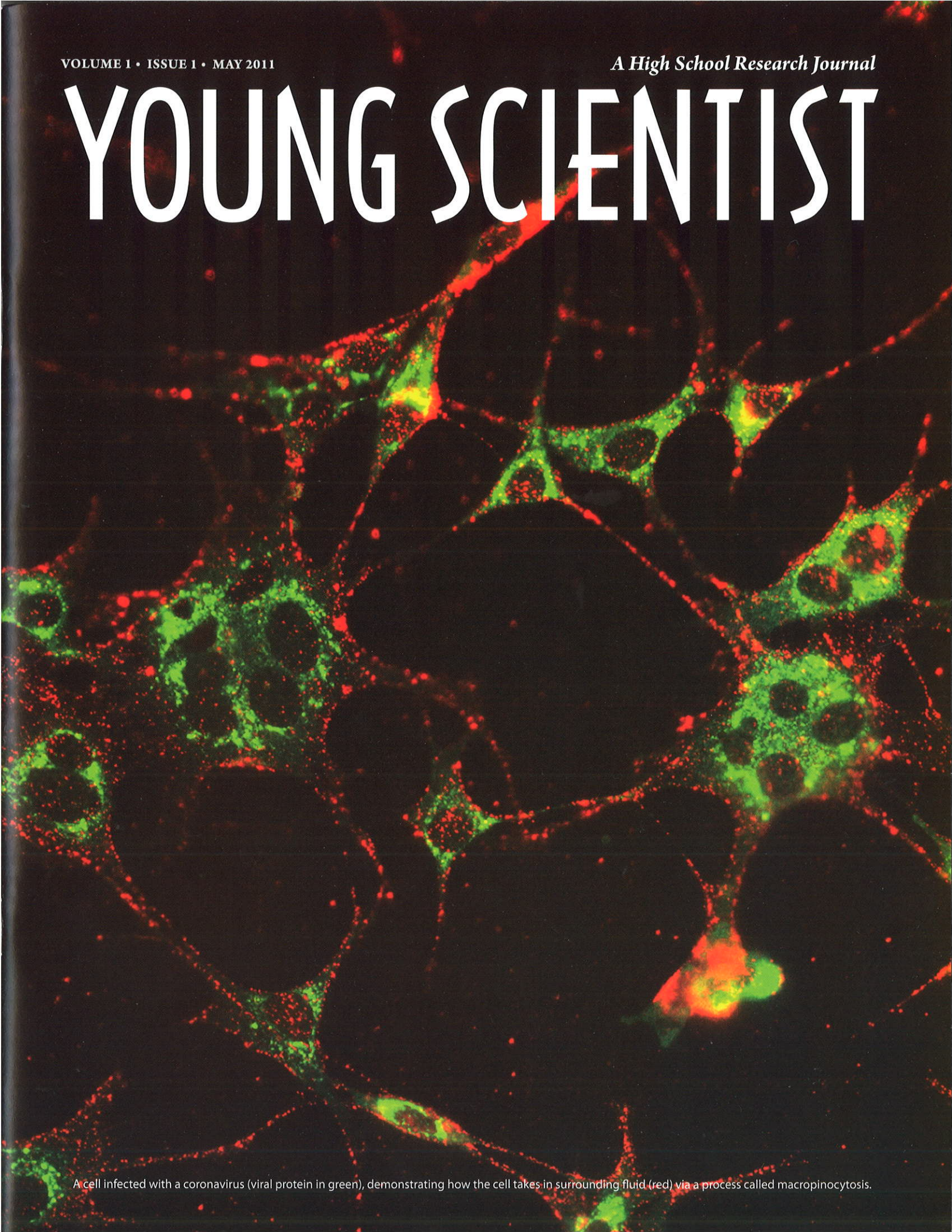
VOLUME 1 • ISSUE 1 • MAY 2011

A High School Research Journal

YOUNG SCIENTIST



A cell infected with a coronavirus (viral protein in green), demonstrating how the cell takes in surrounding fluid (red) via a process called macropinocytosis.

