromolecular Biochemistry Lab Activity. *Journal of Chemical Education*, 94, 7 (2017/07/11 2017), 903-906.
- [13] Rowe, L. Green Fluorescent Protein-Focused Bioinformatics Laboratory Experiment Suitable for Undergraduates in Biochemistry Courses. *Journal of Chemical Education*, 94, 5 (2017/05/09 2017), 650-655.
- [14] Ship, N. J. and Zamble, D. B. Analyzing the 3D Structure of Human Carbonic Anhydrase II and Its Mutants Using Deep View and the Protein Data Bank. *Journal of Chemical Education*, 82, 12 (2005/12/01 2005), 1805.
- [15] León, D., Uridil, S. and Miranda, J. Structural Analysis and Modeling of Proteins on the Web: An Investigation for Biochemistry Undergraduates. *Journal of Chemical Education*, 75, 6 (1998/06/01 1998), 731.
- [16] Petrović, D. and Zlatović, M. Modeling Human Serum Albumin Tertiary Structure To Teach Upper-Division Chemistry Students Bioinformatics and Homology Modeling Basics. *Journal of Chemical Education*, 92, 7 (2015/07/14 2015), 1233-1237.
- [17] Tsai, C. S. A Computer-Assisted Tutorial on Protein Structure. *Journal of Chemical Education*, 78, 6 (2001/06/01 2001), 837.
- [18] Ditty, J. L., Kvaal, C. A., Goodner, B., Freyermuth, S. K., Bailey, C., Britton, R. A., Gordon, S. G., Heinhorst, S., Reed, K., Xu, Z., Sanders-Lorenz, E. R., Axen, S., Kim, E., Johns, M., Scott, K. and Kerfeld, C. A. Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS biology*, 8, 8 (Aug 10 2010), e1000448.
- [19] Kossida, S., Tahri, N. and Daizadeh, I. Bioinformatics by Example: From Sequence to Target. *Journal of Chemical Education*, 79, 12 (2002/12/01 2002), 1480.
- [20] Schneider, T. L. and Linton, B. R. Introduction to Protein Structure through Genetic Diseases. *Journal of Chemical Education*, 85, 5 (2008/05/01 2008), 662.
- [21] Ladson-Billings, G. Reading between the Lines and beyond the Pages: A Culturally Relevant Approach to Literacy Teaching. *Theory Into Practice*, 31, 4 (1992), 312-320.
- [22] Ladson-Billings, G. Toward a Theory of Culturally Relevant Pedagogy. *American Educational Research Journal*, 32, 3 (1995), 465-491.
- [23] Ladson-Billings, G. Culturally Relevant Pedagogy 2.0: a.k.a. the Remix. *Harvard Educational Review*, 84, 1 (2014), 74-84.
- [24] Ojodu, J., Hulihan, M. M., Pope, S. N. and Grant, A. M. Incidence of sickle cell trait--United States, 2010. *MMWR. Morbidity and mortality weekly report*, 63, 49 (Dec 12 2014), 1155-1158.
- [25] Crick, F. Central Dogma of Molecular Biology. *Nature*, 227, 5258 (08/08/print 1970), 561-563.
- [26] Piel, F. B., Steinberg, M. H. and Rees, D. C. Sickle Cell Disease. *The New England journal of medicine*, 376, 16 (Apr 20 2017), 1561-1573.
- [27] Chou, P. Y.  $\beta$ -Sheet aggregation proposed in sickle cell hemoglobin. *Biochemical and Biophysical Research Communications*, 61, 1 (1974/11/06/ 1974), 87-94.
- [28] Finch, J. T., Perutz, M. F., Bertles, J. F. and Dobler, J. Structure of sickled erythrocytes and of sickle-cell hemoglobin fibers. *Proceedings of the National Academy of Sciences of the United States of America*, 70, 3 (Mar 1973), 718-722.
- [29] Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I. and Stockinger, H. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res*, 40, Web Server issue (Jul 2012), W597-603.

- [30] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research*, 28, 1 (2000), 235-242.
- [31] Guex, N. and Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18, 15 (Dec 1997), 2714-2723.
- [32] Hughes-Darden, C.; Ellington, R.M.; Zaveri, J.; Bapna, S.; Akil, L.; Hargett, S.; Bhattacharya, P.; Emdad, A.; Nkwanta, N. *Interventions Addressing Recruitment and Retention of Underrepresented Minority Groups in Undergraduate STEM Disciplines*. Emerald Publishing, Bingley, UK, 2019.[32a] [Alchemigus]. (2014, June 21). PDB RCSB Introduction/Tutorial [Video File]. Retrieved from: [https://youtu.be/\\_ok64nYRiCY](https://youtu.be/_ok64nYRiCY)
- [33] Goodsell, D. S., Dutta, S., Zardecki, C., Voigt, M., Berman, H. M. and Burley, S. K. The RCSB PDB "Molecule of the Month": Inspiring a Molecular View of Biology. *PLoS biology*, 13, 5 (May 2015), e1002140.
- [34] Shaanan, B. Structure of human oxyhaemoglobin at 2.1 Å resolution. *Journal of molecular biology*, 171, 1 (Nov 25 1983), 31-59.
- [35] Harrington, D. J., Adachi, K. and Royer, W. E., Jr. The high resolution crystal structure of deoxyhemoglobin S. *Journal of molecular biology*, 272, 3 (Sep 26 1997), 398-407.
- [36] Crowe, A., Dirks, C. and Wenderoth, M. P. Biology in bloom: implementing Bloom's Taxonomy to enhance student learning in biology. *CBE life sciences education*, 7, 4 (Winter 2008), 368-381.

# The State of Undergraduate Computational Science Programs

Steven I. Gordon  
The Ohio State University  
Columbus, OH  
gordon.1@osu.edu

Katharine Cahill  
Ohio Supercomputer Center  
Columbus, OH  
kcahill@osc.edu

## ABSTRACT

A number of efforts have been made to introduce computational science in the undergraduate curriculum. We describe a survey of the undergraduate computational science programs in the U.S. The programs face several challenges including student recruitment and limited faculty participation in the programs. We describe the current state of the programs, discuss the problems they face, and discuss potential short- and long-range strategies that might address those challenges.

## Keywords

Computational science and engineering education; modeling and simulation; model curricula; undergraduate computational science education

## 1. INTRODUCTION

Modeling and simulation has become an integral part in the advancement of knowledge in science and engineering along with theory and experimentation. Computer modeling allows for the exploration of systems that are too complex, too large or sensitive for experiments, or too small to instrument. A majority of large companies in the U.S. use modeling and simulation to produce goods faster and cheaper. The ability to use this technology is essential to commercial competitiveness. Recognition of the importance of computational modeling has led to a widespread call to educate students on the principals of modeling and simulation and the use of computational tools and algorithms to address those modeling needs.

SIAM (The Society for Industrial and Applied Mathematics) formed a working group on computational science in 1998. Subsequently, a seminal article by Yasar and Landau provided one of the first, comprehensive descriptions of the nature of the field and curricular elements necessary to provide students with the appropriate expertise [1]. A further SIAM task force completed a comprehensive report in 2006 [2].

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright ©JOCSE, a supported publication of the Shodor Education Foundation Inc.

© 2020 Journal of Computational Science Education  
DOI: <https://doi.org/10.22369/issn.2153-4136/11/2/2>

There have also been a number of national science and technology groups that have cited computational science as a key to future discoveries in science and engineering as well as crucial to the competitiveness of US industry.

In 2005, the President's Information Technology Advisory Committee highlighted the importance of computational science to the national economy and cited the lack of qualified personnel to fill the needs of both research and commercial enterprises [3].

The National Science Foundation Blue Ribbon Panel on Simulation-Based Engineering Science indicated that this discipline is "central to advances in biomedicine, nanomanufacturing, homeland security, microelectronics, energy and environmental sciences, advanced materials, and product development." [4] They went on to say that the education of engineers and scientists in the use of simulation techniques is a major challenge.

There have been many efforts to insert computational science into the undergraduate curriculum in an attempt to meet these needs. The National Computational Science Institute (NCSI), developed by the Shodor Education Foundation aimed to develop a national community of faculty interested in incorporating computational science into their undergraduate curriculum [5]. Thomley and Searcy provide a brief overview of this effort along with a comprehensive review of the history of computational science education [6].

Searcy and Thomley [7] completed an evaluation of the Shodor program which points to a number of barriers to the implementation of new academic programs. They surveyed 768 individuals that attended the NCSI workshops. The respondents reported a number of issues with implementing computational science into their courses.

"Between one quarter and one half highlighted the following issues: staying current with changes in technology (49 %), deciding where to make a big shift in their department's curriculum (43 %), lack of available computational science educational materials (39 %), making choices of which technologies/software to use (35 %), having no one else to discuss computational science ideas within their department (32 %), implementing computational science in the face of indifference from other faculty in their department (31 %), trying to incorporate other disciplines' content into a course (31 %), lack of understanding of how software package(s) work (28 %), trying to coordinate content across multiple professors and/or multiple sections of a course (26 %), and trying to coordinate programmatic changes across departments (26 %)." [7 page 3].

Other efforts at integrating computational science in the curriculum have focused on the introduction of formal emphasis or minor programs that include four to six courses focused on the tools and techniques used in the field. Gordon, Carey, and Vakalis [8] review some of these efforts. They then summarize their efforts to start computational science programs at multiple institutions in Ohio. Those efforts included the creation of a set of competencies for undergraduate students in computational science. Those competencies have been updated and are part of the efforts of the XSEDE education program to help other institutions start computational science programs [9].

A number of grants by the National Science Foundation and other institutions have supported efforts such as those cited above. Yet, the number of formal undergraduate computational science programs has grown very slowly. A recent web based search for such programs yielded a list of only 29 programs in the U.S. [10]. The program links on that site were checked and a further search conducted by the authors to ensure that the list is up-to-date.

There are many questions related to the state of these programs. How have the programs fared in producing graduates with computational science knowledge and skills? Are there continuing barriers to the integration of these important skills in the undergraduate curriculum? What are the institutional, personnel, and resource issues that have contributed to the success and/or limitations on the programs? What sorts of institutional or environmental changes that might help to scale up the programs? In order to address these questions, we conducted a survey of the existing programs focusing on the current state of their efforts as well as continuing barriers to program implementation. That is the subject of this paper.

## 2. THE SURVEY

For each of the undergraduate programs in the updated web list, contact information for the program advisors was assembled either as indicated on a program webpage or via a phone call to the appropriate person. For the 29 programs at 26 institutions, a survey of 11 questions addressing some of the motivations and barriers to program implementation was assembled and sent via an email link to the program lead. Of those, ten programs responded.

The survey questions were based on the previous work by Thomley and Searcy as well as informal discussions with program coordinators as part of the work on the XSEDE education program. The survey was distributed via email using the Qualtrics survey tools. A copy of the basic survey can be found in Appendix 1. Email reminders were sent to non-respondents weekly over a one month period.

Follow-up in-depth interviews were also made with three of the respondents to gain additional insight into the state of their programs. Two of the programs selected for interviews were of long-standing and were selected to provide insights into the barriers to program implementation as well as possible changes that would enhance program success. The third is a program that was recently started, hoping to gain insights into continuing motivations and barriers to program initialization. Those interviews were conducted by telephone and guided by a series of the follow-up questions. Those questions are also shown in the appendix. Those interviews were more open-ended, asking for the broader opinions of the program directors.

## 3. SURVEY RESULTS

### 3.1 Nature of the Programs

Of the 29 programs, 10 (34%) responded to the survey. The responding programs surveyed varied in their composition (4 department level programs, 3 college-wide, and 2 university-level programs) and all have been active for more than 5 years.

Table 1 shows the distribution of respondents and non-respondents. Those responding appear a reasonable representation of the population across types of institutions, public or private, and Carnegie classification.

**Table 1. Comparison of Respondents and Non-Respondents**

	Carnegie Classification				
	Public	Private	R1/R2	M1/M2	Other
Responded	6	4	5	3	2
No Response	8	9	8	5	4

All of the responding programs are modest in size. The average number of students completing the program annually ranged from 1 (two programs) to 3-5 (four programs) to 15 (two programs). Administratively, two of the programs are university-wide programs. Three of the programs are college-wide while three others are departmentally based.

Most of the programs are marketed through announcements in basic, required courses. One depends entirely on a website for recruitment.

Respondents were asked what percentage of their students went on to graduate school or professional jobs. Several did not answer this question or answered inconsistently. Of those who responded, there was a range of 33 to 70 percent that go on to graduate school or an average of 41 percent. Similarly, an average of 44% go to professional jobs with a similar range. However, the inconsistency in the responses leads one to believe that there the institutions are not fully able to track what happens to their graduates.

### 3.2 Program Challenges

A series of questions focused on some of the program challenges that were cited in the study of the Shodor program. These are shown in Table 2. The first two questions focused on student recruitment. Here, the majority of respondents indicated that student recruitment is a major problem. Seventy-one percent of respondents strongly agree or somewhat agree that it is difficult to recruit students into their programs. Likewise, the same percentage somewhat disagree or strongly disagree that that they have little or no difficulty getting students into their program.

A second problem facing the programs is the burden of instruction. Here again, 71% of the respondents indicated that the burden for teaching courses falls to too few faculty.

Most programs found their students were prepared in math (78% strongly or somewhat agree) but less in programming (44% strongly or somewhat agree).

People were then asked to indicate the top three problems impacting their programs. Almost 30% indicated that recruiting students was a major problem. Next was engaging faculty in other departments to participate (22%), distribution of teaching loads (17%), and getting advisors to recommend the program to students early in their careers (17%).

**Table 2. Responses to Program Challenges Survey Questions**

	Strongly Agree	Somewhat Agree	Neither Agree Nor Disagree	Somewhat Disagree	Strongly Disagree
It is difficult to recruit students to enroll in our program.	4	2	2	0	1
We have little or no difficulty getting a sufficient number of students in our program courses.	1	1	0	4	3
The burden for teaching the courses at our institution falls to only a few faculty.	3	2	1	2	1
Students enrolling in our program have the required pre-requisite skills and knowledge in mathematics.	2	5	0	2	0
Students enrolling in our program have the required pre-requisite computing skills.	1	3	0	4	1

Follow-up telephone interviews were conducted with three programs. Two were programs of long standing while the third has just started their program. All three indicated that student recruitment was a major problem.

The telephone interviews focused on possible policies that might alleviate some of the barriers to program success. One suggestion that was discussed was a possible national effort to publicize the need for computational scientists like that given to computer science. The response was that this might be helpful, but it still may not address the overall problem faced while trying to recruit students: what is the job or career path that this program will prepare me for?

Another possible boon to student recruitment would be a university requirement that all science and engineering majors take an introductory modeling and simulation course. This, along with the need for an introductory programming course for non-computer science majors, would potentially increase student interest in computational science.

Addressing the issue of “what is the job,” the respondents were asked whether stronger ties with businesses that use computational science would assist in improving program numbers. Such ties would be welcomed but the one respondent to this question indicated that they did not have enough contacts or time to make those connections.

All of the telephone interviews echoed the problems of recruiting students. Indications were that some students still take one or two of the core computational science courses but do not complete the entire program.

#### 4. DISCUSSION

The response rate to the survey of 34% was disappointing, yet it is a decent response rate to an online survey. Nevertheless, we believe that those long-standing programs that did respond are emblematic of the problems facing computational science education. We can only surmise whether the non-respondents represent programs that are inactive, are run by faculty with too little time to make responding a priority, or some combination of these and other factors.

Based on the responses, it is clear that current, undergraduate computational science education efforts are making only modest progress in helping to build a workforce competent in this area. They continue to face ongoing problems in student recruitment, in the engagement of the full range of disciplines for which computational science is important, and the active engagement of

businesses that likewise are seeking graduates ready to contribute to their computational science endeavors.

Student recruitment efforts are foremost among the problems facing the existing programs. There appear to be a number of reasons for this. First, the interdisciplinary nature of the field makes it difficult to point to particular career paths associated with computational science. That confusion may also carry forward to academic advisors that may not fully understand what computational science is and therefore do not advise students to look into those programs early in their academic careers. Recruiting of students is also hampered by the fact that most of the programs are minors that require an additional 15-24 credit hours of additional courses. Students may take part of the sequence but do not complete the program. Should they start the program later in their academic career, they may not have enough time to complete all of the courses. At a time when the costs of higher education are high, delays in graduation pose a significant barrier for students to take on supplementary course work for a career path that is fuzzy at best.

Computational science continues to suffer from the limitations on the number of faculty that are prepared to participate in the programs. A combination of lack of expertise along with the required teaching loads for more traditional courses is probably to blame.

There are no simple solutions to these program impediments. One possible approach might be to introduce a university-wide course that introduces modeling and simulation to all students. This could be done without requiring pre-requisite programming expertise. For example, the University of California at Berkeley has developed a “data science for all” course for freshman students (11). A modeling and simulation course could be one of several alternatives in this vein, introducing students to the area early in their careers and promoting their continued interests.

If modeling and simulation expertise is truly required in the current workforce, then both business and the research community must also play a larger role in supporting program development. Businesses will need to more actively engage with academic institutions providing internships and help with student recruitment by publicizing the need for modeling and simulation skills. Grant programs that encourage the integration of computational science into the curriculum should be put in place in parallel with the efforts for computer science and data science.

Getting widespread participation for computational science across the faculty is probably the most difficult challenge. It may take a generational change in the faculty to fully address the problem. A much larger proportion of recent Ph.D. graduates in science and engineering are using modeling and simulation as part of their research and thus more likely to embrace the integration of those skills in the curriculum. Currently the science and engineering labor force is aging with 33% of the workforce in the ages between 51 and 75 years while only 16% was in the under the 30 age group (12). This implies a high rate of retirement and replacement in the coming years. Perhaps that will help the community to fully embrace the need for computational science expertise.

## 5. REFERENCES

- [1] Yasar, O. and Landau, R. H. (2003). Elements of computational science and engineering education. *Review*, 45(4), 787-805.
- [2] SIAM Working Group on CSE Undergraduate Education. (2006). Undergraduate Computational Science and Engineering Education. [https://www.siam.org/Portals/0/Publications/Reports/CSE\\_Report.pdf?ver=2018-03-16-161618-620](https://www.siam.org/Portals/0/Publications/Reports/CSE_Report.pdf?ver=2018-03-16-161618-620). Accessed November 19, 2018.
- [3] Reed, D., Bajcsy, R., Fernandez, M., Griffiths, J., Mott, R., Dongarra, J., Johnson, C., Inouye, A., Miner, W., Matzke, M., and Ponick, T. (2005). Computational Science: Ensuring America's Competitiveness.: The President's Information Technology Advisory Committee, National Coordination Office for Information Technology
- [4] SBES 2006. Simulation Based-Engineering Science: Report of the NSF Blue Ribbon Panel on Simulation-Based Engineering Science, NSF, February 2006. Available online at [https://www.nsf.gov/pubs/reports/sbes\\_final\\_report.pdf](https://www.nsf.gov/pubs/reports/sbes_final_report.pdf)
- [5] National Computational Science Institute. (n.d.). About NCSI. Shodor/NCSI web site: <http://www.computationalscience.org/about>. Accessed November 19, 2018.
- [6] Jill Tomley and Mary Searcy, (2009). Computational Science: Not Just for Researchers Anymore. *The International Journal of Science in Society*, 1(3), 27-42.
- [7] Searcy, ME and Thomley, JE. (2007). Barriers to interdisciplinary computational innovations in education. In Milková, E, and Prázák, P (Eds.), *Proceedings of the 8th International Conference on Technology in Mathematics Teaching*. Prague, Czech Republic: Faculty of Informatics and Management, University of Hradec Králové.
- [8] Gordon, S., Carey, K., and Vakalis, (2008). A shared, interinstitutional undergraduate minor program in computational science. *Computing in Science and Engineering*, 10(5), 12-16.
- [9] Minor Program in Computational Science ompetency/Topic Overview. <http://hpcuniversity.org/educators/undergradCompetencies/> Accessed May 28, 2019.
- [10] Undergraduate Programs in Computational Science. <http://hpcuniversity.org/students/undergraduatePrograms/>. Accessed on May 28, 2019.
- [11] Foundations of Data Science: A Data Science Course for Everyone <https://data.berkeley.edu/education/courses/data-8>. Accessed on May 28, 2019.
- [12] National Science Board. Science and Engineering Indicators 2018 (NSB-2018-1). January 2018. <https://www.nsf.gov/statistics/2018/nsb20181/figures/fig03-22>. Accessed on May 28, 2019.

