Supplementary Materials for "Deep Data Analytics for Genetic Engineering of Diatoms Linking Genotype to Phenotype via Machine Learning"

Artem A. Trofimov[1], Alison A. Pawlicki[1], Nikolay Borodinov[1], Shovon Mandal[2], Terry J. Mathews[2], Mark Hildebrand[3†], Maxim A. Ziatdinov[1], Katherine A. Hausladen[4], Paulina K. Urbanowicz[4], Chad A. Steed[4], Anton V. Ievlev[1], Alex Belianinov[1], Joshua K. Michener[5], Rama Vasudevan[1], Olga S. Ovchinnikova[1*]

1. Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, United States
2. Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830
3. Scripps Institute of Oceanography, University of California, San Diego, La Jolla, CA, 92093
4. Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830
5. Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830

† Author is deceased: date of death August 9th, 2018

* Author to whom correspondence should be addressed.

Olga S. Ovchinnikova

Center for Nanophase Materials Sciences

Oak Ridge National Laboratory

1 Bethel Valley Rd

Oak Ridge TN, 37831-6493

ovchinnikovo@ornl.gov

Genetic manipulations and testing.

Figure S1 shows a schematic of the gen modification process, as it is described in Methods section

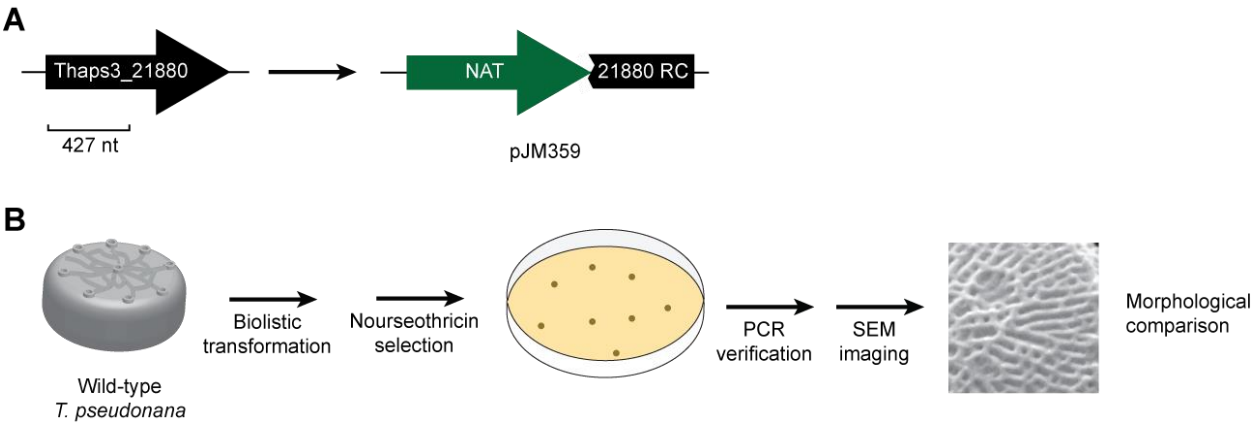

Figure S1. Schematic of the gen modification process. (A) To construct the antisense RNA construct, the first 427 nucleotides of the gene encoding Thaps3_21880 were placed in the reverse orientation downstream of the nourseothricin resistance marker NAT to make plasmid pJM359. (B) The engineered plasmid was transformed into *T. pseudonana* by particle bombardment, followed by selection for nourseothricin resistance. Resistant colonies were verified by colony PCR and then imaged by scanning electron microscopy (SEM).

Class activation maps (CAMs) explaining operation of NN 1.