YOUNG SCIENTIST
VOLUME 1 • ISSUE 1 • MAY 2011

Dear Reader,

Welcome to this inaugural issue of *Young Scientist*, the science research journal of the Vanderbilt Center for Science Outreach. The student authors of these journal articles are the next generation of scientists who are challenging themselves by asking and solving important research questions. Each article has been independently and anonymously reviewed by two graduate students at Vanderbilt University to ensure the integrity and quality of this publication. Topics in this issue of *Young Scientist* span a broad range of interests, including finding better biomarkers for the early detection of pancreatic cancer, understanding preventative methods against coronavirus, and studying speech-language development in kindergarten children.

These outstanding students have dedicated countless hours in the laboratory learning what it means to be a scientist while also contributing unique findings and novel techniques to the scientific community. To this end, I thank the students who have gone above and beyond to submit their work to the *Young Scientist*—your hard work and insight are truly inspirational.

I would also like to take a moment to thank the patrons, listed at left, for their support of this journal. Additionally, I offer gratitude to the principal investigators, listed on the back cover, who mentored these students during their research experience. These individuals have encouraged the enthusiastic pursuit of science and knowledge demonstrated by these authors. Finally, I would like to thank the editorial board, Chris Vanags, Ph.D., and Jens Meiler, Ph.D., whose vision and leadership have given these outstanding students the opportunity to be recognized for their contributions.

I hope you enjoy the articles presented here. If you would like to know more about *Young Scientist*, submit an article for future consideration, or view any supporting information, please visit *www.youngscientistjournal.org*.

Thanks again and enjoy!

Mary E. Loveless, Ph.D.
Editor

***Young Scientist* Patrons**

We recognize and thank the following financial supporters of this volume of *Young Scientist*:

The Community Foundation of Middle Tennessee
The Orrin H. Ingram Fund

The Frist Foundation

The Shayne Foundation
Jon and Ann Shayne

Kenneth L. Roberts
In honor of Harvey Sperling

Harvey and Catherine Sperling

Financial Executives International
In honor of Jon Shayne

Cover image courtesy of Christopher T. Peek

YOUNG SCIENTIST

Table of Contents

Volume 1 Issue 1 Spring 2011

SUPPORTING INFORMATION.

Figure S1. Image sequence illustrating laser ablation and wound healing in a *Drosophila* embryo.
Figure S2. Illustration of Radial Profile.
Movie S1. Time-lapse video of normal *Drosophila* development.
Movie S2. Time-lapse video of micro-laser ablation and wound healing of *Drosophila*.

REFERENCES.
1. M. D. Abramoff, P. J. Magelhaes, S. J. Ram, *Biopho Inter.* **11**, 36. (2004).
2. M. S. Hutson *et al.*, *Science.* **300**, 145. (2003).
3. X. Ma et al., *Phy Bio.* **6**, 036004. (2009).
4. G. M. Odell et al., *Dev Bio.* **85**, 446. (1981).
5. F. Schöck, N. Perrimon, *Dev Bio.* **248**, 29. (2002).

Hannah Asbell is a student at Hillsboro High School and enrolled in the School for Science and Math at Vanderbilt.

Protein Structure Prediction Using Rosetta

Deandre Q. Baynham, Mert Karakaş, and Jens Meiler

KEYWORDS. Protein, structure prediction, Rosetta

BRIEF. The accuracy of protein structure prediction was evaluated using a Rosetta program.

ABSTRACT: Proteins have various functions in the human body that can be better understood with an accurate model for their structure. There are several methods to determine the structure of a protein experimentally, but these methods are not applicable to all proteins. If the structure of a protein cannot be determined experimentally, computational tools can be applied to predicted structure. In this project, the Rosetta protein structure prediction program was tested on several proteins to determine the accuracy of this protocol for predicting protein structure. The primary sequence of the proteins were input to several programs for secondary structure prediction., then Rosetta created models for tertiary structure using this information. Success of the method was determined by computing the root mean square distance (RMSD) between atoms in the model and in the experimental structure. It appears that smaller proteins have lower RMSD values than the larger ones. This indicates that the protocol is most effective at modeling small proteins, normally less than 150 amino acids in length.

INTRODUCTION.