## A. SURVEY INSTRUMENTS

### A.1 Undergraduate Computational Science Programs

Q1 Is your undergraduate computational science program still active?

- Yes (1)
- No (2)

Q2 How many credit hours are required for students to complete the program?

Q3 What is the average number of students that complete the program each year?

Q4 For each of the following questions, please indicate the degree to which you agree or disagree with the statement.

*(Strongly agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Strongly disagree)*

- It is difficult to recruit students to enroll in our program. (1)
- We have little or no difficulty getting a sufficient number of students in our program courses. (2)
- The burden for teaching the courses at our institution falls to only a few faculty. (3)
- Students enrolling in our program have the required prerequisite skills and knowledge in mathematics. (4)
- Students enrolling in our program have the required prerequisite computing skills. (5)

Q5 What proportion of students graduating from your program go to:

- Graduate school (1)
- Professional Jobs (2)
- Other (3)
- Not sure (4)

Q6 What do you think are the major problems associated with maintaining your program?

Q7 Administratively, where is your program located?

- University-wide program (1)
- College-wide program (2)
- Departmental program (3)

Q8 Please list all of the departments that play a role in teaching courses in your program.

Q9 How many years has your program been operating?

Q10 How do students find out about your program?

- Announced in various basic courses in related disciplines (1)
- Website (2)
- Program brochures (3)
- Listed in university catalog (4)
- Other (5)



Q11 Please choose what you see as the top three problems with your undergraduate computational science program.

- Difficulty recruiting students (1)
- Limited dedicated resources for program implementation (2)
- Distribution of teaching loads (3)
- Availability of relevant hardware and software (4)
- Getting advisors to recommend the program to their students early in their careers (5)
- Engaging faculty in other departments to participate (6)
- Other (7)

## A.2 Follow Up Questions for Respondents

We would like to thank you for responding to our survey about your computational science program. We would like to take a few minutes to follow-up with you on some questions that arose from the survey. Do you have a few minutes to speak now or can we set a time that is more convenient for you?

A number of problems associated with maintaining a program were cited by those completing the survey. We would like to get your thoughts on these problems and actions that might help to reduce them.

The first problem is cited is the difficulty of recruiting students. Is this a major problem for your program?

Which of these actions might help to alleviate that problem:

- National attention given to the need for scientists and engineers to understand modeling and simulation similar to that given recently to computer science?
- University requirement for introductory modeling and simulation class for all science majors
- An introductory computer coding class oriented for non-computer science majors
- Working with high schools to bring computational science into HS courses
- Other?

A second problem noted is the lack of resources from the university to offer courses and related limitations on the number of a faculty who can offer courses. What do you see as possible solutions to these problems?

Which of these might help:

- On-going funds the college or university to “buy” courses from other faculty to release them to teach a computational science course
- Funding for adjunct (is this the right word) faculty from industry to teach or co-teach some of the courses
- Sharing course instruction with other institutions using distance learning infrastructure

Several programs require an internship or research experience as part of their programs. Does your program have such a requirement? What do you see as problems managing this program? How might these problems be overcome?

- Deeper connections with businesses that use computational science to employ your students as interns
- A central database of internship opportunities at a national scale University program for undergraduate research opportunities

# Development of a Molecular Model for Understanding the Polymer-metal Interface in Solid State Pumps

Jaime D. Guevara

Micron School of Materials Science and Engineering  
Boise State University  
Boise, ID  
jaimeguevara@u.boisestate.edu

Peter Müllner

Micron School of Materials Science and Engineering  
Boise State University  
Boise, ID  
petermullner@boisestate.edu

Matthew L. Jones

Micron School of Materials Science and Engineering  
Boise State University  
Boise, ID  
MattyJones@boisestate.edu

Eric Jankowski\*

Micron School of Materials Science and Engineering  
Boise State University  
Boise, ID  
ericjankowski@boisestate.edu

## ABSTRACT

Medical micropumps that utilize Magnetic Shape Memory (MSM) alloys are small, powerful alternatives to conventional pumps because of their unique pumping mechanism. This mechanism—the transfer of fluid through the emulation of peristaltic contractions—is enabled by the magneto-mechanical properties of a shape memory alloy and a sealant material. Because the adhesion between the sealant and the alloy determines the performance of the pump and because the nature of this interface is not well characterized, an understanding of sealant-alloy interactions represents a fundamental component of engineering better solid state micropumps in particular, and metal-polymer interfaces in general. In this work we develop computational modeling techniques for investigating how the properties of sealant materials determine their adhesive properties with alloys. Specifically, we develop a molecular model of the sealant material polydimethylsiloxane (PDMS) and characterize its behavior with a model Ni-Mn-Ga surface. We perform equilibrium molecular dynamics simulations of the PDMS/Ni-Mn-Ga interface to iteratively improve the reliability, numerical stability, and accuracy of our models and the associated data workflow. To this end, we develop the first model for simulating PDMS/Ni-Mn-Ga interfaces by combining the Optimized Potentials for Liquid Simulations (OPLS) [21] force field with the Universal Force Field [5], and show promise for informing the design of more reliable MSM micropumps. We also reflect on the experiences of Blue Waters Supercomputing intern Guevara (the first author) to identify key learning moments during the one-year internship that can help guide future molecular simulation training efforts.

\*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright ©JOCSE, a supported publication of the Shodor Education Foundation Inc.

## KEYWORDS

molecular dynamics, organic polymers, coarse-graining, magnetic-shape memory alloys, magnetic-shape memory, materials science, micropumps, BWSIP

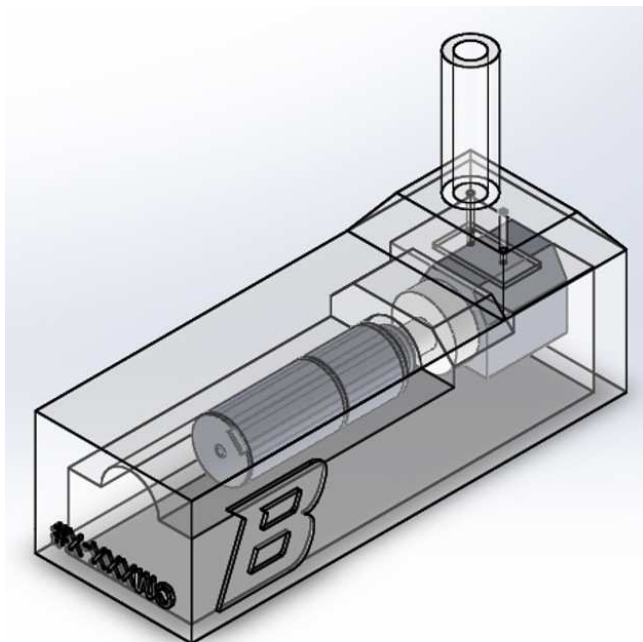
## 1 INTRODUCTION

MSM micropumps (Figure 1) represent a new paradigm of microfluidic mechanism, enabling accurate delivery of fluids over a wide range of densities and pressures [33]. MSM alloys including Ni-Mn-Ga enable such pumping through deformation under the influence of a magnetic field, which influences properties including twin motion deformation [38], strain [30], stress [29], magnetic and thermal activation [4, 38], operating temperatures [11, 35], magnetic permeability [22], and electric resistivity [32]. These properties make MSM alloys advantageous for use as actuators, channels, and pump membranes. In the case of MSM pumps, the aim is to exploit a localized constriction in the material (Figure 2) to encapsulate and propagate the working fluid against a sealant material as shown in Figures 2 and 3. Polydimethylsiloxane (PDMS) gel is a common sealant due to its low cost, bio-compatibility, moldability [8]. As the alloy is actuated it pulls away from the PDMS sealant, allowing fluid to be drawn in through the inlet port. Subsequently, as the fluid-filled constriction propagates toward the outlet due to the rotating magnetic field, the sealant re-adheres to the Ni-Mn-Ga surface, closing the inlet.

### 1.1 Simulations of Materials

The main focus of this manuscript is to advance understanding of PDMS/Ni-Mn-Ga interfaces by investigating the adhesion characteristics of the polymer sealant. Characterizing the nanoscale interface between PDMS and Ni-Mn-Ga (as seen in Figure 3) is challenging experimentally, so we use molecular simulations to explore PDMS/Ni-Mn-Ga adhesion. Molecular models for PDMS and Ni-Mn-Ga are individually available, but until this work no model exists that combines and describes interactions between both materials. Therefore, a focus of this work is creating this model.

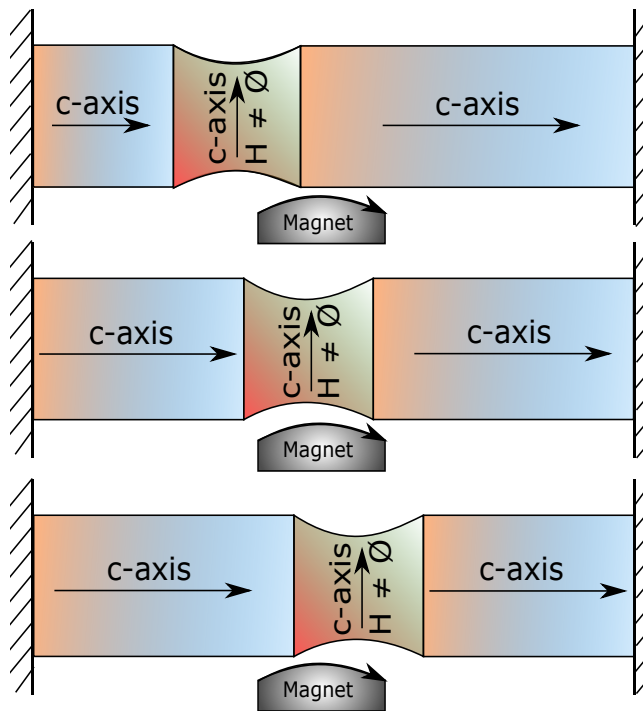
A key component of the model is the “force field” defined by potential energy functions describing bonded and non-bonded interactions between each type of particle in the system. Bonded



**Figure 1:** Model of a Ni-Mn-Ga micropump, showing placement of the MSM element, which is encapsulated in PDMS sealant, below the inlet/outlet reservoir.

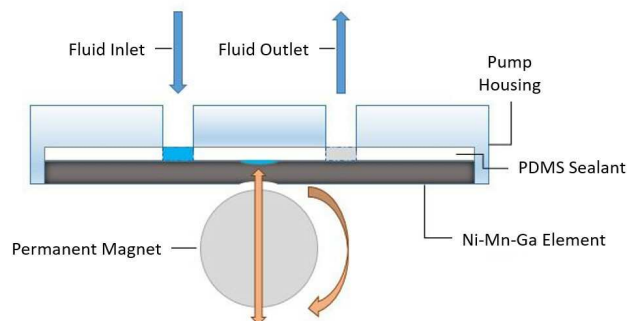
interactions are described as a set of constraints between bonded pairs (bond), triplets (angle) and quadruplets (dihedral) of particles that belong to the same molecule. Different forcefields vary in the functional forms used to model these constraints, so it is important to keep in mind any conversion factors when mapping from one forcefield to another and when combining multiple forcefields as we do here. Small differences in forcefield parameters can give rise to qualitatively different phase behavior and equilibrium structure [1].

We base our model on the Optimized Potentials for Liquid Simulations - United Atom model [21] (OPLS-UA) and UFF [5] force fields. There are several considerations to be made when constructing a new model. For instance, while some polymer models can describe how the conformation of a polymer chain evolves in a pristine thin film, varying the number of chains in the simulation, the average chain length, and polydispersity can strongly influence the interactions of the chains with themselves and other materials in the system [26]. Also, many alloy models are parameterized for a specific crystal structure and unit cell lattice parameters that are not necessarily transferable to other unit cells or crystallographic planes. Finally, it is important to ensure that the forcefields being combined are compatible with each other. As an example, the non-bonded Lennard-Jones interaction potentials used by Rappe *et al.* [5] in UFF are parameterized based on general hybridization rules, whereas Elliott and Akerson [7] rescale based on the bond dissociation energy, resulting in a thousand-times stronger interactions. While both approaches are self-consistent in that they correctly



**Figure 2:** Three side views of the MSM micropump element illustrating the translation of the constriction along the *c*-axis as the magnetic field (*H*) is rotated. In a MSM pump, fluid is held between this constriction and the sealant material.

describe the difference in interaction potentials between atoms defined within, naïvely combining the two would lead to the Elliott



**Figure 3:** Schematic of a MSM pump interface showing the pocket of fluid bounded by the MSM and PDMS in transit from the inlet to outlet port of the pump.

forcefield dominating. Care must therefore be taken when combining interactions from different forcefields in the same simulation volume and so we iteratively test our model as we proceed and compare to observations in the literature to validate our work.

## 2 METHODS

We perform molecular dynamics (MD) simulations using HOOMD-Blue [2, 10] on NVIDIA Tesla K20X Graphic Processing Units (GPUs) at the Blue Waters supercomputer at the National Center for Supercomputing Applications [31] and P100 GPUs on our local cluster “Fry”.

### 2.1 Molecular Dynamics

MD simulations can efficiently sample the equilibrium structure of molecular systems by iteratively calculating the forces between neighboring particles and numerically updating positions using velocity and force information over discretized time steps. Here we employ the velocity-Verlet algorithm [1] and a tree-based neighbor list [15] that provides performance benefits for the non-isotropic systems studied here.

We perform MD simulations in the canonical (constant number of simulation elements  $N$ , volume  $V$ , and temperature  $T$ ) ensemble. The number of time steps to relax to equilibrium and to subsequently sample equilibrium microstates must be determined empirically as they depend on the forcefield, ensemble, and initial conditions. The initial configuration is specified with an XML file, which describes all of the particle properties including position, mass, atom type, charge, and velocity. The force field is then used to calculate the total force acting on each particle due to the interactions with their neighbors, which becomes an acceleration of the particle according to Newton’s second law of motion,  $F = ma$ , which in turn updates the velocity of the particle for the time step, which in turn updates the particle’s position. After particle positions are updated, these steps are repeated for each subsequent time steps. Individual microstates (specifying instantaneous positions and velocities) are written to a trajectory file (here, every 500 steps) as the simulation progresses. The trajectory is useful for visualizing system evolution and for post-processing material properties. Instantaneous potential energy and kinetic temperature measurements are recorded to quantify equilibration and check for unphysical behavior.

### 2.2 Computational Resources

The theoretical peak performance of K20X GPUs is 1.22 TeraFLOPS (or  $1.22 \times 10^{15}$  floating-point operations per second) for double (64-bit) precision performance and 2.9 TeraFLOPS for single (32-bit) precision performance. P100 GPUs have a peak performance of 4.7 TeraFLOPS for double-precision performance and 9.3 TeraFLOPS for single precision performance. When referring to the performance of these GPUs, double and single precision refers to the number of bits used to represent each floating point number used in the calculations. Generally, for these types of molecular simulations, single point float precision is sufficient for numerical stability without compromising computational efficiency [2]. Each MD simulation performed herein used a single CPU core driving a single

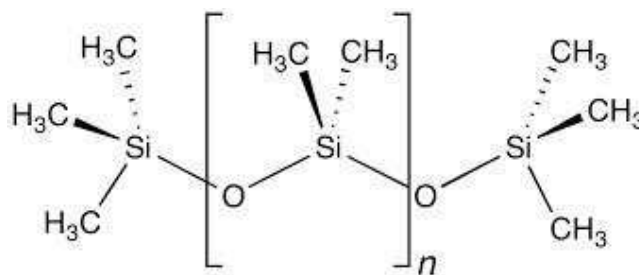
GPU, using Blue Waters’ K20X and Fry’s P100 hardware. Simulation benchmarking on both GPUs is presented in the Section 5.

## 3 MODEL

We specify the bonded and non-bonded PDMS interactions with the OPLS-UA force field. When running the surface interaction simulations with Ni-Mn-Ga we describe the polymer-surface interactions using UFF force field and polymer-polymer interactions with OPLS-UA. Surface-surface interactions are omitted from our investigation, and the Ni-Mn-Ga degrees of freedom are not integrated. This avoids modeling fast degrees of freedom of metallic bonds, which would require additional force fields such as the Embedded Atom Model [6] and more computational cost. Intrinsic to this constraint is the assumption that surface fluctuations are negligible on polymer fluctuation timescales. Here we describe the implementation details of the two material models individually and in combination.

### 3.1 Construction of PDMS

We construct PDMS topologies using the Avogadro [3] chemical drawing software. Using the tool, we draw the repeating monomer units (5 repeats shown in Figure 5), until reaching 20 (1.6 kDa), after which end caps (Figure 4 are added).



**Figure 4: Chemical formula for Polydimethylsiloxane (PDMS) [8].**

We then refine the 20-mer model from above in Jupyter notebook [24] that utilize mBuild [23] and Foyer [17], toolkits for managing molecule creation and forcefields. In Figure 5 and Figure 6 each sphere represents one United Atom (UA) simulation element. Each UA simulation element represents a “heavy” atom such as Carbon, Oxygen, or Silicon and its associated Hydrogens. This simplification accelerates sampling of the molecular dynamics without sacrificing structural accuracy [12, 27, 28].

PDMS has been studied extensively in the literature, and there is a wealth of information on possible simulation forcefields. For this investigation, we describe the non-bonded inter-molecular, and bonded intra-molecular interactions of PDMS using parameters obtained from the works of *Frischknecht and Curro* [9] and *Tamai et al* [36]. The complete set of parameters is given in Table 1.

The bonded interaction parameters are described as parameterizations of the following equations [9]:

$$V_b(r) = k_b(r - r_0)^2 \quad (1)$$

$$V_a(\theta) = k_\theta(\theta - \theta_0)^2 \quad (2)$$

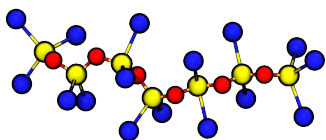


Figure 5: 5-mer chain of PDMS built using a Jupyter notebook, with energy minimized through Avogadro [3], and visualized with VMD [16]

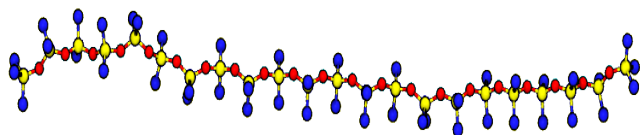


Figure 6: 20-mer chain of PDMS visualized with VMD [16]

$$V_t(\phi) = k_t[1 + \cos(n\phi)] \quad \text{and} \quad \phi = \pi \quad (3)$$

where  $V_b$ ,  $V_a$ , and  $V_t$  describe the pairwise bond, triplet angle, and quadruplet dihedral potentials, and are functions of separation,  $r$ , in-plane angle,  $\theta$ , and out-of-plane angle,  $\phi$ , respectively. These potentials are dependent on the constraint coefficients  $k_b$ ,  $k_\theta$ , and  $k_t$ , which are taken from the work of *Frischknecht and Curro*.