Figure S2 shows additional examples of CAMs of the diatoms. Images are displayed in pairs (*e.g.*, a and b, c and d, etc.) showing the same frame of the diatom image. Top images always highlight the regions that prompt wild activation class (AC), while bottom images emphasize the regions that prompt modified AC. Whiter color corresponds to higher AC intensity, as shown by the heatbar. CAMs reveal that the NN 1 pays attention to the pores as one of the main features to distinguish wild and modified diatoms and, in such a manner, it emphasizes the importance of

pore-related parameters. It is important to note that achieved accuracy of NN 1 is 94 %. Therefore, as shown in Figure S2a and S2b, there is a small chance of NN 1 making a mistake when assigning a label (wild or modified) to an image. Despite that, NN 1 still highlights pores as a main feature of the morphology.

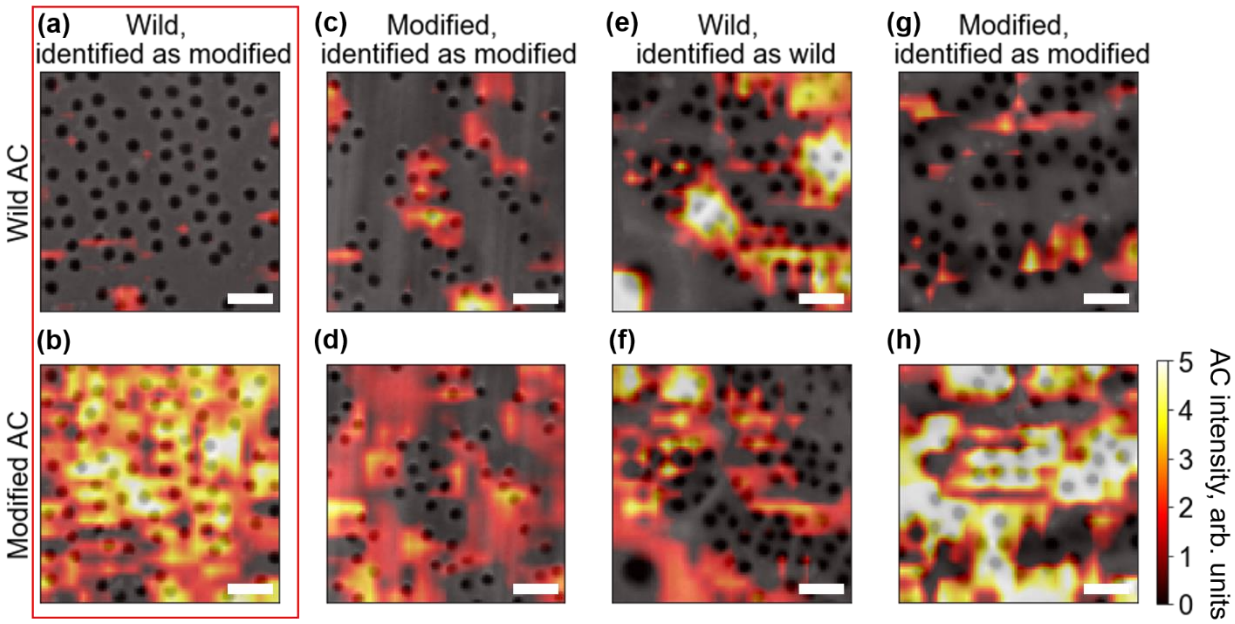

Figure S2. Class activation maps explaining operation of NN 1. Top row shows regions of the diatom images that prompt wild activation class (AC); bottom row shows regions of corresponding images that prompt modified AC. Scale bar is 50 nm.

Additional SEM images of wild diatoms.