Since the early years of biochemistry proteins have been the focal point of the field. After the discovery of diastase, the first known enzyme found by Anselme Payen in 1833, many scientists made huge progress in this field, trying to understand the chemical processes inside organisms. Proteins are biological macromolecules that are very important to functions in the body, and are still the main focus at the center of biochemical studies today, including the folding of proteins how folded proteins interact with one another, substrates, drugs, DNA or RNA. Proteins are synthesized in our body from monomeric units called amino acids. Therefore amino are also energy metabolites and essential nutrients.

To understand proteins, one must first consider how the genetic code memorized in a DNA molecule becomes a protein. The first step of this process is known as transcription and begins with the unwinding of DNA by a protein from a group called helicase. Next, a strand of RNA is attached to the unwound DNA, and starts to replicate its nucleotides. After this process is finished, the new messenger RNA molecule has the nucleotides that correspond with those that were on the original DNA molecule. This messenger RNA molecule that is then read out in the ribosome while the transfer in a process known as translation. During this process, RNA recruits amino acids to the messenger RNA. Each group of three nucleotides (codon) of the messenger RNA determines a specific transfer RNA and thereby a specific amino acid. Typically between 80 and 300 amino acids are translated to form one protein. However, smaller proteins with as few as 30-40 amino acids are known, as well as larger proteins with more than 1000 amino acids [1].

Most proteins are made of 20 standard amino acids, except few cases where non-natural counterparts. Although all of the standard amino acids differ a little, they all have the same principal core structure. In all of these amino acids, except proline, there is a carboxylic acid (COO^-) group, a primary amino group (NH_3^+), and a variable R group bonded to a central carbon atom. Since amino acids have a carboxylic group and the primary amino groups, they can act as either an acid or a base.

These amino acids react to form a polypeptide which than folds into secondary and tertiary structure. The most important types of secondary structure include α-helices β-strands. Other regions of the polypeptide stay mostly flexible forming loop sections. The secondary structure elements then come together to form the tertiary structure, or folded structure, of the protein.

The tertiary structure of a protein dictates its function. Therefore, obtaining an atomic resolution tertiary structure model of a protein is crucial in order to get a better understanding of its dynamics and continue further biological studies. However, many proteins of interest evade experimental methods such as X-Ray crystallography and Nuclear Magnetic Resonance (NMR). In such cases, computational programs are used to predict a structural model for these proteins. There are several ways to predict the tertiary structure of a protein computationally. In this project, the accuracy of tertiary structure prediction was evaluated using Rosetta, a commonly used program, via benchmarking over a set of proteins of variety of topologies [2].

MATERIALS AND METHODS.

The primary sequence forms the starting point for protein tertiary structure prediction. By convention the primary sequence is represented as a FASTA file which contain one-letter codes of all amino acids in the sequence. Since Rosetta will be tested on proteins for which the structures were determined experimentally, the experimental structure for each protein in the benchmark set was obtained from the Protein Data Bank (PDB). It, represent the native structure of a protein by individual coordinates of each atom.

Once the FASTA files were obtained, the BLAST program was used, which aligns the sequence to all known sequences and calculates for each position in the sequence how frequently that amino acid was substituted with another

amino acid in a similar sequence. The end product of this program is a profile for each position in the sequence, giving the likelihood for observing each of the 20 natural amino acids in that position.

Once the BLAST profile was generated for the sequence of interest, several secondary structure prediction programs were run on each sequence to predict which stretches of the sequence are likely to be an α-helix or β-strand. These programs were PSIPRED and JUFO, which utilize artificial neural networks, in addition to SAM which uses Hidden Markov Models. The BLAST profile is an input to these programs.

After the completion of secondary structure prediction, the last task required before running Rosetta is the generation of "fragments". The program iterates over each overlapping three and nine residue stretch of the sequence of interest, then looks for similar stretches of sequences, thus fragments, from proteins with experimentally determined structures, and picks 200 of such fragment conformations for each position in the sequence. These fragments are collected in a large file that forms the fragment database.

Following this, the Rosetta program was used to predict 50,000 structural models for each protein in the benchmark set using the secondary structure predictions and fragment files.

The structural models generated by Rosetta were evaluated by looking at root mean square deviation (RMSD) values. The RMSD value of a model indicates how close it is to the native structure. First the RMSD values for 50,000 models were converted to histograms in order to look at not only how close the models got to the native structure, but also how frequently such good models were obtained. In addition, by comparing the histograms for different proteins, it can be analyzed if for proteins of a certain size and topology, Rosetta generated more accurate models.