The non-bonded pairwise interactions were taken from *Tamai et al.*, describing the balance between short-range atomic repulsion and long-range Van der Waals attraction as a Lennard-Jones potential:

$$V(r) = -4\epsilon \left( \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right), \quad (4)$$

where  $\sigma$  and  $\epsilon$  control the shape of the interaction at varying separation  $r$ , and are dependent on the types of atoms in the pair being considered.

bonds	$r_o$ [Å]	$k_b$ [kJ/(mol nm <sup>2</sup> )]
Si-O	1.647	146490.2
Si-CH <sub>3</sub>	1.866	79337.0
angles	$\theta_o$ [rad]	$k_\theta$ [kJ/(mol rad <sup>2</sup> )]
Si-O-Si	2.547	59.162
O-Si-O	2.076	395.388
CH <sub>3</sub> -Si-CH <sub>3</sub>	1.844	209.074
O-Si-CH <sub>3</sub>	1.875	209.074
dihedrals	$n$	$k_t$ [kJ/mol]
Si-O-Si-O	1	1.8828
Si-O-Si-CH <sub>3</sub>	3	0.083638
nonbonded	$\sigma$ [Å]	$\epsilon$ [kJ/mol]
Si	3.385	2.4480
O	2.955	0.8493
CH <sub>3</sub>	3.786	0.7532
Ni	2.522	0.2510
Mn	2.635	0.2176
Ga	3.901	6.9454

Table 1: Potential Parameters for PDMS and Ni-Mn-Ga [5, 9, 36]

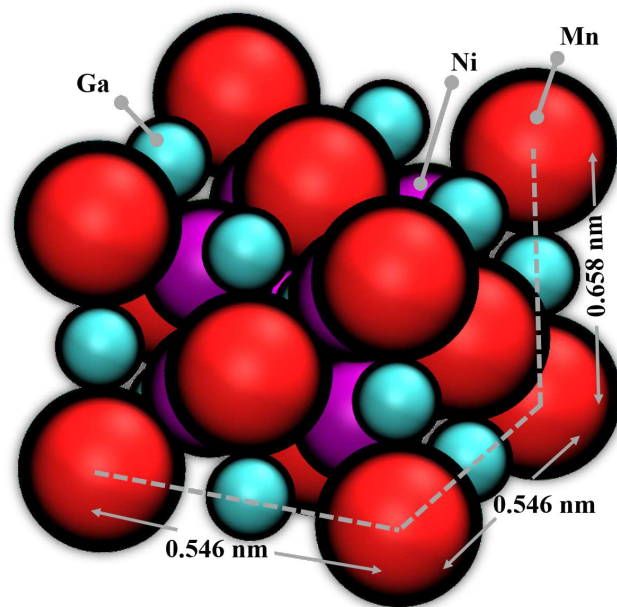


Figure 7: Unit cell of Ni-Mn-Ga with lattice parameters for a Non-Modulated (NM) crystal, visualized using VMD [16]

As per the molecular dynamics algorithm, interaction forces are calculated at each discrete time step ( $\delta t$ ). For our initial PDMS simulations, we use  $\delta t = 0.001$  fs. After calibrating our model, we found we could increase this time step without resulting in numerical inconsistencies. Therefore, for the final PDMS simulation and for all the interface simulations, we use  $\delta t = 1$  fs. The increase in simulation time step is beneficial as it permits a longer period of material simulation per minute of elapsed “wall-clock” time.

### 3.2 Construction of Ni-Mn-Ga

We use Avogadro [3] to construct the Ni-Mn-Ga surface, as it includes crystallography tools to position the atoms of the unit cell. We employ lattice constants

$$a = b = 0.546 \text{ nm}$$

$$c = 0.658 \text{ nm.}$$

as specified in *Sozinov et al.* [34].

The atomic coordinates themselves are taken from *Hickel et al.* [13]. The resulting unit cell of Ni-Mn-Ga, as shown in Figure 7, is then used as a building block for surfaces of varying sizes based on arranging multiple cells in a lattice, the size of which can be easily tuned as an input.

We employ UFF developed by *Rappé et al* [5] to provide interaction parameters for PDMS with Ni-Mn-Ga. However, because UFF lacks bonded constraint parameterizations for PDMS we must simultaneously employ OPLS-UA and UFF to simultaneously model surface-PDMS and PDMS-PDMS interactions. This is the first work to our knowledge to combine these force fields to investigate PDMS on Ni-Mn-Ga. The complete list of nonbonded interaction parameters for Ni-Mn-Ga used in our investigation can be found in Table 1.

## 4 DATA WORKFLOWS

Key open-source software packages used in this work include: mBuild [23], Foyer [17], MorphCT [18, 20], and Rhaco [19]. mBuild and Foyer are MD toolkits developed by the Molecular Simulation Design Framework team, aiding molecule building and atom typing, respectively, used here to create XML files input into HOOMD. MorphCT and Rhaco are used for checking periodic particle image information and initializing molecules on surfaces. As part of this work, Rhaco was generalized in order to accept a Ni-Mn-Ga unit cell as the surface template allowing us to initialize PDMS in its vicinity and to control the MD simulation of these two materials.

Therefore, the key contributions of this paper are the following:

- **MSM-Polymer Simulations** -

These are the first MD simulation of Ni-Mn-Ga with PDMS. This work also differs from prior MSM work that focuses on twin boundary movement within the alloy and the mechanisms by which this occurs. This paper reports the simulated interactions of this alloy with PDMS.

- **Educational Value** (J.Guevara) -

Working on this project has exposed me to many areas of both research and Computational Science, from learning how to gather the necessary data to begin simulations, to knowing how to collaborate with others. I gained significant experience in how to use Python for scripting, programming, and statistical analysis, as well as significant knowledge on the interaction of organic polymer and metal interfaces from a molecular standpoint. Furthermore, this experience has made me want to continue to pursue, and learn more about, High Performance Computing.

## 5 RESULTS

### 5.1 Performance

Initial PDMS simulations containing 7,120 particles took 45 minutes to complete  $10^7$  time steps on Blue Waters XK nodes (over 3,700 time steps per second (TPS)). These simulations informed additional, larger simulations on P100 GPUs on Fry. With P100 cards, we perform simulations of 36,900 particles at 3,600 TPS – a factor of 5 increase in system size with similar run time. In total, we submitted 60 pure (“neat”) PDMS jobs (25 on Blue Waters and 35 on Fry) and 14 jobs with PDMS/Ni-Mn-Ga interfaces (all on Fry) for a total of 33 hours and 14 minutes of wall clock time for simulations with analyzable trajectories. We estimate many hundreds of unsuccessful simulation jobs were submitted (unphysical initial conditions,  $\delta t$  too large, missing packages in the software stack, etc).

### 5.2 Neat PDMS

Neat PDMS systems containing 80 20-mer chains (7,120 particles) are simulated using HOOMD for  $10^5$  time steps at a temperature of  $T = 294$  K. However, for transferability of simulations and numerical accuracy, HOOMD records the temperature in reduced units,  $T^*$ , based on a reduced energy scaling parameter,  $\epsilon$ . For the initial PDMS volumes  $\epsilon = 0.585$  kcal/mol. The following equation can be used to map between the real and dimensionless temperatures:

$$T^* = \frac{k_B T}{\epsilon}, \quad (5)$$

where  $k_B$  is Boltzmann’s constant. We encountered some issues simulating our initial PDMS volumes, which are detailed in Section 7.1.

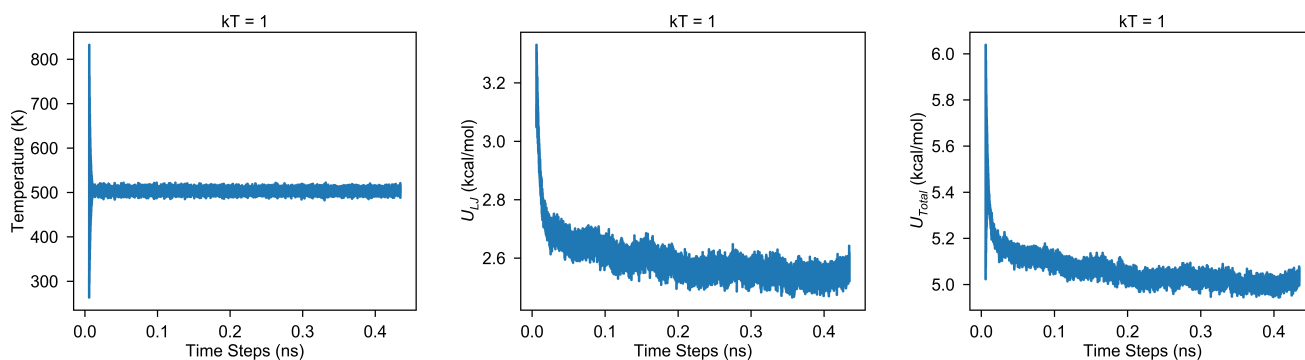
Figure 8 shows the simulation temperature, nonbonded potential energy, and total potential energy for our PDMS system after fixing issues corresponding to volume packing, image, and interaction parameter corrections. The kinetic temperature (8a) equilibrates quickly, with non-periodic temperature fluctuations of around 20-30K. Additionally, the potential energies presented in 8b and c exhibit a roughly-exponential relaxation as molecules reorganize into free energy minimizing configurations, followed by a period of constant average energy signifying equilibration of the system.

Another useful tool for analyzing the output of our simulation is the visualization software VMD (Visual Molecular Dynamics) [16]. VMD allows us to view the trajectories, helping to understand how the system evolves over time and what the final relaxed systems looks like. Visualizing the final trajectory can provide useful insights into any issues with the initialization of our simulations, or any bugs in the code, as described in Section 7.1.

### 5.3 PDMS with Ni-Mn-Ga Surface

Several issues needed to be addressed before equilibrated PDMS on Ni-Mn-Ga configurations could be sampled (Section 7.1). After fixing these bugs, we perform simulations of 100 PDMS chains in the presence of a  $20 \times 20 \times 1$  unit cell Ni-Mn-Ga slab at 294K. These simulations took around 40 seconds to simulate  $10^4$  time steps, and around 45 minutes to simulate  $10^7$  time steps - a similar wall-clock time to those noted in our initial neat PDMS tests. The resultant simulation trajectories after are shown in Figure 9. We also present the temperature and potential energy evolution of the simulations in Figure 10, demonstrating the small non-periodic fluctuations in both, and the eventual equilibration of the system according to the potential energy. That the kinetic temperature, nonbonded, and total energies all independently stabilize after  $\approx 2$  ns supports our interpretation that these systems reach equilibrium (or at least a metastable state). The PDMS is observed to quickly aggregate (within the first 50 frames of our visualization) and then sticks to the surface to form a film over the alloy layer.

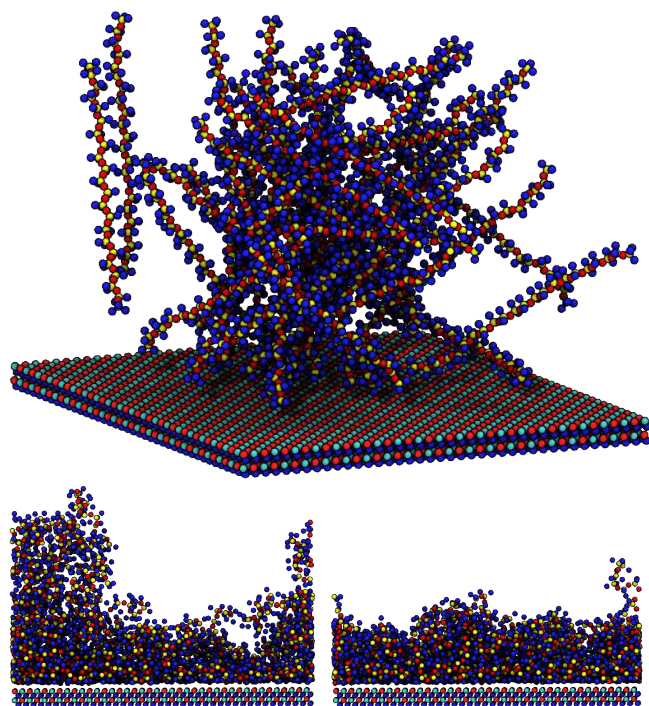
To test our simulation model and Rhaco further, we simulate the same system at a higher temperature: 500 K (approx.  $227^\circ\text{C}$ ). The increased thermal energy should translate to an increase in the speed at which the PDMS atoms move. Additionally, since the boiling point of PDMS is around 473 K, we would expect the vaporization of the polymer, leading to it filling the simulation volume rather than forming a liquid film at the alloy surface. Indeed, the visualization of the simulation shows the molecules interacting faster with each other, and Figure 11 shows the PDMS filling the simulation volume and adhering less to the alloy at 500 K compared to 294 K. Figure 11 also shows a rather visually drastic change in the PDMS, however the chains do stick together and seem far more drastic since we are compressing  $10^5 - 10^7$  timesteps into a much shorter 500 simulation frames during visualization. Figure 12 shows the equilibration of the temperature and potential energy of the high-temperature system. Note that the fluctuations are larger and the system equilibrates sooner than in the liquid case at 294 K.



**Figure 8: Energy profiles for an 80 20-mer volume of PDMS, for the same simulation temperature,  $k_B T = 1$ , approximately 500 K. Potential energies (both Lennard-Jones and total) are per mole.**

## 6 DISCUSSION

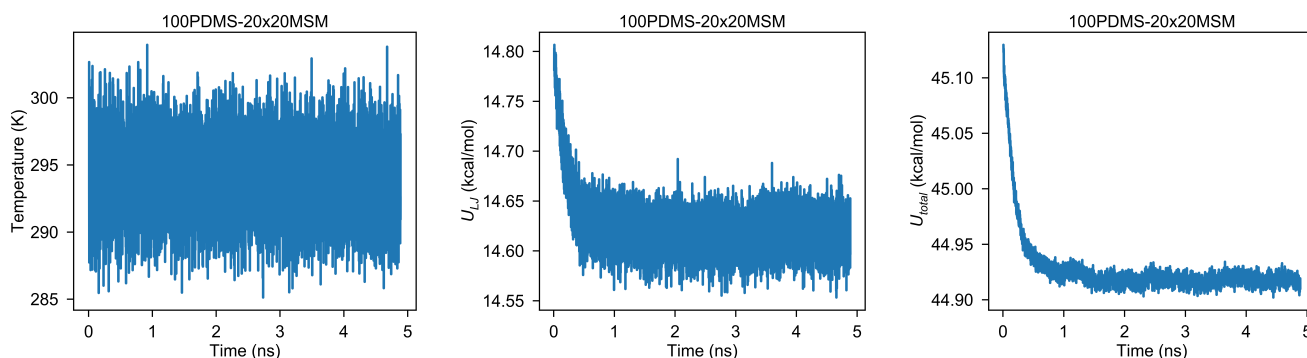
The final model presented here permits the equilibration and sampling of PDMS on Ni-Mn-Ga efficiently across a range of state points. Identifying this model required iterative testing checking for artefacts including periodic oscillation of temperature, or inaccurate combinations of forcefields. Early test showed that PDMS did not adhere as the sealant on the alloy, instead aggregating away from the surface (Section 7.1). The final model instead shows the



**Figure 9: First time step (top) and the contrast of the 20<sup>th</sup> time step (bottom-left) and the last time step (bottom-right), of our interface.**

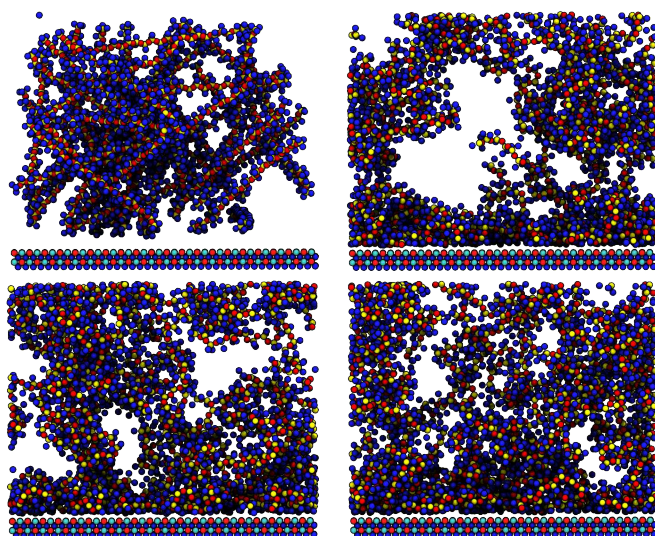
two materials to attract and permits surface structure to be investigated, therefore providing a starting point for understanding the sealant-alloy interface.

We now turn to comparing our simulation results to literature, in order to validate our model's accuracy. While there are not our knowledge experimental values of PDMS and Ni-Mn-Ga surface energies to compare against there have been molecular investigations with PDMS interacting with other materials. Comparisons of molar energies and surface energies with such studies, despite their differences, provide useful benchmarks for future improvements to the present work. For instance, the work of *Liu et al.* [25] shows the MD simulation of PDMS interacting with a Zeolite (ZSM-5) surface using the COMPASS force field. We note that ZSM-5's structure is dissimilar to Ni-Mn-Ga, however we are interested in the fact that this is an MD simulation of PDMS with another material for which energy values are provided in a manner that we can easily calculate and compare to ours. In the manuscript, the authors report the total nonbonded interaction energy of PDMS as being -1885.6 kcal/mol. This interaction energy is a total energy of the system as a whole, in contrast to the energies per mole reported in Figures 10 and 12. Reporting energies per mole allows for a direct comparison between the 7,120 particle PDMS-only simulations and the 21,700 particle combined PDMS Ni-Mn-Ga simulations. The simulations in *Liu et al.* contain five chains, each made up of four repeating units, which we calculate to contain 235 total atoms in the system. The per-atom PDMS nonbonded potential energy from *Liu et al.* is therefore 8.02 kcal/mol, which is in good agreement with our simulated value of 2.6 kcal/mol for our simulated volume of PDMS by being within the same order of magnitude. We would not expect our non-bonded values to align perfectly, given that the biggest difference in simulation is the simulation temperatures: 303K for *Liu et al.* vs. 500K for ours. We can perform similar analysis for the PDMS-Zeolite simulations, according to the reported total nonbonded potential energy of -13073.4 kcal/mol. Using the same number of atoms, and an estimated 288 atoms present in the authors' Zeolite unit cell, we obtain a per-atom nonbonded PDMS-Zeolite potential energy of 25.00 kcal/mol, which is within a factor of two of the 14.62 kcal/mol energy calculated from our simulations. We also note a consistent ratio between the neat PDMS interaction



**Figure 10: Energy profiles for our Ni-Mn-Ga-PDMS interface at a set simulation temperature of 294K. Both potential energies, Lennard-Jones and total, are per mole.**

energies and the surface energies in our work compared to *Liu et al.* The ratio of PDMS-Ni-Mn-Ga interaction energies to neat PDMS is 5.6. Similarly, the ratio between PDMS-Zeolite interaction energies and neat PDMS in *Liu et al.* is 3.1. The similarity and order-of-magnitude quantitative agreement in these calculated energy levels lends additional confidence to our hybrid OPLS-UA/UFF forcefield, while providing additional versatility over COMPASS by parameterizing interactions between any elemental atoms in the periodic table. Furthermore, it must be noted that the substrates in the simulations of *Liu et al.* and ours, Zeolite and Ni-Mn-Ga respectively, are dissimilar which is why despite the energy ratios not being the same, the fact that they're within an order of magnitude gives us confidence that our model is within the realm of credibility. This



**Figure 11: Progression of our interface model at a temperature of 500 K. First frame (top-left), 100<sup>th</sup> frame (top-right), 250<sup>th</sup> frame (bottom-left), and final frame (bottom-right). We observe PDMS self-aggregating at less 500 K than at 294 K (Fig. 9), as expected.**

is not to say that the model cannot be more finely tuned to more closely yield results to those of *Liu et al.*, but rather that by being within such a factor and order of magnitude our simulation does not, at this current first glance, require more than some parameter modifications as opposed to a full rebuild. *Tsige et al.* [37] present a similar simulation of neat PDMS, containing 100 20-mers of the elastomer with SiO<sub>2</sub>. This simulation is important, not because of the substrate, but rather because the PDMS used in this simulation covers the exact same chain length and number of chains as our simulation, making it another great point of reference. The authors report a nonbonded potential energy of 6 kcal/mol at a separation of 4 Å, in good agreement with the current model. This is to say that this reported value lies within the range of nonbonded values that we see in our model: 2.6 kcal/mol for PDMS by itself and 14.62 kcal/mol for PDMS-Ni-Mn-Ga, where our metal substrate has more of an influence than SiO<sub>2</sub>.