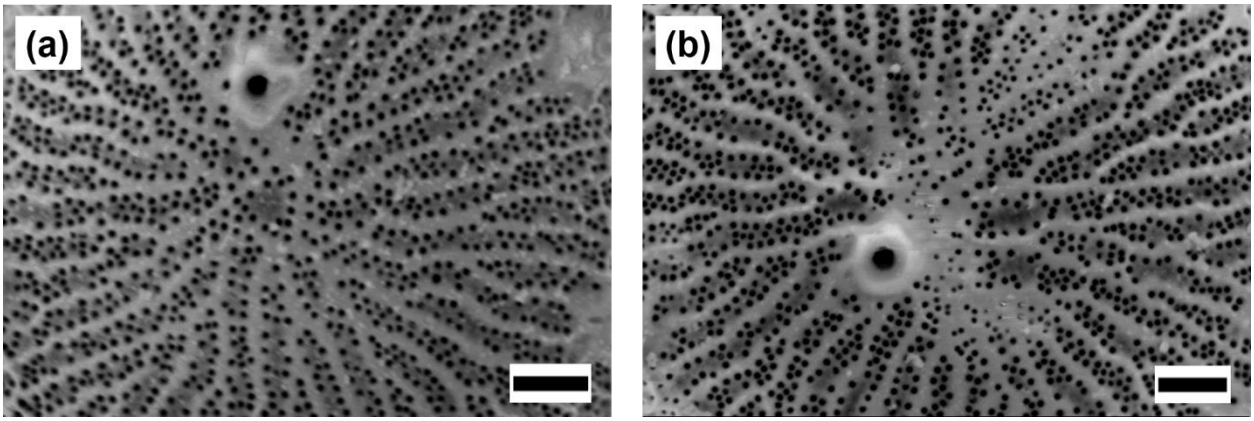

Figure S3. SEM images of wild diatoms. (a) Example 1, (b) Example 2. Scale bar is 300 nm

SEM images of modified diatoms.

In addition to the SEM images of wild diatoms shown in Figure 1, examples of SEM images of modified diatoms can be seen in Figure S4.

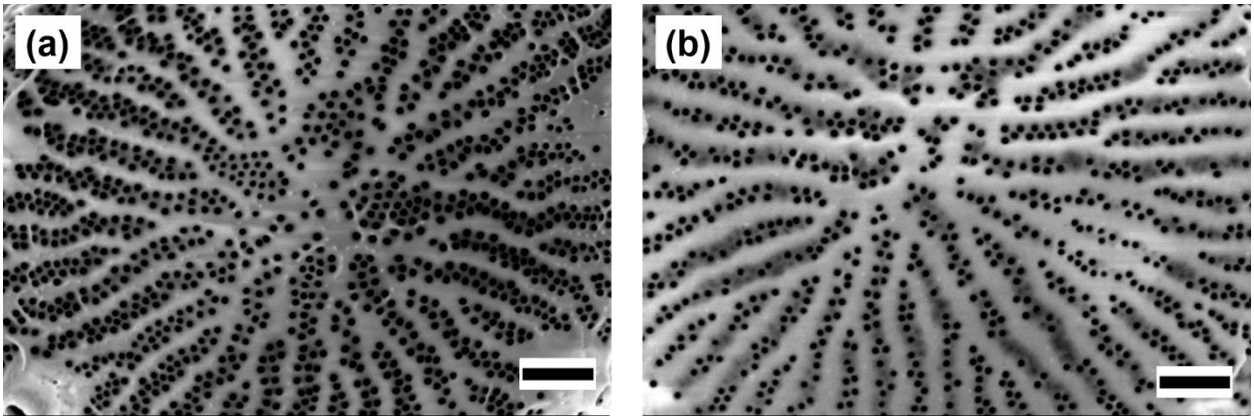

Figure S4. SEM images of modified diatoms. (a) Example 1, (b) Example 2. Scale bar is 300 nm

Acknowledgements

The research was partially conducted at the Center for Nanophase Materials Sciences, which is a DOE Office of Science User Facility. Work by CAS was sponsored by the Department of Energy under the Scientific Discovery through Advanced Computing (SciDAC) RAPIDS project. Work by KAH and PKU was enabled through the Oak Ridge High School Math Thesis Program. The research by JKM, TJM, OSO, AAP, AAT, SM, and MH was sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy. This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC0500OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for the United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).