One would also expect Rosetta to rank good models by RMSD with good scores, in order to be able pick such accurate models out of the 50,000 models. In order to see if this happened, RMSD values of all 50,000 models were plotted against the score that was calculated for that model by Rosetta.

RESULTS.

The protocol was completed for six benchmark proteins. The figures below show the native structure, the best Rosetta generated model, and an RMSD distribution plot for each protein (Figure 1-3). 50,000 models exhibit a large range of RMSD and energy values which is expected for protein structure prediction.

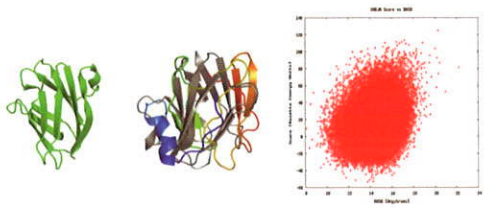


Figure 1. The native structure of 1AAJ, (green), the model with the best RMSD, 8.2Å, (rainbow) superimposed over the native structure (gray), and the RMSD versus energy plot for all 50,000 models.

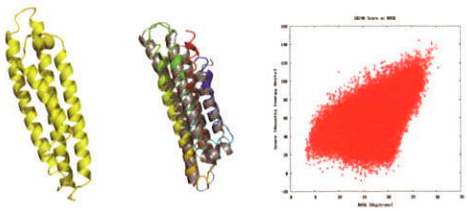


Figure 2. The native structure of 1BZ4 (yellow), the model with the best RMSD, 3.9Å, (rainbow) superimposed over the native structure (gray), and the RMSD versus energy plot for all 50,000 models.

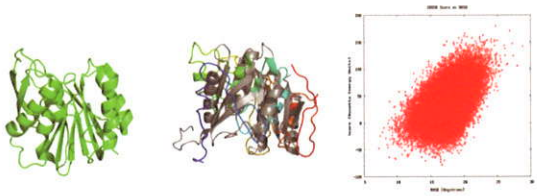


Figure 3. The native structure of 1DUS (green), the model with the best RMSD, 7.3.Å, (rainbow) superimposed over the native structure (gray), and the RMSD versus energy plot for all 50,000 models.

DISCUSSION.

The results indicate that Rosetta can accurately model the topology of smaller proteins. The accuracy of the models generated correlates with the complexity of the protein topology and the sequence length. This is an expected behavior, since for larger proteins more conformations have to be sampled by Rosetta. The accuracies ranged from as low as 3.9Å for 1BZ4 to as high as 12.5Å for 1CHD. Although very accurate models were generated for the smaller proteins, the remainder of models had a wide range of RMSD values. At the same time, the plots indicate that most of these high RMSD value models can be removed using the energy as a filter.

In the future, a more detailed benchmark will be executed including 54 proteins and the evaluations are done also using measures other than RMSD. An additional analysis would be to evaluate how likely good RMSD models can be selected through clustering or energy in the absence of information regarding the native structure. This analysis will also be completed for an alternative protein structure prediction program BCL::Fold, which is being currently developed in Meiler lab, in order to assess the strengths and weakness of both programs.

ACKNOWLEDGMENTS. Thanks to Mert Karakaş for the mentorship and guidance throughout the project. Extended thanks to Dr. Jens Meiler for allowing me to work in his laboratory. I thank the reviewers of my manuscript for many excellent suggestions and the whole Meiler lab at Vanderbilt University.

SUPPORTING INFORMATION.

Figure S1. The native, best model by RMSD and RMSD versus energy plot for 1BGC.

Figure S2. The native, best model by RMSD and RMSD versus energy plot for 1BJ7.

Figure S3. The native, best model by RMSD and RMSD versus energy plot for 1CHD.

REFERENCES.

1. K.T. Simons, et al. J. Mol. Biol. **268**, 209(1997).
2. P. Bradley, et al. Proteins, **61**, 128(2005).

DeAndre Baynham is a student at Hillsboro High School and enrolled in the School for Science and Math at Vanderbilt.

Vanderbilt Center for Science Outreach

www.scienceoutreach.org

How can I get involved?