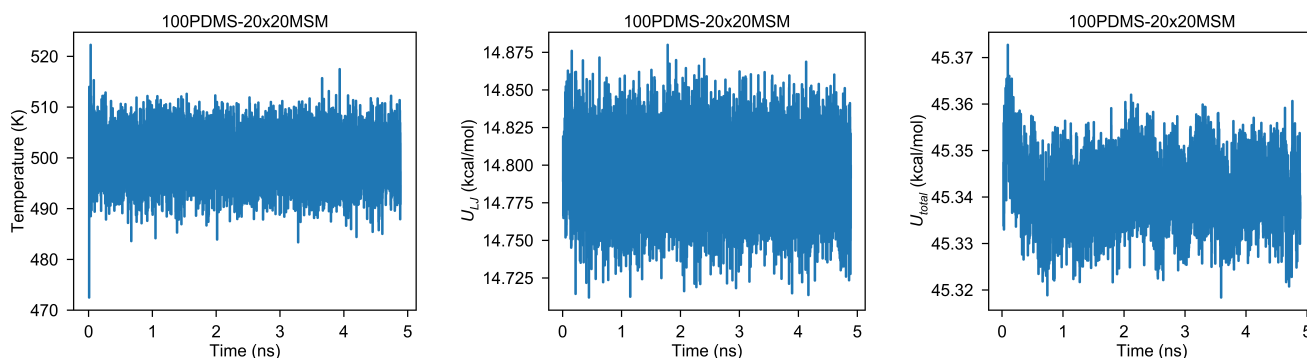
Being able to simulate PDMS and Ni-Mn-Ga together opens up significant opportunities for future study, from both HPC and materials science perspectives. We can continue to hone our toolkits and algorithms to yield more efficient and accurate simulations, reducing computational cost while maintaining good quantitative agreement with experiment. From a materials viewpoint, the model demonstrated here provides an excellent staging ground for understanding surface-sealant interactions for micro-electro-mechanical systems, guiding manufacturing processes and focussing future experimentation efforts.

## 7 INTERNSHIP REFLECTION

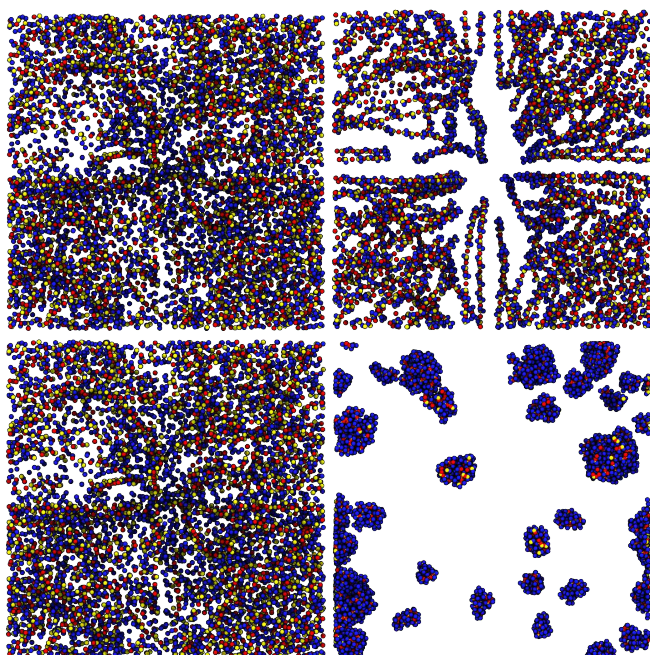
To meet the JOCSE aim of improving computational science education, we reflect on pivotal learning moments over the course of the Blue Waters Student Internship Program (BWSIP).

A problem that can arise through these simulations comes when particles are too tightly bound within a simulation space. If the user doesn't take care in adjusting time steps and integrator time constants to control thermostating, numerical instabilities can result in unphysical particle displacements (sometimes referred to as the simulation "exploding").





**Figure 12: Energy profiles for our Ni-Mn-Ga-PDMS interface at a set simulation temperature of 500K. Both potential energies, Lennard-Jones and total, are per mole.**



**Figure 13: Larger time steps advance simulations faster. (top row) After the first time step  $\delta t = 10^{-3}$  fs (left) and  $\delta t = 1$  fs (right). (bottom row) final snapshots after  $10^3$  steps.**

## 7.1 Troubleshooting

After the two basic structures for our PDMS molecule, and our Ni-Mn-Ga alloy, were finished they were then put into simulations. Previously, it was mentioned that PDMS was chosen to be constructed first, as it would allow for better early testing. This testing was not only meant to be of the material, but also of the code we wrote to initialize and perform these simulations. Over the course of the last year many issues were resolved with the PDMS simulations, the biggest of these problems being the following:

**7.1.1 Timestep Selection.** In Figure 13 we can see a problem resulting when the time step is too small. The snapshots in the left

column show after the 1st (top) and 1000th (bottom) step when  $\delta t = 0.001$  fs. The snapshots in the right column show after the 1st (top) and 1000th (bottom) step when  $\delta t = 1$  fs. Even after the first snapshot there is a visible difference between the 0.001 fs and 1 fs cases, and after 1000 steps the 1 fs case shows PDMS aggregating while the 0.001 fs case has now changed. Here, a step size of 0.001 fs is so small that position changes fall below the rounding threshold in 32-bit precision addition during integration, and the system is essentially frozen. Visualizing the trajectory quickly identifies this problem, but such a problem may go undetected if not visualized because the 0.001 fs system is numerically stable and quickly “equilibrates”.

**7.1.2 Image Correction.** To enable a finite number of particles to represent bulk interfacial areas, periodic boundary conditions are employed. These boundaries are not boundaries: when a simulation element moves out of the left-hand face of the box, it enters again from the right, and particle interactions are considered across periodic boundaries along each axis. HOOMD therefore needs to keep track of which periodic image that the particle is in, in order to calculate how far it has traveled in the system. Initially, every particle in the system is located in the original simulation volume and has image indices  $[0, 0, 0]$ . However, our initial XML files did not contain these indices even after we had established the other parameters in our XML, leading to numerical instabilities in our simulation. This problem was solved through by using MorphCT to manipulate our initial XML file to obtain the correct formatting for HOOMD.

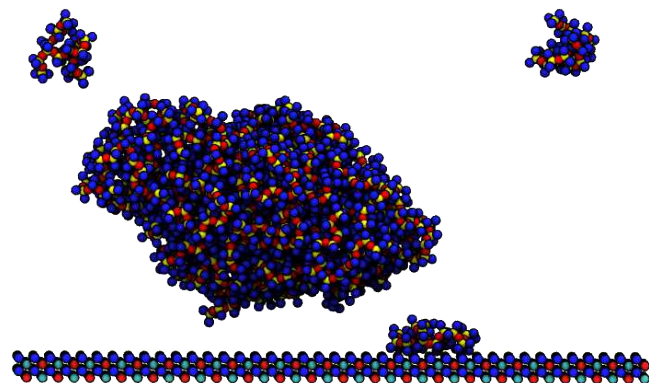
**7.1.3 Coefficient Correction.** When trying to simulate a larger volume of PDMS from the molecules we first produced, an error with the bond, angle, and dihedral coefficients was encountered: They were missing from the XML file. Without these coefficients, the conformational structure of PDMS could not be maintained and the atoms in the molecules would act as if they were not bonded to anything. This would undoubtedly lead to incorrect particle trajectories and unreliable results about the interaction of PDMS with the Ni-Mn-Ga surface. This problem was solved by using mBuild’s builder package to populate the bond, angle, and dihedral coefficients accordingly based on a neighbor list it created for each

of the particles present in the system. Furthermore, we also needed to rescale the system in order to correct incorrect, and inconsistent, distances along with the energy that the system takes in from the forcefield used.

In the initial surface simulations, we noticed very strong polymer-polymer interactions but significantly weaker polymer-surface adsorption as shown in Figure 14. At first this was attributed to PDMS desorbing over long timescales. Further investigation revealed that a UFF unit conversion problem resulting in negligible interactions between the PDMS and the Ni-Mn-Ga. Reworking the potential parameters to the units Rhaco expected resolved this issue: After the OPLS-UA and UFF potentials were calibrated, PDMS integrated was successful and adsorption onto Ni-Mn-Ga was observed.

**7.1.4 Volume Packing.** This problem presented early on when trying to produce a sample volume of PDMS where the number of molecules being packed into our early systems exceeded the size of the system itself. This then would continuously lead to our simulations “exploding” - a colloquial term used when atoms are placed so close to each other that they experience strong repulsion from the Lennard-Jones equation, often many orders of magnitude greater than the forces expected throughout the rest of the system. The large forces lead to large velocities of particles, and in a single time step an atom can move distances many times larger than the simulation volume, causing the program to crash. This issue was resolved by changing our input of molecules by using a reduced number, this then let mBuild pack our molecules at a lower density than the physical material.

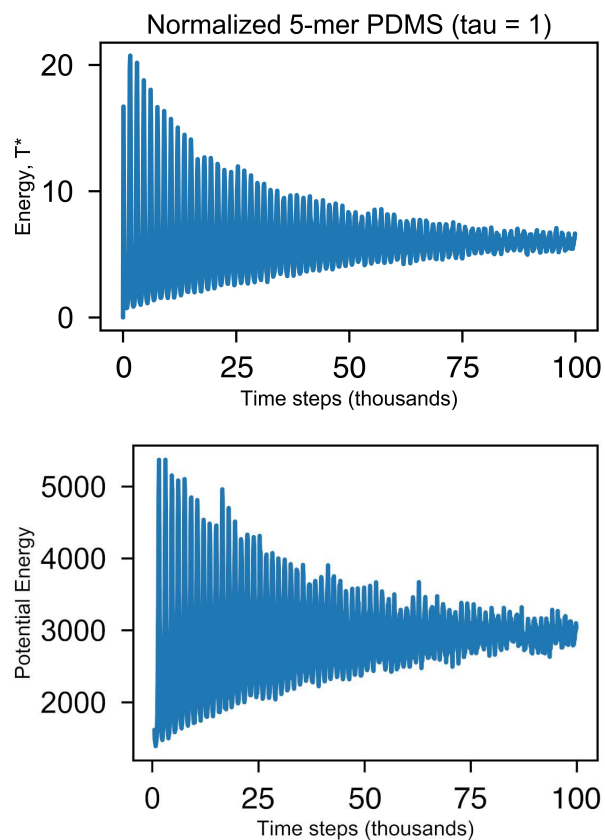
**7.1.5 Integrator ringing.** The evolution of the system temperature in Figure 15 for this initial simulation shows that the simulation needs to run for at least 100,000 time steps in order for it to stabilize around the set point. Additionally, periodic “ringing” of the temperature can be observed, as the thermal energy of the system oscillates around the set point. HOOMD regulates its temperature through the use of the Nosé-Hoover thermostat [14], which couples to an infinite thermal bath and injects (removes) thermal energy to (from) the simulation by increasing (decreasing) the velocities of the atoms in the simulation, regulating its temperature. The strength of this coupling can be modified using the parameter  $\tau$ . Selecting an



**Figure 14:** End state of initial testing of PDMS and Ni-Mn-Ga through the use of Rhaco

appropriate  $\tau$  is extremely important to avoid unphysical ringing from the integrator. If  $\tau$  is too high, then the timescale of energy control is large and initial deviations from the set point may take millions of steps before stabilizing. If  $\tau$  is too low it permits the integrator to “overshoot” the set point, and the ringing seen above occurs as the integrator iteratively removes and adds kinetic energy over short times, oscillating about the setpoint. Periodic oscillations of the temperature are indicative of an unbalanced  $\tau$  value, and it is important to modify tau incrementally to obtain consistent, random fluctuations around the set point temperature.

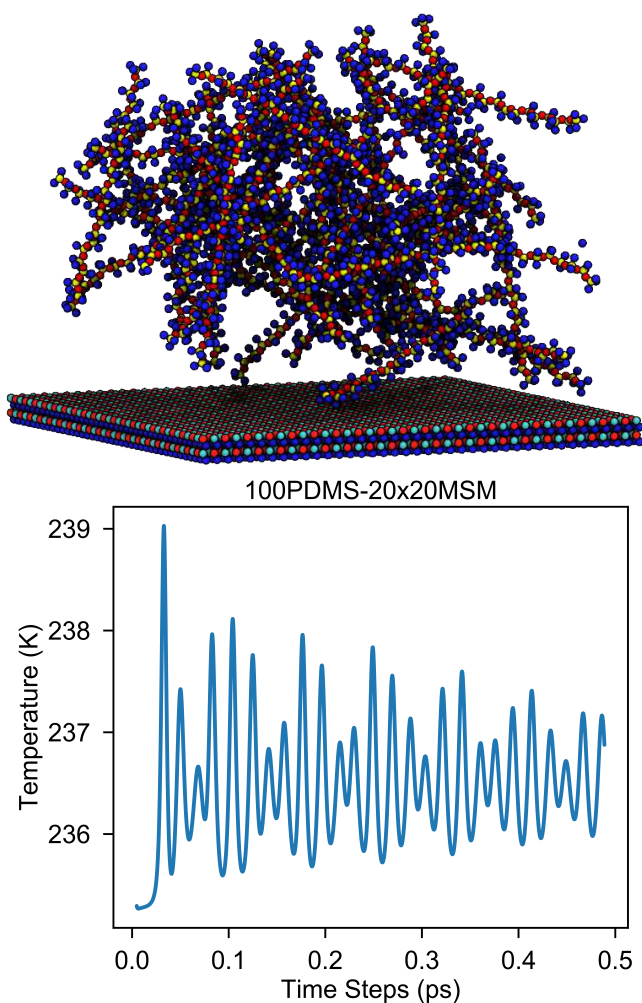
In addition to having a stable temperature, the stability of the potential energy of the system is also important for molecular dynamics simulations. As the molecules relax in the system, the potential energy decreases as the simulation attempts to find the global minimum of free energy (balancing potential energy minimization with entropy maximization and therefore the most likely conformations of molecules thermodynamically). When the potential energy no longer evolves, the equilibrium energy has been found. Generally, small systems and those at high temperature equilibrate quickly, whereas larger systems or those at low temperature can take millions of time steps for the potential energy to stop evolving. It is therefore imperative to consider both the



**Figure 15:** Energy plots for our initial 5-mer chain. Temperature/Kinetic (top), and Potential (bottom)

temperature and the potential energy before reporting structural or energetic results for a molecular dynamics simulation.

As an example, initial simulations of the PDMS and alloy yielded some promising results, as shown in Figure 16 for a system of 36,100 particles. However, the temperature profile of the simulation also demonstrates integrator ringing. This indicates that the value of  $\tau$  is unbalanced, and could also suggest that longer runtimes are needed for this simulation. In fact, on further review, these data identified a critical bug in Rhaco. While Rhaco was successfully able to generate the interface observed in Figures 16 and 14, the conversion calculation between the dimensionless simulation temperature and the real temperature in Kelvin was incorrect and needed to be fixed. This error arose from the surface atoms being assigned initial velocities, even though those atoms were omitted from the integrator and never actually moved in the simulation. The additional kinetic energy in the system was interpreted as an increased temperature, offsetting our reported temperatures by several hundreds of Kelvin.



**Figure 16: Initial testing of the PDMS-Ni-Mn-Ga interface showing some attempt at temperature stabilization.**

## 8 CONCLUSIONS

While being able to fully simulate the interaction of PDMS and Ni-Mn-Ga alloy inside of a microfluidic pump is a significant challenge, we here demonstrate the first steps towards achieving that goal. We develop and present a successful model and toolkit for simulating the interactions of the sealant at an alloy surface. Our model is versatile in terms of input parameters, allowing us to test systems of different sizes, PDMS densities, alloy thickness, and processing temperatures, and can be easily extended to include additional materials such as solvents and dyes in the simulation volume. For the future we hope to develop additional analysis tools that can calculate surface energies more rigorously, include additional infrastructure to simulate twin-boundary dislocations in the alloy and ascertain the effect this has on the PDMS adsorption, as well as identify other, superior sealant candidates for this application. Pedagogically, this has been a great introduction into High Performance Computing (HPC) and computation in general, especially in how to best optimize code: The original simulation code took 40-45 minutes to produce a trajectory for a relatively small amount of neat PDMS, whereas the final tool could simulate the interactions large amounts of both PDMS and Ni-Mn-Ga alloy in approximately the same wall-clock time, representing a simulation time of several nanoseconds. Although my allocation on Blue Waters has come to an end for this project, the lessons learned over my time working with this system will definitely aid JG in the future, be it in continuing with this work, testing different versions of this interface, or with HPC as a whole.

## 9 ACKNOWLEDGEMENTS

This work was supported by a grant from the Shodor Education Foundation through the Blue Waters Student Internship Program (BWSIP) and by the National Center for Supercomputing Applications, who provided hardware used in this project. National Science Foundation through project NSF-DMR 1710640. This work used the computing resources supported by Boise State College of Engineering Information Technology services.

