**Running Title:** Protein Sequence to Structural Conservation

**Title:** **From Protein Sequence to Structure: The Next Frontier in Cross Species Extrapolation for Chemical Safety Evaluations**

\*, aCarlie A. LaLone, abDonovan J. Blatz, acMarissa A. Jensen, dSara M.F. Vliet, aeSally Mayasich,ad Kalie Z. Mattingly, gThomas Transue, gWilson Melendez, gAudrey Wilkinson, gCody W. Simmons, hCarla Ng, iChengxin Zhang, ijYang Zhang

aU.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division, Duluth, Minnesota, USA

bOak Ridge Institute for Science and Education, Duluth, Minnesota, USA

cUniversity of Minnesota Duluth, Swenson College of Science and Engineering, Department of Biology, Duluth, Minnesota, USA

dU.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Scientific Computing and Data Curation Division, Duluth, Minnesota, USA

eUniversity of Wisconsin-Madison Aquatic Sciences Center at U.S. EPA GLTED

fSpecPro Professional Services, Contractor to the US EPA GLTED

gGeneral Dynamics Information Technology, Research Triangle Park, North Carolina, USA

hDepartments of Civil & Environmental Engineering and Environmental and Occupational Health, University of Pittsburgh

iDepartment of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109 USA

jDepartment of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

\*To whom correspondence should be addressed

**Supplemental Data S2**, Text to describe Perl Script for processing

A PERL wrapper script was developed to facilitate the use of I-TASSER and TM-align. The purpose of the script is to automate the gathering of necessary inputs, the formatting of these inputs, along with assisting the submission of I-TASSER and TM-align jobs to the High-Performance Computing (HPC; <https://www.epa.gov/hesc/high-performance-computing-computational-resources-behind-agencys-environmental-modeling-and>) system within the US Environmental Protection Agency Environmental Modeling and Visualization Laboratory.

The first part of this script imports a list of proteins and associated results from a SeqAPASS Level 1 report. Using this information, default priorities are determined based on the hit protein accession IDs, protein names, and taxonomic groups from the SeqAPASS report. These default priorities are used to determine which accessions should be run in I-TASSER to generate protein models. The user can edit these default priorities in order to customize which accessions are run.

This wrapper script also maintains the queue of accessions to run in I-TASSER and enables the user to submit them in batches, of user determined size, to the HPC system. The status of these batched runs can be checked at any time, allowing the progress of these jobs to be monitored from creation to completion. If more HPC computation time is needed, jobs can be resubmitted and will pick up where they left off.

The script uses the results from I-TASSER runs to create the input needed for TM-align and allows the user to submit TM-align runs. Once TM-align runs are completed, the script gathers the results and organizes them for further examination. This initial script provides the bases for development and testing of the pipeline that will be integrated in the SeqAPASS web-based tool interface during the current round of development.