If you are a middle or high school student interested in research…

- The School for Science and Math at Vanderbilt (SSMV) is a joint venture between Vanderbilt University and Metropolitan Nashville Public Schools (MNPS). The School offers high school students a four-year, interdisciplinary, research-centered learning experience. To learn more, go to *http://theschool.vanderbilt.edu.*
- The research internship program (R.I.P.) offers motivated, rising high school seniors the opportunity to participate in a six-week scientific research internship during the summer at Vanderbilt University, centering on full immersion in a Vanderbilt University research laboratory.
- Middle school students may attend summer camps for both girls (G.A.S.) and boys (B.E.S.T.) to learn scientific investigation and foster confidence in scientific achievement. To learn more, go to *www.scienceoutreach.org.*

If you are a Vanderbilt graduate student or post-doctoral fellow…

- *Young Scientist* is always looking for reviewers for the next issue. For more information please contact Mary Loveless, Ph.D., at *mary.loveless@vanderbilt.edu.*
- Scientist in the Classroom offers graduate students and postdoctoral fellows an opportunity to collaborate with Nashville middle or high school educators in developing and implementing hands-on, inquiry-based activities and assisting student research projects. To learn more, go to *www.scientistintheclassroom.org.*

If you are a Vanderbilt principal investigator…

- You can get involved by mentoring a student in your laboratory. The CSO can help integrate broader impact statements into your research proposals. For more information, contact Jennifer Ufnar, Ph.D., at *jennifer.ufnar@vanderbilt.edu.*
- You can share your research and laboratory experience with students at the School for Science and Math at Vanderbilt (SSMV). For more information, contact the director, Angela Eeds, Ph.D., at *angela.eeds@vanderbilt.edu.*

If you would like to be a Young Scientist patron in future volumes…

- To those individuals who recognize the importance of supporting these students' endeavors and successes and would like to make a financial contribution, please email the editor directly at *mary.loveless@vanderbilt.edu.*

About us…

Young Scientist is a publication of the Vanderbilt Center for Science Outreach (CSO) under the directorship of Virginia Shepherd, Ph.D. The goal of this publication is to recognize the outstanding achievements of high school scientists. The Vanderbilt CSO is dedicated to enhancing scientific and technological literacy through the establishment of unique partnerships between Vanderbilt University scientists, K–12 educators and students, and the local and global science communities.

GIRLS & SCIENCE CAMP AT VANDERBILT UNIVERSITY

BEST Boys Exploring Science & Technology at Vanderbilt University

SCIENTIST IN THE CLASSROOM

SCHOOL FOR SCIENCE AND MATH AT VANDERBILT

Vanderbilt Center for Science Outreach
806 Light Hall
Nashville, TN 37232-0670

YOUNG SCIENTIST

We recognize and thank the following Vanderbilt University principal investigators and laboratories published in this volume of *Young Scientist:*

M. Shane Hutson, Ph.D.
Department of Physics

Jens Meiler, Ph.D.
Department of Chemistry

Scott Guelcher, Ph.D.
Department of Chemical and Biomolecular Engineering

Dan Funk, Ph.D.
Department of Biological Science

Ariel Deutch, Ph.D.
Department of Psychiatry

Pampee Young, M.D., Ph.D.
Department of Pathology

Kevin Seale, Ph.D.
Department Biomedical Engineering

Hal Moses, M.D.
Department of Cancer Biology

Maureen Gannon, Ph.D.
Department of Medicine

Mark Denison, M.D.
Department of Pediatrics

Terry Page, Ph.D.
Department Biological Sciences

Sohee Park, Ph.D.
Department of Psychology

David McCauley, Ph.D.
Department of Biological Sciences

Willi Honegger, Ph.D.
Department of Biological Sciences

Melanie Schuele, Ph.D.
Department of Hearing and Speech Sciences

John Ayers, Ph.D.
Department of Earth and Environmental Sciences

We would like to acknowledge the following funding agencies:

- Vanderbilt University Medical Center
- Metropolitan Nashville Public Schools
- Nashville Alliance for Public Education
- National Center for Research Resources (NCRR), National Institute of Health (NIH)
  Grant Number 1R25RR024261-01