## REFERENCES

- [1] Mp Allen. 2004. Introduction to Molecular Dynamics Simulation. *Comput. Matter From Synth. Polym. to ...* 23, 2 (2004), 1–28. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3254&rep=rep1&type=pdf>
- [2] Joshua A. Anderson, Chris D. Lorenz, and A. Traveset. 2008. General Purpose Molecular Dynamics Simulations Fully Implemented on Graphics Processing Units. *J. Comput. Phys.* 227, 10 (may 2008), 5342–5359. <https://doi.org/10.1016/j.jcp.2008.01.047>
- [3] Avogadro. 2012. An open-source molecular builder and visualization tool. Version 1.2.0. <http://avogadro.openmolecules.net/>
- [4] V. A. Chernenko, V. L'vov, J. Pons, and E. Cesari. 2003. Superelasticity in high-temperature Ni-Mn-Ga alloys. *J. Appl. Phys.* 93, 5 (2003), 2394–2399. <https://doi.org/10.1063/1.1539532>
- [5] K. S. Colwell, C. J. Casewit, W. A. Goddard, W. M. Skiff, A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. 2005. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* 114, 25 (dec 2005), 10024–10035. <https://doi.org/10.1021/ja00051a040> arXiv:1305.7119
- [6] Murray S. Daw and M. I. Baskes. 1984. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Phys. Rev. B* 29, 12 (1984), 6443–6453. <https://doi.org/10.1103/PhysRevB.29.6443>
- [7] Ryan S. Elliott. 2018. Efficient 'Universal' Shifted Lennard-Jones Model for All KIM API Supported Species Developed by Elliott and Akerson (2015) v003. <https://doi.org/10.25950/962b4967>
- [8] Elveflow. 2005. PDMS: A Review. <https://www.elveflow.com/>

- [9] Amalie L. Frischknecht and John G. Curro. 2003. Improved United Atom Force Field for Poly(dimethylsiloxane). *Macromolecules* 36, 6 (mar 2003), 2122–2129. <https://doi.org/10.1021/ma025763g>
- [10] Jens Glaser, Trung Dac Nguyen, Joshua A. Anderson, Pak Lui, Filippo Spiga, Jaime A. Millan, David C. Morse, and Sharon C. Glotzer. 2014. Strong Scaling of General-purpose Molecular Dynamics Simulations on GPUs. *Comput. Phys. Commun.* 192 (dec 2014), 97–107. <https://doi.org/10.1016/j.cpc.2015.02.028> arXiv:1412.3387
- [11] Oleg Heczko and Ladislav Straka. 2003. Temperature dependence and temperature limits of magnetic shape memory effect. *J. Appl. Phys.* 94, 11 (2003), 7139–7143. <https://doi.org/10.1063/1.1626800>
- [12] Michael M Henry, Matthew Lewis Jones, Stefan D Oosterhout, Wade A Braunecker, Travis W Kemper, Ross E Larsen, Nikos Kopidakis, Michael F Toney, Dana C Olson, and Eric Jankowski. 2017. Simplified Models for Accelerated Structural Prediction of Conjugated Semiconducting Polymers. *J. Phys. Chem. C* 121, 47 (nov 2017), 26528–26538. <https://doi.org/10.1021/acs.jpcc.7b09701>
- [13] Tilmann Hicckel, Matthé Uijttewaal, Ali Al-Zubi, Biswanath Dutta, Blazej Grabowski, and Jörg Neugebauer. 2012. Ab Initio-Based Prediction of Phase Diagrams: Application to Magnetic Shape Memory Alloys. *Adv. Eng. Mater.* 14, 8 (aug 2012), 547–561. <https://doi.org/10.1002/adem.201200092>
- [14] William G. Hoover. 1985. Canonical Dynamics: Equilibrium Phase-space Distributions. *Phys. Rev. A* 31, 3 (1985), 1695–1697. <https://doi.org/10.1103/PhysRevA.31.1695>
- [15] Michael P. Howard, Joshua A. Anderson, Arash Nikoubashman, Sharon C. Glotzer, and Athanassios Z. Panagiotopoulos. 2016. Efficient Neighbor List Calculation for Molecular Simulation of Colloidal Systems Using Graphics Processing Units. *Comput. Phys. Commun.* 203 (2016), 45–52. <https://doi.org/10.1016/j.cpc.2016.02.003>
- [16] William Humphrey, Andrew Dalke, and Klaus Schulten. 1996. VMD: Visual molecular dynamics. *J. Mol. Graph.* 14, 1 (feb 1996), 33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5)
- [17] Christopher R Iacovella, Janos Sallai, Christoph Klein, and Tengyu Ma. 2017. Idea Paper: Development of a Software Framework for Formalizing Forcefield Atom-Typing for Molecular Simulation. <https://github.com/mosdef-hub/foyer>
- [18] M. L. Jones. 2018. MorphCT - DOI: 10.5281/zenodo.1243843. <https://doi.org/10.5281/zenodo.1243843>
- [19] M. L. Jones. 2018. Rhaco - DOI: 10.5281/zenodo.1308187. <https://doi.org/10.5281/zenodo.1308187>
- [20] Matthew Lewis Jones and Eric Jankowski. 2017. Computationally connecting organic photovoltaic performance to atomistic arrangements and bulk morphology. *Mol. Simul.* 43, 10-11 (mar 2017), 1–18. <https://doi.org/10.1080/08927022.2017.1296958>
- [21] William L. Jorgensen and Julian. Tirado-Rives. 1988. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimization for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* 110, 6 (mar 1988), 1657–1666. <https://doi.org/10.1021/ja00214a001>
- [22] D. Kikuchi, T. Kanomata, Y. Yamaguchi, H. Nishihara, K. Koyama, and K. Watanabe. 2004. Magnetic properties of ferromagnetic shape memory alloys Ni<sub>2</sub>Mn<sub>1-x</sub>Fe<sub>x</sub>Ge. *J. Alloys Compd.* 383, 1-2 (2004), 184–188. <https://doi.org/10.1016/j.jallcom.2004.04.053>
- [23] Christoph Klein, János Sallai, Trevor J Jones, Christopher R Iacovella, Clare McCabe, and Peter T Cummings. 2016. A Hierarchical, Component Based Approach to Screening Properties of Soft Matter. *Found. Mol. Model. Simul.* (2016). [https://doi.org/10.1007/978-981-10-1128-3\\_5](https://doi.org/10.1007/978-981-10-1128-3_5)
- [24] Thomas Kluyver, Benjamin Ragan-kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. 2016. Jupyter Notebooks - A Publishing Format for Reproducible Computational Workflows. *Position. Power Acad. Publ. Play. Agents Agendas* (2016), 87–90. <https://doi.org/10.3233/978-1-61499-649-1-87>
- [25] Gongping Liu, Fenjuan Xiangli, Wang Wei, Sainan Liu, and Wanqin Jin. 2011. Improved performance of PDMS/ceramic composite pervaporation membranes by ZSM-5 homogeneously dispersed in PDMS via a surface graft/coating approach. *Chem. Eng. J.* 174, 2-3 (2011), 495–503. <https://doi.org/10.1016/j.cej.2011.06.004>
- [26] Evan Miller, Matthew Jones, and Eric Jankowski. 2018. Tying Together Multiscale Calculations for Charge Transport in P3HT: Structural Descriptors, Morphology, and Tie-Chains. *Polymers (Basel)*. 10, 12 (dec 2018), 1358. <https://doi.org/10.3390/polym10121358>
- [27] Evan D Miller, Matthew Lewis Jones, Michael M Henry, Paul Chery, Kyle Miller, and Eric Jankowski. 2018. Optimization and Validation of Efficient Models for Predicting Polythiophene Self-Assembly. *Polymers (Basel)*. 10, 12 (nov 2018), 1305. <https://doi.org/10.3390/polym10121305>
- [28] Evan D Miller, Matthew Lewis Jones, and Eric Jankowski. 2017. Enhanced Computational Sampling of Perylene and Perylothiophene Packing with Rigid-Body Models. *ACS Omega* 2, 1 (jan 2017), 353–362. <https://doi.org/10.1021/acsomega.6b00371>
- [29] P. Müllner, V. A. Chernenko, M. Wollgarten, and G. Kostorz. 2002. Large cyclic deformation of a Ni-Mn-Ga shape memory alloy induced by magnetic fields. *J. Appl. Phys.* 92, 11 (2002), 6708–6713. <https://doi.org/10.1063/1.1513875>
- [30] S. J. Murray, M. Marioni, S. M. Allen, R. C. O’Handley, and T. A. Lograsso. 2000. 6% magnetic-field-induced strain by twin-boundary motion in ferromagnetic Ni-Mn-Ga. *Appl. Phys. Lett.* 77, 6 (2000), 886–888. <https://doi.org/10.1063/1.1306635>
- [31] NCSA. 2018. About Blue Waters. <http://www.ncsa.illinois.edu/>
- [32] C. Seguí, V. A. Chernenko, J. Pons, E. Cesari, V. Khovailo, and T. Takagi. 2005. Low temperature-induced intermartensitic phase transformations in Ni-Mn-Ga single crystal. *Acta Mater.* 53, 1 (2005), 111–120. <https://doi.org/10.1016/j.actamat.2004.09.008>
- [33] Aaron R. Smith, Andrey Saren, Jari Järvinen, and Kari Ullakko. 2015. Characterization of a High-Resolution Solid-State Micropump that can be Integrated into Microfluidic Systems. *Microfluid. Nanofluidics* 18, 5-6 (may 2015), 1255–1263. <https://doi.org/10.1007/s10404-014-1524-6>
- [34] A. Sozinov, A.A. Likhachev, and K Ullakko. 2002. Crystal Structures and Magnetic Anisotropy Properties of Ni-Mn-Ga Martensitic Phases with Giant Magnetic-Field-Induced Strain. *IEEE Trans. Magn.* 38, 5 (sep 2002), 2814–2816. <https://doi.org/10.1109/TMAG.2002.803567>
- [35] L. Straka, A. Soroka, H. Seiner, H. Hänninen, and A. Sozinov. 2012. Temperature dependence of twinning stress of Type I and Type II twins in 10M modulated Ni-Mn-Ga martensite. *Scr. Mater.* 67, 1 (2012), 25–28. <https://doi.org/10.1016/j.scriptamat.2012.03.012>
- [36] Yoshinori Tamai, Hideki Tanaka, and Koichiro Nakanishi. 1994. Molecular Simulation of Permeation of Small Penetrants through Membranes. 1. Diffusion Coefficients. *Macromolecules* 27, 16 (aug 1994), 4498–4508. <https://doi.org/10.1021/ma00094a011>
- [37] Mesfin Tsigie, Thomas Soddemann, Susan B. Rempe, Gary S. Grest, Joel D. Kress, Mark O. Robbins, Scott W. Sides, Mark J. Stevens, and Edmund Webb. 2003. Interactions and structure of poly(dimethylsiloxane) at silicon dioxide surfaces: Electronic structure and molecular dynamics studies. *J. Chem. Phys.* 118, 11 (2003), 5132–5142. <https://doi.org/10.1063/1.1545091>
- [38] K. Ullakko, J. K. Huang, C. Kantner, R. C. O’Handley, and V. V. Kokorin. 1996. Large magnetic-field-induced strains in Ni<sub>2</sub>MnGa single crystals. *Appl. Phys. Lett.* 69, 13 (1996), 1966–1968. <https://doi.org/10.1063/1.117637>

# Using Blue Waters to Assess Tornadoic Outbreak Forecast Capability by Lead Time

Caroline MacDonald  
Department of Geosciences  
Mississippi State, MS  
cnm306@msstate.edu

Andrew Mercer  
Department of Geosciences  
Mississippi State, MS  
a.mercer@msstate.edu

## ABSTRACT

Severe weather outbreaks come with many different hazards. One of the most commonly known and identifiable outbreaks are those with tornadoes involved. There has been some prior research on these events with respect to lead time, but shifts in model uncertainty by lead time has yet to be quantified formally. As such, in this study we assess tornado outbreak model uncertainty by lead time by assessing ensemble model precision for outbreak forecasts. This assessment was completed by first identifying five major tornado outbreak events and simulating the events using the Weather Research and Forecasting (WRF) model at 24, 48, 72, 96, and 120-hours lead time. A 10-member stochastically perturbed initial condition ensemble was generated for each lead time to quantify uncertainty associated with initialization errors at the varied lead times. Severe weather diagnostic variables derived from ensemble output were used to quantify ensemble uncertainty by lead time. After comparing moment statistics of several convective indices, the Energy Helicity Index (EHI), Significant Tornado Parameter (STP), and Supercell Composite Parameter (SCP) did the best job of characterizing the tornadoic outbreaks at all lead times. There was good consistency between each case utilizing these three indices at all five lead times, suggesting outbreak model forecasting confidence may be able to extend up to 5 days for major outbreak events. These results will be useful for operational use by forecasters in forecast ability of tornadoic events.

## Keywords

Tornado outbreaks, stochastic initial condition perturbation, numerical weather prediction, ensemble forecasting

## 1. INTRODUCTION

Tornado outbreaks are one of the premier atmospheric phenomena that cause substantial damage across the United States each year. While considerable research has been performed on these outbreaks [1, 2, 3], many questions remain regarding outbreak predictability and model uncertainty at increasing forecast lead time.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright ©JOCSE, a supported publication of the Shodor Education Foundation Inc.

© 2020 Journal of Computational Science Education  
DOI: <https://doi.org/10.22369/issn.2153-4136/11/2/4>

Previous research defines a tornado outbreak as ten or more tornadoes associated with a single synoptic-scale system and confined to a specific geographic area [1]. Importantly, 73% of deaths associated with historic tornadoes were associated with tornadoic outbreak events [1]. On average, over 1,000 tornadoes occur across the United States each year (NCDC), a majority of which are associated with outbreaks. Additionally, tornadoes have the highest ten-year average fatality rate among weather related hazards from 2008 to 2017 [5].

Given the hazardous nature of tornadoic events, forecasting of these events several days in advance is vital. Frequently, convective and kinematic indices are utilized to quantify outbreak characteristics and predict outbreak mode at varying lead times [3]. Many previous studies have studied short and medium-term outbreak model uncertainty using these indices [3, 6, 7, 8, 9, 10]. Research has shown that there is a statistical drop off in forecast skill at lead times of 72 hours or longer [7], which is typical in most forecasting research. Keeping these limitations in mind, little research has been conducted looking beyond 72 hours in advance of tornadoic outbreaks.

Recently, improvements in numerical weather prediction models such as the Weather Research and Forecasting model (WRF) have resulted in improved forecasts of tornado outbreaks [9]. Such improvements result from prior knowledge and pattern recognition, and the availability of the WRF to reproduce historic major events has helped forecasters identify important meteorological details in these events. Thompson and Edwards [11] recognized that large scale outbreaks may not be as evident through numerical model output as previously thought. Their work showed that forecasts 12 to 24 hours in advance of the major 3 May 1999 tornado outbreak relied heavily on poor model output, which led to greater forecast uncertainty. An additional study, [12] relied on the WRF model to conduct similar case study research at varying lead times, noting forecast uncertainty increases with lead time. This suggests more work into model uncertainty is needed at both short and medium lead times. Clearly, a need for improved understanding of model uncertainty in tornado outbreak forecasts exists.

The objective of this research is to quantify uncertainty in tornado outbreak convective indices by lead time, which in turn will allow forecasters to have greater confidence when forecasting major tornadoic outbreaks. Uncertainty quantification via ensemble modeling is a primary way to assess predictability as increasing ensemble spread relates directly to reduced model precision. Following the methods of previous research [13], five major tornado outbreaks were selected and WRF simulations conducted at lead times of 24, 48, 72, 96, and 120-hours were completed. Previous work suggests that short-term forecast uncertainty will be significantly less than medium-term (96 to 120 hour) forecasts, which is the primary hypothesis of this study.

This student paper summarizes the purpose and results from research conducted through the Blue Waters Student Internship Program. Through this program, computational resources were utilized via the Blue Waters supercomputing facility. Section 2 will describe the data and methods utilized in this study as well as the computer and statistical modeling that aided in this research. Section 3 will discuss findings from this research at varying lead time intervals and the education impact related to this research. Additionally, this section will discuss the overall student goal and challenges associated with this study. The 4<sup>th</sup> section will refer to reflections and discussion about what the student learned through this research and the computing knowledge the student gained through this program. Section 5 will highlight the summaries and draw conclusions from this study and highlight the needs of future research.

## 2. DATA AND METHODS

### 2.1 Data

As the primary goal of this research was to analyze the importance of lead time for tornadic severe weather outbreak forecasting, five major tornado outbreaks were selected from the Storm Event database [14] from the National Climatic Data Center. A major tornado outbreak was selected based on the number of tornadoes produced and overall damage extent of such tornadoes. Each of the five cases produced millions of dollars in damage, had several large, significant tornadoes, and frequently resulted in multiple deaths. Table 1 displays such statistics for each of the five outbreak cases [15-19].

**Table 1. Displays of case date and statistical information for all five tornado outbreak cases utilized within this study**

Date	Total # of Tornadoes	EF/F 3+ Tornadoes	Deaths	Damage (millions)
10 April 1979	29	5	50	63
28 March 1984	24	12	57	600
31 May 1985	43	14	89	600
3 May 1999	75	9	46	1500
17 June 2010	74	8	4	100

A robust database of three-dimensional meteorological analyses was required to complete WRF simulations on the five outbreaks. Given the spatial scale of the outbreak events and the dates of the events, data was obtained from a mesoscale reanalysis analysis dataset, the North American Regional Reanalysis [20]. The NARR are provided on a 32-km Lambert conformal North American grid with 3-hourly temporal resolution and 29 vertical levels, spanning 1979 to present. NARR fields were retained such that lead times of 24, 48, 72, 96, and 120 hours could be simulated using the WRF.

### 2.2 Model Configuration and Simulations

In this study, the WRF model was used to formulate the outbreak forecasts at varying lead times. The WRF is highly configurable, and in this study the configuration was kept consistent with previous research [13]. In particular, the WRF was configured using the WRF Single-Moment six class microphysics scheme [21], the Grell-Freitas convective scheme [22], the Yonsei University boundary layer scheme [23], the Rapid Radiative Transfer Model radiation scheme for longwave [24], and the Dudhia scheme [25] for shortwave radiation.

The WRF model (version 3.9) was utilized to run simulations of five tornadic outbreak events listed in Table 1. Since the primary goal of this research was to relate outbreak lead time intervals to forecast uncertainties as it relates to short and medium-term forecasting of tornadic outbreak events, each case was run at the five previously described lead time intervals. In this study, the simulation period was established by defining the end of the outbreak as 1200 UTC the day after the event. Lead times were then taken relative to this time, so that a 24-hour forecast started at 1200 UTC the day of the event, and so forth. All outbreaks had a peak activity time between 2100 UTC and 0600 UTC, meaning this simulation period sufficiently captured the active tornadic activity during the event.

Following the WRF simulations, severe weather diagnostics were computed on the simulation output, including Convective Available Potential Energy (CAPE), Significant Tornado Parameter (STP), Supercell Composite Parameter (SCP), Energy Helicity Index (EHI), and Storm Relative Helicity (SRH). CAPE and SRH are commonly utilized meteorological parameters for severe weather forecasting, where CAPE describes available thermodynamic potential energy (in  $J/kg$ ) and SRH describes streamwise vorticity (in  $m^2/s^2$ ) available to be tilted vertically and contribute to mesocyclone development. The remaining parameters were derived from a combination of SRH and CAPE. The three main equations discussed in further depth below are typically utilized with right-moving supercells, which is the predominant storm motion type within tornadic outbreaks utilized in this research.

EHI combines CAPE and SRH into a single convective index:

$$EHI = \frac{CAPE * SRH}{160,000 \frac{kg s^2}{J m^2}} \quad (1)$$

Larger values of EHI are thought to be indicative of environments supporting large tornadoes and tornadic supercells.

The STP is an index used to highlight the co-existence of ingredients favoring strong, damaging tornadoes [26]. This index was developed at the Storm Prediction Center to help identify meteorologically suitable regions for significant (EF2 or stronger) tornado activity [26]. The STP is given by:

$$STP = \left( \frac{mCAPE}{1500 \frac{J}{kg}} \right) \left( \frac{ESRH}{150 \frac{m^2}{s^2}} \right) \left( \frac{EBWD}{12 \frac{m}{s}} \right) \left( \frac{2000 - mLCL}{1000} \right) \left( \frac{mLCINH + 200}{150} \right) \quad (2)$$

Here, MLCAPE represents mixed-layer (ML) CAPE, a measure of CAPE derived from a parcel originating from a well-established mixed-layer. ESRH represents the SRH computed over the effective inflow layer [26]. EBWD represents the effective bulk wind difference (the wind vector difference between the top and bottom of the effective inflow layer). MLLCL represents the lifted

condensation level of a parcel originating from the mixed layer. Smaller MLLCL values (ideally less than 1000 m) are typical in tornadic environments. Finally, MLCINH represents the mixed-layer (ML) convective inhibition (CINH), which is needed to inhibit early-morning convection and create a buildup of boundary layer potential energy for afternoon explosive convective development.

The SCP is also a commonly used parameter in forecasting tornadic outbreak events, as it was developed to highlight areas where supercell development is favorable [26]. SCP was developed at the Storm Prediction Center in Norman, Oklahoma as well. The equation is as follows:

$$SCP = \left( \frac{\mu\text{CAPE}}{1000 \text{ J/kg}} \right) \left( \frac{\text{ESRH}}{50 \frac{\text{m}^2}{\text{s}^2}} \right) \left( \frac{\text{EBWD}}{10 \frac{\text{m}}{\text{s}}} \right) \quad (3)$$

Here, MUCAPE represents the most-unstable (MU) convective available potential energy (CAPE), based on the most unstable parcel in the lowest 300 mb of the atmosphere, and all other terms have been defined previously.

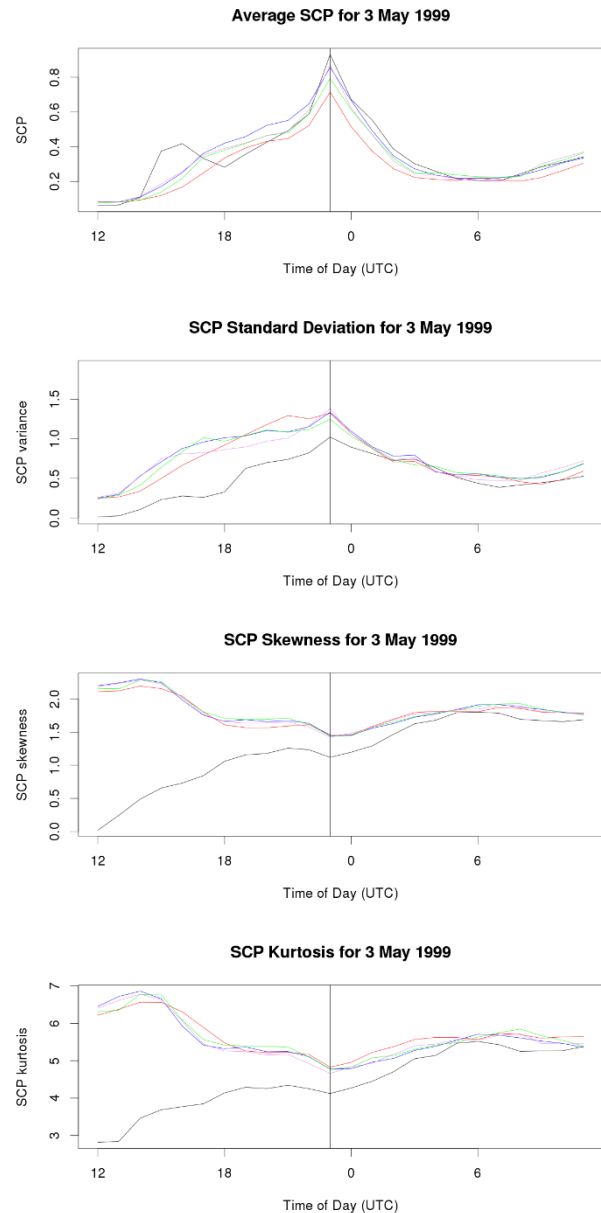
### 2.3 Simulation Analysis

To quantify forecast uncertainty, a spread in the forecast output was required. This spread was obtained by using the Stochastic Kinetic Energy Backscattering Scheme (SKEBS) [27] within the WRF configuration. The SKEBS provides initial condition perturbations to potential temperature and streamfunction values, essentially introducing random thermodynamic and kinematic noise into the model boundary conditions, which in turn creates model spread as the simulation progresses. Here, ten simulations for each case and lead time (a total of 250 simulations for all cases, perturbations, and lead times) were completed to assess WRF output uncertainty in the severe weather diagnostic fields. Uncertainty and distribution shape among the 10 perturbed simulations was identified through moment statistics (mean, variance, skewness, and kurtosis). The ten-member ensemble output was bootstrap-resampled 1000 times for each forecast timestep and moment statistics were formulated on the bootstrap replicates. The median bootstrap replicate for each moment statistic was then retained to assess ensemble member performance. The results are described below.

## 3. DISCUSSION AND LESSONS LEARNED (EDUCATIONAL IMPACT)

The primary goal of this study was to quantify uncertainty within tornado outbreak simulations with increasing lead time (24, 48, 72, 96, and 120-hours). A multi-member stochastic perturbation initial condition ensemble was used to create simulation uncertainty, while outbreak characteristics were quantified using various severe weather indices. Overall, several indices stood out compared to other indices which aligned with previous research [7, 9]. SCP, STP, and EHI appear to be three of the convective indices that seemed to have the smallest uncertainty for on tornadic outbreak events at varying lead times, as indicated by the moment statistics calculated for the five test cases. Thermodynamic indices such as CAPE had a much larger spread of data as compared to these three indices. The ensembles and moment statistics for these three variables appeared to be more consistent and uniform. While previous research [3, 6, 7, 8, 9, 10] has shown some variations in these indices as you increase lead time, this study seemed to show more consistent results at all five lead times. Figure 1 shows moment statistics from the May 3, 1999 tornado outbreak case utilizing the SCP guidance. For this case, peak outbreak time occurred at 00Z on May 4<sup>th</sup> (denoted by the vertical bar in the

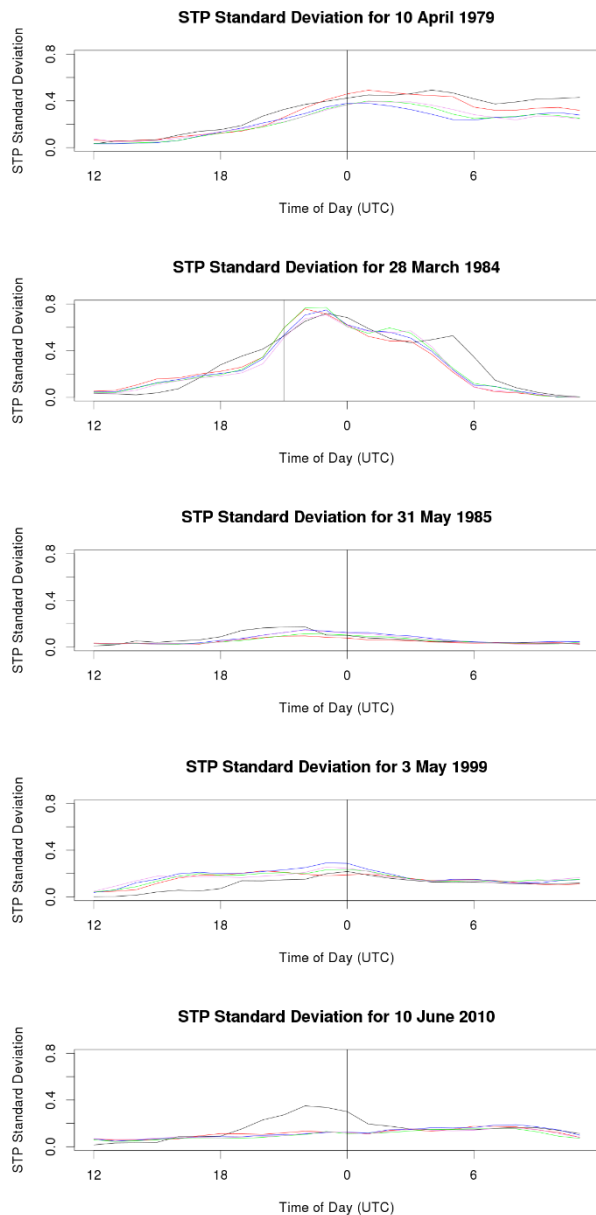
figures). At this time, all five lead time mean intervals peaked at nearly the same SCP value and provided minimal positive skewness and some leptokurtic behavior, suggesting a gamma-distribution type shape to the ensemble members that was consistent among all lead times.



**Figure 1. The four moment statistics (mean, standard deviation, skewness, and kurtosis) for the May 3, 1999 case are shown. All four moment statistics show similar results across all five lead time intervals. The only outlier is the 24-hour lead time in the skewness and kurtosis plots, which could be due to the small amount of time as compared to the remaining four lead time intervals. Peak outbreak time is represented by the vertical black line.**

These results suggest uncertainty remains consistent by lead time for these events, suggesting similar predictability at 120 hours exists at 24 hours.

Kinematic indices such as SCP, STP, and EHI appear to be more beneficial in medium-term forecasting of tornadic outbreak events. These three indices had similar distribution characteristics at all five lead times and had the smallest spread, suggesting high ensemble precision in the kinematic fields, as shown in Figure 2.



**Figure 2. Standard Deviation (SD) statistics for all five tornadic outbreak cases in order of: April 10, 1979, March 28, 1984, May 31, 1985, May 3, 1999, and June 17, 2010 outbreak cases. Peak outbreak time is displayed by the vertical black line in each graph.**

These results are not surprising as kinematic fields within tornado outbreaks are typically governed by the highly predictable synoptic-scale background flow patterns. Interestingly, the only major differences were observed in the 24-hour ensemble forecasts, suggesting predictability above 24-hours lead time remains consistent for tornado outbreak events. Thermodynamic indices

exhibited more variability, but the general trend of 48-hour lead times and beyond performing consistently remained intact.

This research was conducted through the Blue Waters Student Internship Program which utilized the Blue Waters supercomputing center. Through this program, the student was exposed to high-performance computing research in the atmospheric sciences and gained many valuable educational lessons. For example, the student learned that scientific research requires flexibility in the research timeline. On numerous occasions timelines were altered to allow for unforeseen complications, including compilation issues, big data issues, and data transfers that took longer than originally anticipated.

Along with the computation and educational lessons learned, the student gained valuable experience in big data problems, which are becoming the standard in atmospheric science research, particularly research involving numerical weather prediction. In particular:

1. The student learned that computational resources must be modified at times to meet the needs of the research that is trying to be done. Due to a compilation issue, the student had to transfer all simulation output between an in-house computer and the Blue Waters system for post-processing. This required tedious organization to ensure all cases were stored and post-processed properly and required some cases to be redone. The in-house computer was also a shared resource, which further hindered progress and required the student to be fully apprised of time and resource issues.
2. Time management played a huge role within this research, particularly accounting for run-times of some of the longer simulations. Many of the outbreak simulations took hours to run, so the student needed to set up distinct times for running these codes and a lot of time to transfer the data between an in-house computer and the Blue Waters computer. Additionally, transferring the cases for post-processing added an additional time component that was not foreseen at the beginning of this research.
3. The student gained valuable knowledge into the difficulties and limitations with tornadic outbreak modeling, as well as some of the meteorological advantages and disadvantages of diagnostic fields in outbreak research. While there are many convective indices that are utilized, and data can be collected for, not all of these indices are good at predicting such outbreaks at varying lead-time intervals.

Despite these computational challenges, this research would not have been able to have been completed without the computational power provide by the Blue Waters supercomputing center. Each simulation required roughly 16 GB of RAM, and each simulation produced roughly 20 GB of output after post-processing (leading to a final output size for all cases exceeding 2 TB). The project would not have scaled well to systems at the student's university and would have required much longer than one year to complete, simply owing to the volume of simulations required.

#### 4. REFLECTIONS

With any undergraduate research project, student inexperience with research and computing can increase the difficulty in a successful outcome. While the research done through this project was challenging at times, the student gained valuable experience with programming and high-performance computing, equipping the student for future research in big data and supercomputing,



currently among the highest priorities in meteorological research. In particular, supercomputing has become highly prevalent in multiple public and private disciplines of atmospheric sciences. This research opportunity helped set the student apart from their peers by providing critical experience in a High-Performance Computing environment which will have further benefits as the student continues their education. The lessons and processes learned through this research will aid in the student's graduate studies beginning this fall. This research, in conjunction with the Blue Waters Student Internship Program and Summer Experience has given the student diligent training and preparedness for future graduate work and eventual career working with big data in atmospheric sciences.

## 5. SUMMARY AND CONCLUSIONS

It has been well established that tornadic outbreak events can be extremely difficult to forecast, especially several days in advance [1, 2, 3]. While efforts have been made to enhance such forecasts [3, 6, 7, 8, 9, 10], there is still room for growth within this field, particularly in medium-range outbreak forecasting. The primary goal of this research was to obtain an improved understanding of forecast uncertainty within tornado outbreaks at increasing lead times out to 5 days (considered medium-range forecasting). To accomplish this task, five major tornado outbreaks were simulated using stochastic perturbation ensembles at lead times out to 120 hours. Each of the five tornado outbreak cases chosen produced many tornadoes (over 20 per case), with many of them being violent EF2 or greater intensity tornadoes.

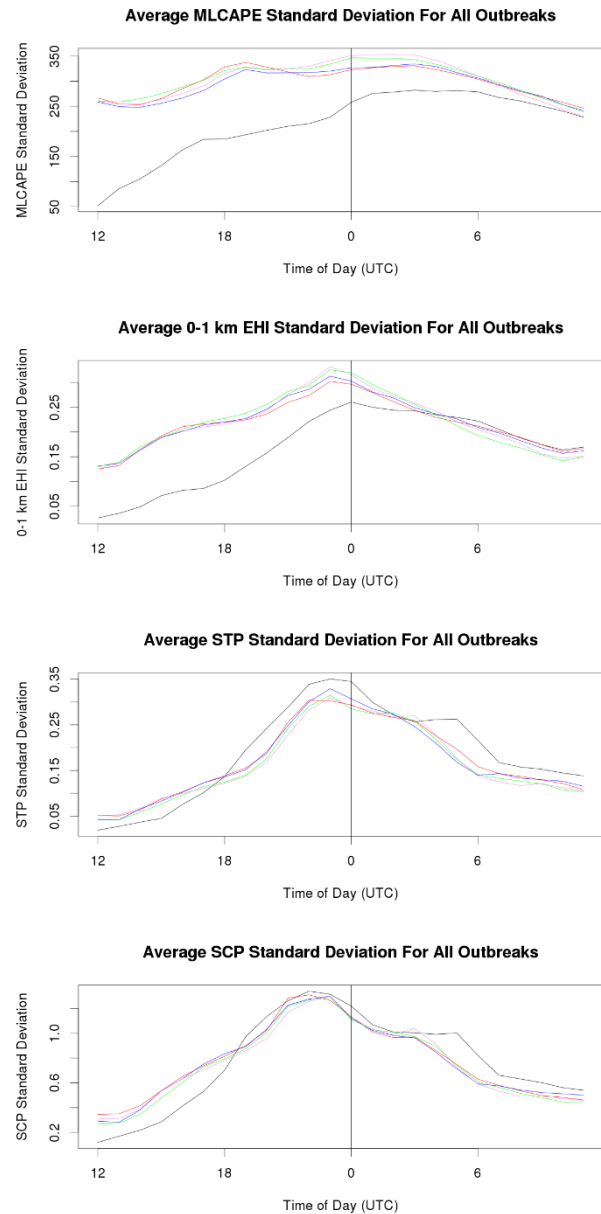
Once the test cases were chosen, utilizing input data from NARR, this data was run through ensemble simulations of WRF at 24, 48, 72, 96, and 120-hour lead time intervals. Each case and lead time included a ten-member stochastically perturbed ensemble to help generate (and thereby quantify) uncertainty in the 250 simulations. Post-processing of the data collected from WRF output files allowed many convective indices to be assessed using moment statistics, including MLCAPE, MLCINH, 0-3 km SRH, 0-1 km EHI, SCP, and STP.

Overall, thermodynamic convective indices seemed to show more variability across cases than index indices such as SCP, STP, and EHI. Moment statistics, such as standard deviation, show a much larger spread of data through thermodynamic convective indices like CAPE than with index indices like the three listed above. As shown in Figure 3, thermodynamic parameter such as CAPE show a bigger shift from the 24-hour forecast period to additional lead times, though the 48-hour to 120-hour lead times still behave similarly. This bigger shift from the 24-hour forecast to later lead times suggests outbreak predictability using CAPE alone is not sufficient, which has been well established in previous work [7].

The major finding in this study was that both thermodynamic and kinematic indices all behaved similarly at 48, 72, 96, and 120-hour lead times, regardless of the moment statistic selected. This was a surprising result given previous research, which typically noted a substantial increase in uncertainty after 72 hours lead time. This suggests outbreak predictability for these major events may not change significantly between 48 hours and 120 hours, which may give forecasters additional potential for predicting these events at longer lead times.

Furthermore, these results can be utilized in the future to continue drawing conclusions about forecasting techniques at short-term and medium-term ranges for tornado outbreaks. Future research is needed to ascertain the generalization of these results to all major tornado outbreak events. Some future work would include utilizing

more test cases and more ensemble WRF forecasts to have more data to compare across tornadic outbreak cases. Additionally, previous research [13] has been done assessing medium-term forecast uncertainty of non-tornadic outbreak cases, and those results could be compared with the results of this study to gain a broader understanding of outbreak predictability across the full spectrum of outbreak events.



**Figure 3. Average standard deviation (SD) statistics for one thermodynamic indices (MLCAPE) and for three kinematic indices (EHI, STP, and SCP).**

## 6. ACKNOWLEDGMENTS

This research was funded through the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. This project was a result of the Shodor Blue Waters Undergraduate Student Internship Program and we are

extremely thankful for their financial, technical, and educational support throughout the project.

## 7. REFERENCES

- [1] Joseph G. Galway. 1977. Some Climatological Aspects of Tornado Outbreaks. *Monthly Weather Review*, 105, 477-484.
- [2] Russell S. Schneider, Joseph T. Schaefer, and Harold E. Brooks. 2004. Tornado outbreak days: An updated and expanded climatology (1875–2003). Preprints, 22nd Conf. on Severe Local Storms, Hyannis, MA, Amer. Meteor. Soc., P5.1.
- [3] Chad M. Shafer, Andrew E. Mercer, Michael B. Richman, Lance M. Leslie, and Charles A. Doswell III. 2012. An assessment of areal coverage of severe weather indices for severe weather outbreak diagnosis. *Weather and Forecasting*, 27, 809-831.
- [4] NOAA National Centers for Environmental Information (NCEI) National Climatic Data Center (NCDC). 2018. U.S. Tornado Climatology (1991-2010). Retrieved from <https://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology>.
- [5] National Weather Service Office of Climate, Water, and Weather Services. 2018. NWS Weather Fatality, Injury and Damage Statistics. Retrieved from <http://www.nws.noaa.gov/om/hazstats.shtml>.
- [6] Charles A. Doswell III and Jeffry S. Evans. 2003. Proximity sounding analysis for derechos and supercells: an assessment of similarities and differences. *Atmospheric Research*, 67-68., 117-133.
- [7] Andrew E. Mercer, Chad M. Shafer, Charles A. Doswell III, Lance M. Leslie, and Michael B. Richman. 2009. Objective classification of tornadic and nontornadic severe weather outbreaks. *Monthly Weather Review*, 137, 4355-4368.
- [8] Andrew E. Mercer, Chad M. Shafer, Charles A. Doswell III, Lance M. Leslie, and Michael B. Richman. 2012. Synoptic composites of tornadic and nontornadic outbreaks. *Monthly Weather Review*. 140, 2590-2608.
- [9] Chad M. Shafer, Andrew E. Mercer, Charles A. Doswell III, Michael B. Richman, and Lance M. Leslie. 2009. Evaluation of WRF forecasts of tornadic and nontornadic outbreaks when initialized with synoptic-scale input. *Monthly Weather Review*, 137, 1250-1271.
- [10] Chad M. Shafer, Andrew E. Mercer, Lance M. Leslie, Michael B. Richman, and Charles A. Doswell III. 2010. Evaluation of WRF model simulations of tornadic and nontornadic outbreaks occurring in the spring and fall. *Monthly Weather Review*, 138, 4098-4119.
- [11] Richard L. Thompson and Roger L. Edwards. 2000. An overview of environmental conditions and forecast implications of the 3 May 1999 tornado outbreak. *Weather and Forecasting*, 15, 682-699.
- [12] Yussouf, N., D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley. 2015. Storm-scale data assimilation and ensemble forecasts for the 27 April 2011 severe weather outbreak in Alabama. *Monthly Weather Review*, 143, 3044-3066.
- [13] Taylor Prislovsky and Andrew E. Mercer. 2017. Using Blue Waters to Assess Non-Tornadic Outbreak Forecast Capability by Lead Time. *The Journal of Computational Science Education*, 8, 30-35, DOI: <https://doi.org/10.22369/issn.2153-4136/8/3/5>.
- [14] Storm Data. 2018. National Climatic Data Center, April 2018.
- [15] NOAA National Severe Storms Laboratory (NSSL). 2018. May 3, 1999 Oklahoma/Kansas Tornado Outbreak. Retrieved from <https://www.nssl.noaa.gov/about/history/may3rd/>.
- [16] NOAA National Weather Service Norman, OK WFO. 2018. The Red River Valley Tornado Outbreak of 1979. Retrieved from <https://www.weather.gov/oun/events-19790410>.
- [17] NOAA National Weather Service State College, PA WFO. 2018. May 31, 1985 Tornado Outbreak. Retrieved from [https://www.weather.gov/ctp/TornadoOutbreak\\_may311985](https://www.weather.gov/ctp/TornadoOutbreak_may311985).
- [18] NOAA National Weather Service Twin Cities, MN WFO. 2018. Summary of June 17, 2010 Record Tornado Outbreak. Retrieved from [https://www.weather.gov/mpx/20100617\\_MN\\_Record\\_Tornado\\_Outbreak](https://www.weather.gov/mpx/20100617_MN_Record_Tornado_Outbreak).
- [19] NOAA National Weather Service Wilmington, NC WFO. 2018. Carolinas Tornado Outbreak: March 28, 1984. Retrieved from <https://www.weather.gov/ilm/CarolinasOutbreak>.
- [20] Fedor Mesinger, Geoff Dimego, Eugenia Kalnay, Kenneth Mitchell, Perry C. Shafran, Wesley Ebisuzaki, Dusan Jovic, Jack Woolen, Eric Rogers, Ernesto H. Berbery, Michael B. Ek, Yun Fan, Robert Grumbine, Wayne Higgins, Hong Li, Ying Lin, Geoff Mankin, David Parrish, and Wei Shi. 2006. The North American Regional Reanalysis. *Bulletin of the American Meteorological Society*, 87, 343-360.
- [21] Song-You Hong and Jeong-Ock Jade Lim. 2006. The WRF Single-Moment 6-Class Microphysics Scheme (WSM6). *Journal of the Korean Meteorological Society*, 42, 129-151.
- [22] G. A. Grell and S. R. Freitas. 2014. A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, 14, 5233-5250.
- [23] Song-You Hong, Yign Noh, and Jimy Dudhia. 2006. A new vertical diffusion package with an explicit treatment of the entrainment processes. *Monthly Weather Review*, 134, 2318-2341.
- [24] Eli J. Mlwaer, Steven J. Taubman, Patrick D. Brown, Michael J. Iacono, and Shepard A. Clough. 1997. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysics Res.*, 102, 16663-16682.
- [25] Jimy Dudhia. 1989. Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *Journal of Atmospheric Science*, 46, 3077-3107.
- [26] Richard L. Thompson, Roger Edwards, and Corey M. Mead. 2004. An update to the supercell composite and significant tornado indices. Retrieved from [https://www.spc.noaa.gov/publications/thompson/stp\\_scp.pdf](https://www.spc.noaa.gov/publications/thompson/stp_scp.pdf)
- [27] Judith Berner, G. J. Shutts, M. Leutbecher, and T. N. Palmer. 2009. A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of Applied Meteorology*, 66, 603-62

# Improvement of the Evolutionary Algorithm on the Atomic Simulation Environment Though Intuitive Starting Population Creation and Clustering

Nicholas Kellas

Department of Chemistry and Biochemistry  
California State University Fullerton  
Fullerton, CA  
nickellas094@csu.fullerton.edu

Michael N. Groves

Department of Chemistry and Biochemistry  
California State University Fullerton  
Fullerton, CA  
mgroves@fullerton.edu

## ABSTRACT

The Evolutionary algorithm (EA), on the Atomic Simulation Environment (ASE), provides a means to find the lowest energy conformation molecule of a given stoichiometry. In this study we examine the ways in which the initial population of molecules affect the success of the EA. We have added a set of rules to the way in which the molecules are created that leads to more chemically relevant structures using chemical intuition. We have also implemented a clustering program that selects molecules that differ from each other from a large pool of molecules to form the initial population. Through testing of EA runs with and without clustering and intuitive population creation, the following success rates were obtained; no intuition and no clustering, 28±3%, no intuition with clustering, 31±4%, with fixed intuition but without clustering, 49±5%, with fixed intuition and clustering, 49±4%, with variable intuition and without clustering, 47±4%, and with variable intuition and clustering, 50±3%. A significant increase in success rate was found when implementing intuitive population creation while clustering the initial population seems to marginally help as the population becomes more diverse.

## Keywords

Evolutionary Algorithm, Agglomerative Clustering

## 1. INTRODUCTION

Determining the structure of a molecule is an important heuristic in understanding its function. As chemical structures become larger, predicting the global minimum becomes increasingly challenging. To find global minimum structures, several strategies have been developed including molecular dynamics [1], Monte Carlo [2], particle swarm optimization [3], random search [4], as well as evolutionary algorithms [5]. Evolutionary algorithms are well suited for chemical structure problems because they can quickly cover large regions of configuration space and can be easily parallelized dramatically improving the search efficiency.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright ©JOCSE, a supported publication of the Shodor Education Foundation Inc.

© 2020 Journal of Computational Science Education  
DOI: <https://doi.org/10.22369/issn.2153-4136/11/2/5>

An Evolutionary Algorithm (EA) is a metaheuristic that uses principles inspired by natural selection to find optimized solutions to complex problems. A population is created from possible solutions and through a process of combining information, optimized solutions can be found [6]. EA's utilize terminology derived from evolution theory such as individual, parent, offspring, and fitness. The EA we use involves 6 steps [7]: first, a starting population of individuals is generated from possible solutions to the complex problem being computed; second, the individuals are evaluated based on the desired metric and ranked from best solution to the problem to worst; third, two of these solutions are then selected to undergo the fourth step, a recombination of information to generate a new solution made from parts of the information from each; fifth, the new solution is compared to the population of solutions, and the worst solution is removed from the population; lastly, a check is done to evaluate if a completion condition has been met, and the selection, crossover, and re-evaluation steps are repeated until one of these conditions is met.



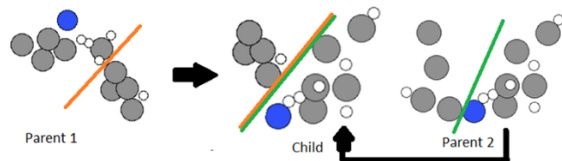
**Figure 1. Steps of an evolutionary algorithm: 1. Generate Starting Population, 2. evaluate the starting population energy, 3. select two candidates to undergo crossover, 4. crossover of two molecules via cut and splice, 5. re-evaluate population to determine if improvement has occurred, and 6. check for completion condition (as the end of the re-evaluation step). This process loops from selection through check for completion until a completion condition is met.**

Originally written on by Alan Turing in 1950, computational scientists have theorized about machines using the principles of evolution to solve problems since the beginning of computational science [8]. By the 1960s “artificial evolution” had become a widely-used optimization method and notably was used by Ingo Rechenberg to generate new aerodynamic wing designs [9]. At the same time, Lawrence J. Fogel developed evolutionary programming in his attempts to create artificial intelligence [10]. The modern evolutionary algorithm was developed by John Henry Holland and published in his 1975 book, "Adaptation in Natural and Artificial Systems" [11]. In the 1980s EA's began to see commercial use by General Electric [12] and Axcelis, Inc. [13] and in 2006 NASA used an EA to develop the evolved antennae [11].

Today, evolutionary algorithms are used in the field of computational chemistry to find the morphology of bulk surface, and nanoparticle systems [14, 15, 16, 17].

The ASE EA begins by generating an initial population of molecules, made from the desired molecular formula by plotting the positions of the atoms as a virtual object. This step can be done stochastically, by assigning the positions of each atom randomly, or semi randomly by specifying properties that the molecules must have such as having set distances between each atom or set angles formed by 3 atoms. By placing restrictions on the kinds of molecules that can be made we are limiting the regions on the potential energy surface that can be represented in the starting population. The algorithm makes a number of these molecules specified by the scientist to form the starting population. Each of these molecules then has its potential energy calculated and they are ranked from highest energy to lowest energy.

Two of the molecules from the initial population are then selected to undergo crossover. This selection is skewed to favor lower energy molecules over higher energy molecules so that more favorable structures are more likely to be replicated [15]. Each of the two molecules are then cut and spliced [7] along a plane and the portion of the molecule on one side of this plane is combined with a counterpart from the other molecule as demonstrated in Figure 2. This process is random and may result in the creation of molecules with incorrect molecular formula. The newly created molecule will either be rejected, and new matched planes will be made, or the molecule will be altered, with extra atoms being removed or added to random locations to correct the molecule.



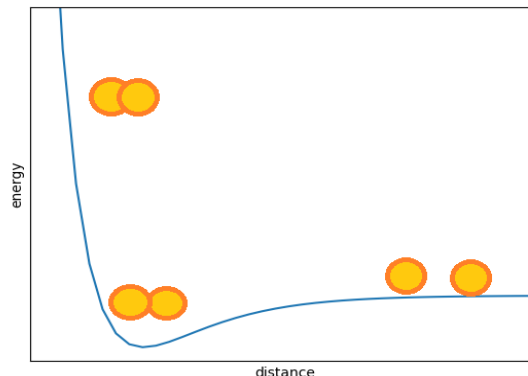
**Figure 2. Crossover of molecules with a portion of the structure of two parent molecules making one child molecule.**

The new molecule is analyzed and compared with the population and the molecule with the highest energy is removed. If the newly created molecule is higher in energy than all the molecules in the population, it is discarded, and the next generation is identical to the first. If a molecule from the population is higher in energy than the newly created individual, the highest energy molecule from the population is removed and the new molecule is added. This new generation is again sorted from low to high energy.

Through this process, the energy of the population will either stay the same or be lowered with each new generation. The selection, crossover, and analysis steps of the EA will be repeated until a set goal is met. This goal can be a number of generations, a desired energy being reached, or convergence, where the same lowest energy molecule is created frequently enough to suggest the population has become trapped into some local minima and cannot escape. This minimum may be the global minima, or it may be a local minimum.

The most computationally intense portion of the EA is the analysis step, in which the candidate molecule is optimized into its nearest local minimum energy structure and its energy is calculated. The energy calculation can be done in many ways depending on the accuracy needed and the complexity of the system being calculated.

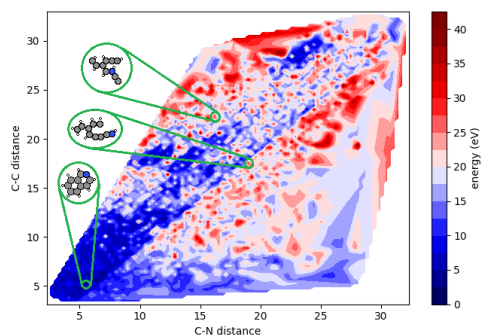
Although energy calculators are effective in locating local energy minima, for large molecules, it becomes increasingly more costly for them to find the global energy minimum. When a molecule is diatomic, containing only two atoms, there is only one optimized shape for it to be in, shown in Figure 3. This shape is the point where the intermolecular attractions most overwhelm the repulsions.



**Figure 3. Morse potential diagram showing the energy of a diatomic molecule in response to the distance between the atoms.**

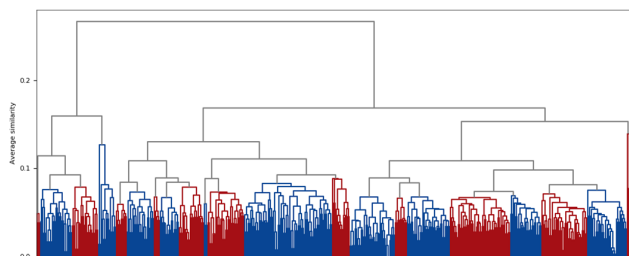
When dealing with multi-atom molecules, the interactions between the atoms becomes increasingly complex and multiple configurations can be created that are lower in energy than all other nearby structures; these are known as local minima. Even though these structures are stable the molecules are still capable of restructuring into more stable configuration. The only point where this is not true is the global minimum, the structure that is the most stable for the given stoichiometry of atoms. The many ways that the molecule can be arranged can be represented in a  $3N$  dimensional matrix where  $N$  is the number of atoms in the molecule. The dimensions are the  $x$ ,  $y$ , and  $z$  coordinates of each atom. This matrix is the potential energy surface of the molecule, and, when energy is added as an extra dimension, it forms a topographical map of all molecular structures involving the target molecular formula.

It is impossible for a human to view a  $3N$  dimensional matrix for any polyatomic molecule, so a simplified view of the potential energy surface must be designed to be understood. For the  $C_9H_7N$  formula, a representation of the surface is shown in Figure 4 from data collected in this study. This 2-dimensional contour graph represents the farthest carbon-nitrogen distance and farthest carbon-carbon distance of each molecule as its  $x$  and  $y$  coordinates with the energy of the molecule shown by its color with red being high energy and blue being low energy. The largest region of low energy occurs when the farthest carbon-carbon and carbon-nitrogen distances are both small, meaning the molecule is coiled into itself in the form of rings. The molecule quinoline is known to be the global minimum of this formula and is composed of two six-member rings meaning it would exist in this low energy area. The region of the graph with greatest carbon-carbon and carbon-nitrogen distance would conversely represent fully stretched linear molecules with nitrogen on one end.



**Figure 4. Simplified Potential Energy Surface for  $C_9H_7N$  using the furthest carbon-nitrogen distance and carbon-carbon distance in a molecule.**

The starting population can be made more diverse by applying clustering [17]. Clustering is the process of grouping individuals based on similarity. In agglomerative hierarchical clustering, each individual is evaluated by analyzing target features. In this case, the features used are the intramolecular carbon-carbon and carbon-nitrogen distances that are compared to each other to determine the similarity between the individuals. When clustering begins each molecule is its own cluster. The program then begins clustering similar molecules together. These clusters are then evaluated based on the averaged characteristics of all the molecules in the cluster. This process continues until a similarity goal is reached, meaning that no clusters are similar enough to be clustered together based on the desired similarity level. This mapping can be seen in Figure 5. When using clustering, a large number of molecules are generated and then a similarity level [7] is found that results in a number of clusters equal to the numbers of individuals required for the starting population. By randomly selecting one molecule from each cluster to form the starting population the diversity of the population will be greater than when the population is made without clustering due to the larger pool of molecules to choose from [18].



**Figure 5. Example agglomerative hierarchical clustering dendrogram at an inconsistency threshold of 2.8.**

When a starting population is made without clustering, there is a likelihood for areas of the potential energy surface to not be represented in the starting population. By generating a large number of molecules and then clustering, the diversity of the starting population can be increased by choosing molecules from all clustered groups. We predict that this increase in diversity will increase the likelihood that molecules similar to the global minimum molecule will be included in the starting population leading to an increase in EA efficiency.

There is, however, a competing element proposed by Oganov [14]. Oganov predicted that when dissimilar molecules undergo cut-and-splice, the offspring will be high in energy due to the merging of incompatible structures. This would mean that the increased diversity of the clustered populations would decrease the efficiency of the EA because of an increased number of unproductive cycles where lower energy molecules are not created and the population does not evolve.

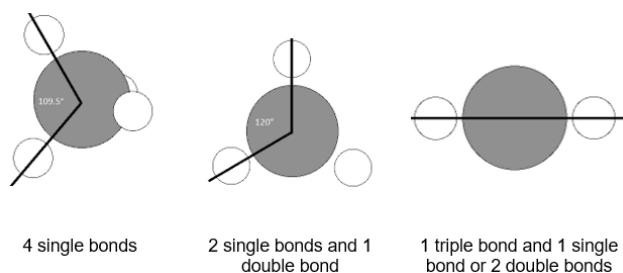
In this study we will show that by building the initial population with chemically intuitive molecules, we can significantly increase the success rate of the GA to find the global minimum. Furthermore, we will show that the role of clustering is more nuanced. In cases where the population is diverse, agglomerative clustering seems to have a minor benefit to the success rate of the GA. When the population is more uniform, agglomerative clustering seems to hinder the success of the GA. This appears to be a demonstration of the trade-off between creating and maintaining a diverse population while being able to form competitive candidate structures which can enter the population.

## 2. EXPERIMENTAL AND COMPUTATIONAL DETAILS

By augmenting the way in which the EA creates its starting population, the efficiency with which it locates the most stable molecular configuration can be improved. For this purpose, the starting population generation was altered to generate molecules using standard hybridized orbital geometries to determine bond angles and bond lengths that would result in more optimized molecules. An agglomerative clustering program was also implemented in which molecules were analyzed for similarity and divided into groups, or clusters.  $C_9H_7N$  was used as the molecular stoichiometry for testing purposes because the global minimum (GM) is known (quinoline), the potential energy surface is well explored, and because the presence of a double ring structure in the GM makes it computationally challenging to find.

The method was written to generate molecules based on rules of molecular geometry using the standard documented orbital hybridization geometries of the atoms. The program begins when it is fed the desired molecular formula. An atom is chosen randomly from those available to fit the formula and placed in position (0,0,0) of the configuration space and then is given a geometry consistent with the element. For example, Carbon can form  $sp^3$  hybridization, where four single bonds are formed 109.5 degrees from each other;  $sp^2$  hybridization, where one double bond and two single bonds are formed 120° apart; or  $sp$  hybridization, where either one triple bond and one single bond, or two double bonds, are formed 180° from each other. Unit vectors signifying these bonds are added to the information stored for this atom. When another atom is chosen, an already-placed atom is selected for it to bond to. The new atom is placed along an available unit vector a distance away from the bonding atom equal to the covalent radii of both atoms added together. The program then checks if the atom is too close to any other atoms, defined as less than 70% of a bond radius as determined above, to prevent energy calculation errors that may occur when atoms are placed inside each other's atomic radii. The new atoms are then similarly assigned a geometry and aligned so that one of the unit vectors is pointed towards the bonded atom. Finally, the unit vectors of bonds that were used in this process are removed from the list of available bonds. Molecules created in this manner have many advantages over molecules created randomly; these molecules require less time to optimize and are far more likely

to form complex, chemically-relevant shapes such as rings and long chains.



**Figure 6. Molecular geometries of carbon.**

Two different forms of intuitive population generation were tested in this study; in fixed hybridization, each hybridization type was given equal likelihood to be chosen for each atom. 20 of these molecules are generated in the non-clustered cases and 500 are created in the clustered cases and then made into 20 clusters with one molecule chosen randomly from each to form the starting population. In the second, called variational hybridization, variable probabilities to select different hybridizations for carbon and nitrogen were utilized in the following manner: the possible hybridizations of both carbon and nitrogen are  $sp$ ,  $sp^2$ , and  $sp^3$ . Each of these hybridizations are assigned a 0, 25, 50, 75, or 100% chance to be selected to be assigned to each atom with the restriction that the total probability to select any of the three is 100%. All potential combinations for the probabilities are sampled. For example, a molecule could be generated with each carbon having a 50% chance to be  $sp^3$  hybridized, a 50% chance to be  $sp^2$  hybridized, and a 0% chance to be  $sp$  hybridized. The nitrogen is independently assigned to have a 100% chance to be  $sp$  hybridized with no chance to be  $sp^2$  or  $sp^3$  hybridized. Not all these probabilities can be used to make a viable molecule with a given stoichiometry. As a result, the program must also be able to identify and skip those cases. When this generation method is used in the non-clustered case, each variation is used to produce 4 molecules, and then 20 are chosen at random from this pool. In the clustered case, 4 molecules are created for each variation and then clustered to generate 20 clusters, and one molecule is chosen from each randomly to form the starting population. These methods are compared to the null case by running random molecule generation trials. For the non-clustered random cases 20 molecules are generated and for the clustered cases 500 molecules are created randomly and then clustered down to 20.

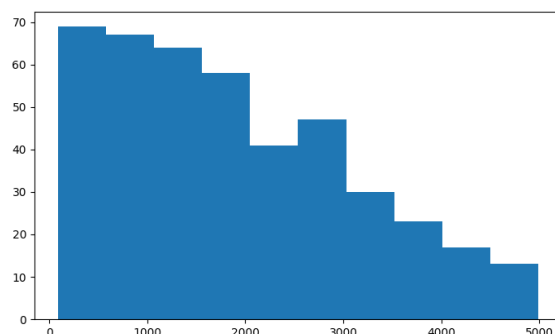
The most obvious alternative to the intuitive population generation method developed here would be one involving the SMILES technique [14] for generating molecular structures. This technique can be used to generate varied molecules easily and with built-in chemical intuition. The reason we chose to write our own program for molecule generation is to have more control over the amount of intuition used in molecule creation and so that the EA can be used to study metallic compounds, as SMILES does not support inorganic complexing [20].

The clustering algorithm takes in a large number of unrelaxed molecules and sorts them based on intermolecular distances. For the test molecular formula,  $C_9H_7N$ , the fact that there was only one Nitrogen was taken advantage of, and the C-N and C-C distances were compared. Hydrogens are ignored because it was found that including C-H and N-H distances did not meaningfully change the

functionality of the clustering process [7]. The factors that will be effective in clustering are strongly dependent on the stoichiometry given, and so additional work must be done to fit clustering to an EA run for different formulas. The molecules are then fitted to centroids as described in the introduction to form a number of clusters equal to the desired starting population size using the methods detailed in Jørgensen et al [7]. One molecule from each cluster is chosen at random to form the starting population.

Molecules were optimized using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [21] and their energy calculated using Density Functional Tight Binding (DFTB) method [22] with the calculator DFTB+ [23]. A more popular energy calculation method is Density Functional Theory (DFT) [13] which calculates the energy of the system using the density of the electron clouds. While this method is a good mix of fast and accurate and is widely used, it is best to have an extremely fast method for testing molecular EAs. For this reason, the empirical method Density Functional Tight Binding (DFTB) is used. DFTB uses a library of ab initio constants to quickly calculate energy at the expense of accuracy. Using this method reduces the time spent optimizing a 17-atom molecule, like the one used for testing this EA, from minutes per molecule to seconds per molecule. Though the method is not highly accurate, DFTB+ was chosen due to the speed at which it calculates energy. This was acceptable to test the hypothesis because high accuracy of calculation was not needed to test the program's ability to move from relatively high energy molecules to low energy molecules [7].

There are two stop conditions for each of these trials. The first is if the EA locates quinolone, and the second is if 5000 cycles have occurred. We set a step count end condition, because we believe that after 5000 cycles the chance of the EA finding quinoline goes down greatly. This is because the EA has likely become stuck in some local minima with all the members of the current population being too similar to find any other structure and escape. This is demonstrated in Figure 7 where we can see that the likelihood of the EA finding quinoline goes down as it reaches higher cycle counts.



**Figure 7. Histogram of the completion steps of the fixed intuitive starting population generation with clustering runs.**

### 3. RESULTS

Each of the six permutations of the experiment were run until they converged to a consistent success rate and are shown in Table 1. The null case, run with random population generation and without clustering, was found to have a  $28\pm 3\%$  success rate.

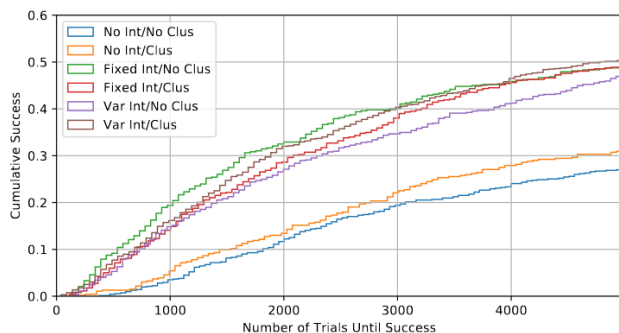
The runs conducted with random population generation and with clustering were found to have a  $31\pm 4\%$  success rate. The runs conducted with fixed intuitive starting population generation and without clustering were found to have a  $49\pm 5\%$  success rate, and the runs conducted with both fixed intuitive starting population generation and clustering were found to have a  $49\pm 4\%$  success rate. For variational intuitive starting population generation without clustering, a  $47\pm 4\%$  success rate was found. Finally, for variational intuitive starting population generation with clustering, a  $50\pm 3\%$  success rate was found.

Data from the above trials is re-expressed in Figure 8 showing the percentage of runs that locate quinoline at or before each cycle count until the 5000<sup>th</sup> cycle. The value of this plot at the 5000<sup>th</sup> iteration corresponds to the Success Rate column in Table 1.

**Table 1. Evolutionary Algorithm Results**

Run type	Completed	Found GM (%)	Success rate (%)	Average	Standard deviation
Random, not clustered	667	185	$28\pm 3$	4295	1330
Random, clustered	462	143	$31\pm 4$	4179	1421
Fixed, not clustered	480	234	$49\pm 5$	3378	1881
Fixed, clustered	618	303	$49\pm 4$	3509	1771
Variational, not clustered	701	330	$47\pm 4$	3615	1761
Variational, clustered	959	484	$50\pm 3$	3431	1811

A representation of the energy of the molecules in the starting populations generated with each method is shown in figure 9. The box plots illustrate the median energy as well as all four quartiles of the energy of the molecules relative to the energy of quinoline. The top three quartiles of the randomly generated starting molecules are in the same range as the top of the highest quartile of the intuitively generated populations. Furthermore, the energy ranges of the intuitively generated starting populations overlap significantly.



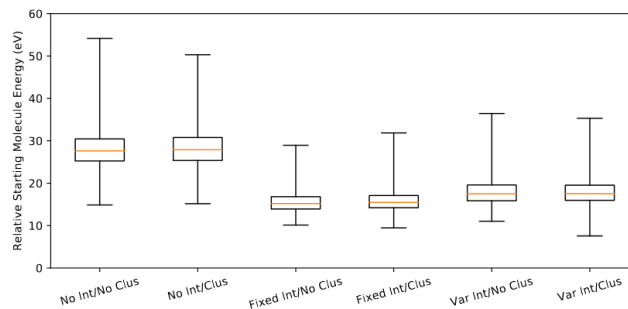
**Figure 8. Success rate for each trial type expressed as percent completed by cycle.**

## 4. DISCUSSION

The implementation of intuitive starting population generation led to a significant increase in success rate, going from a roughly 30% success rate for the two randomly generated cases to roughly 50% success rate for the intuitively generated cases. Intuition led to a sharp increase in the efficiency of the EA by selectively generating molecules in lower energy areas of the potential energy surface. There is also the possibility that the intuitive population generator created molecules with a structure similar to the global minimum. From these results, we can say that seeding the population with chemically intuitive molecules has a positive effect on the EA finding the GM in the case of quinoline.

As reported in Table 1, the average number of candidate structures to find the global minimum was higher for the random cases, but the standard deviation was lower. This lower deviation is theorized to be due to the average completion cycle being closer to 5000 cycles where data collection was set to end. This causes the portion above the average to be cut off at that point, leading to an artificially lower standard deviation.

When clustering is applied to the EA runs without intuitive population generation, the average success rate increases by 3%. When applied to the EA runs with fixed intuitive population generation, little or no change in success rate is found, and when applied to the EA runs with variable intuitive population generation, the success rate increases by 3%. These differences are within the margin of error, so we cannot say that there is a definite difference. We can only remark on the pattern of the differences.



**Figure 9. Box plots for the starting populations of each trial type.**

Figure 8 shows the cumulative success rate as a function of cycle count. When looking at the trials completed using randomly generated starting populations, it can be seen that there was a consistently higher success rate for the clustered case at all cycle counts. Because these molecules were created using less rules than the intuitive populations, we believe they are capable of forming the most diverse populations and as such benefit most from clustering. Clustering was effective in seeding the starting population with molecules whose structures had similarities with the global minimum. It should be noted that the difference between the clustered and non-clustered success rate never varied greater than the margin of error, but the consistent difference in success is indicative of a positive, albeit minor, effect.

For the populations generated using fixed intuitive starting populations, it was seen that the non-clustered trials were much more successful for the first roughly 3000 steps. After this point the clustered runs slowly caught up to them until they had the same success rate. The fixed intuitive populations are the most focused

due to only utilizing one hybridization ratio. This means that the starting population will contain similar molecules. Clustering attempts to maximize diversity, but because this population creation method produces a fairly homogeneous population, it instead creates a population that has a greater challenge forming viable candidates during crossover. This agrees with Oganov's theory [9] that diverse molecules have difficulty producing lower energy offspring when cut-and-spliced. Because of this effect, the non-clustered populations can move toward the global minimum structure more quickly while the clustered populations undergo ineffective crossover until the population filters out its diversity through repeated selection of the more closely related molecules.

For the populations generated using variable intuitive population generation, the clustered populations outperformed the non-clustered cases. These populations were made from many different hybridization ratios, which presumably produced a much more diverse initial population than the fixed intuitive method. Clustering is beneficial in this case to ensure that this diversity is maintained. It appears that Oganov's crossover hindrance overcomes the benefits of clustering only in cases where the population generation method is apparently more uniform.

Clustering initial populations and using crossover to generate new candidate structures appear to be competing processes. Clustering is used to increase the diversity of a population while crossover benefits from homogeneity in the population to form viable candidate structures. The benefit or penalty due to clustering appears to be minor and temporary. In contrast to this, building molecules using chemical intuition appears to significantly increase the success rate of the GA compared to random population generation. This can be attributed to the fact that it takes many cycles for the random population to start creating candidates that exhibit the same geometries that the intuitive generator creates in the starting population. This is a deficit which cannot be overcome. Furthermore, the two different intuitive generation schemes produced different types of starting populations with similar results. This indicates that this process might be robust and tolerant of imperfect schemes which can result in significant improvements in other applications.

Although this experiment was conducted on a single molecular formula, the techniques used are readily applicable to other molecules to test for optimized structures.  $C_9H_7N$  was chosen to test improvements on the genetic algorithm because of the complexity of its structure and the relatively low atom count leading to comparatively low computational time. The technique has been built so that any number of atoms and configurations can be applied including organics and metals. For example, a surface reconstruction of a step edge of  $TiO_2$  is a second test case that has been used previously [7] and would validate the increased cumulative success rate of this methodology. A similar starting population generator can be created based on typical Ti and O hybridizations so that a similar study could be conducted.

## 5. CONCLUSIONS

This study was conducted to examine the effect that intuitive starting population generation and clustering have on an evolutionary algorithm. Testing was conducted using the EA on the Atomic Simulation Environment [24] written for the Python programming language. The effect of intuitive population generation and clustering was tested by running EA's on the molecular formula  $C_9H_7N$  until the known global minima structure, quinoline, was found, or until 5000 cycles were completed.

Intuitive population generation was tested by comparison with random population generation and was split into two forms of intuitive population generation: a form where each hybridization type had equal chance of being selected, and a variable type where molecules were made with various likelihoods of hybridization selection. The effect of clustering on the EA was evaluated by testing each of these population generation methods with and without clustering, creating baseline runs and comparing them to modified runs. It was found that both methods of intuitive population generation had a similar positive effect on the success of the EA, increasing the success rate by approximately 30% in 5000 iterations. Clustering was found to have little effect on the efficiency of the EA, with all results having overlapping confidence intervals between the clustered and non-clustered versions of each generation type. We can remark on trends showing a relationship between the population diversity possible from each molecule generation method and the effect of clustering. The methods that generated the most diverse populations benefited from clustering due to an increased likelihood of the population being seeded with molecules with similar structure to the global minimum structure. However, methods that generate focused populations that include global minimum-like structures were slowed in their ability to find the global minimum due to poor crossover between dissimilar molecules.

## 6. REFLECTION

This project was conducted as part of the Blue Waters Student Internship Program. At the start of the internship, I attended a two-week long computational science institute hosted by the Shodor foundation at the University of Illinois at Urbana-Champaign. There, my fellow interns and I were taught many facets of high performance computing (HPC) including basics like programming in the C programming language and submitting jobs to a supercomputer to compute large problems as well as more complex areas of HPC including many different methods for program parallelization and how to use them to the greatest effect. The information I received from the two-week institute and the internship as a whole have been invaluable for both myself and my fellow researchers in the Groves lab at California State University Fullerton by giving me the knowledge and skill to conduct my project and help others to understand their projects in the field of computational chemistry.

Perhaps just as important to me as the knowledge I received, the friends and connections I have made as a result of this internship have been a massive source of opportunity and community to me. The internship put me into contact with 25 of the most brilliant young computational scientists I have ever met, the staff of Shodor, the National Center for Supercomputing Applications, and the Blue Waters supercomputer and I became part of a community of like-minded individuals that have pushed me to be the best computational scientist I can be, and I can only hope I have done the same for them. The confidence I gained from becoming a part of this community has been the most noticeable benefit of my involvement.

## 7. ACKNOWLEDGMENTS

This study was conducted as part of the Blue Waters Student Internship Program, which is run by Shodor with funding provided by the National Science Foundation. Data was collected using Kepler, a cluster located at California State University, Fullerton. Clustering code was provided by Mathias S. Jørgensen.



## 8. REFERENCES

- [1] Goedecker, S. J. 2004. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* 120, 21, 9911-9917. DOI=10.1063/1.1724816.
- [2] Wales, D. J., and Doye, J. P. K. 1997. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* 101, 28, 5111-5116. DOI= 10.1021/jp970984n.
- [3] Call, S. T., Zubarev, D. Y., and Boldyrev, A. I. 2007. Global minimum structure searches via particle swarm optimization. *J. Comput. Chem.* 28, 7, 1177-1186. DOI= 10.1002/jcc.20621
- [4] Pickard, C. J. and Needs, R. J. 2011. Ab initio random structure searching *J. Phys.: Condens. Matter.* 23, 5, 053201. DOI=10.1088/0953-8984/23/5/053201
- [5] Hartke, B.J. 1993. Global geometry optimization of clusters using genetic algorithms. *J. Phys. Chem.* 97, 39, 9973-9976. DOI=10.1021/j100141a013
- [6] Lohn, J. D., Hornby, G. S., and Linden, D. D. 2005, An evolved antenna for deployment on NASA's Space Technology 5 Mission. In *Genetic Programming Theory and Practice II. Genetic Programming, Vol 8.* U. M. O'Reilly, T. Yu, R. Riolo, and B. Worzel Eds. Springer, Boston, MA.
- [7] Jørgensen, M. S., Groves, M. N., and Hammer, B. 2017, Combining Evolutionary Algorithms with Clustering toward Rational Global Structure Optimization at the Atomic Scale, *J. Chem. Theory Comput.* 13, 3, 1486-1493. DOI=10.1021/acs.jctc.6b01119
- [8] Turing, A. M. 1950, "Computing machinery and intelligence". *Mind.* 59, 236, 433-460. DOI=10.1093/mind/LIX.236.433
- [9] Rechenberg, I. 1973. *Evolutionsstrategie.* Holzmann-Froboog, Stuttgart.
- [10] Fogel, D. B. (editor) 1998. *Evolutionary Computation: The Fossil Record.* IEEE Press., New York, NY
- [11] Holland, J. 1992. *Adaptation in Natural and Artificial Systems.* MIT Press. Cambridge, MA.
- [12] Aldawoodi, N. 2008. *An Approach to Designing an Unmanned Helicopter Autopilot Using Genetic Algorithms and Simulated Annealing.* Doctoral Thesis. University of South Florida
- [13] Evolver: Sophisticated Optimization for Spreadsheets. 2019. Palisade. <https://www.palisade.com/evolver/>
- [14] Lyakhov A.O., Oganov A.R., and Valle M. 2010. How to predict very large and complex crystal structures. *Comp. Phys. Comm.* 181, 9, 1623-1632. DOI= 10.1016/j.cpc.2010.06.007
- [15] Vilhelmsen L. B., and Hammer, B. 2014. A genetic algorithm for first principles global structure optimization of supported nano structures. *J. Chem. Phys.* 141, 4, 044711. DOI= 10.1063/1.4886337.
- [16] Zhai, H., and Anastassia N. Alexandrova. 2017. Fluxionality of Catalytic Clusters: When It Matters and How to Address It. *ACS Catal.* 7, 3, 1905-1911. DOI= 10.1021/acscatal.6b03243
- [17] Lones, M. A., and Tyrrell, A. M. 2007. Regulatory Motif Discovery Using a Population Clustering Evolutionary Algorithm. *IEEE/ACM Transactions On Computational Biology And Bioinformatics.* 4, 3, 403-414. DOI= 10.1109/tcbb.2007.1044
- [18] Lipkowitz, K.B. and Boyd, D.B. 2003. *Reviews in Computational Chemistry, Vol. 18.* John Wiley & Sons, New York, NY.
- [19] Anderson E, Veith GD, Weininger D. 1987. SMILES: A line notation and computerized interpreter for chemical structures. U.S. EPA, Environmental Research Laboratory-Duluth. Duluth, MN
- [20] Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comp. Sci.* 28, 1, 31-36, DOI: 10.1021/ci00057a005
- [21] Fletcher, R. 1987. *Practical methods of optimization* (2nd ed.), John Wiley & Sons, New York, NY
- [22] Elstner, M., Seifert, G. 2014, Density functional tight binding. *Phil. Trans. R. Soc. London, Ser. A.* 372. 20120483. DOI=10.1098/rsta.2012.0483
- [23] Aradi, B., Hourahine, B., and Frauenheim, 2007. Th. DFTB+, a sparse matrix-based implementation of the DFTB method, *J. Phys. Chem. A.* 111, 26, 5678-5684. DOI=10.1021/jp070186p
- [24] Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Jensen, P. B., Kermode, J., Kitchin, J. R., Kolsbjerg, E. L., Kubal, J., Kaasbjerg, K., Lysgaard, S., Maronsson, J. B., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiøtz, J., Schütt, O., Strange, M., Thygesen, K. S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z., Jacobsen, K. W. 2017. The Atomic Simulation Environment—A Python library for working with atoms, *J. Phys.: Condens. Matter.* 29, 27, 273002. DOI= 10.1088/1361-648x/aa680e





April 2020

Volume 11 Issue 2

ISSN 2153-4136 (online)