

Data management and curation of research data in academic scientific research environments

Barrie Hayes^{1,2}, James Harroun^{1,2}, Brenda R. S. Temple³

CONTEXT AND BACKGROUND

- INVESTIGATING LARGER SCOPE: DEVELOPING METHODS OF CONSISTENT RESEARCH DATA MANAGEMENT & CURATION.
- PROJECT MOTIVATED BY UNC HEALTH SCIENCES LIBRARY'S E-SCIENCE INITIATIVE.
- PROJECT INITIATED TO ADDRESS EXPRESSED NEED OF STRUCTURAL BIOINFORMATICS (SBI) CORE FACILITY DIRECTOR.
- EXPLORING NEW LIBRARY ROLES IN DATA MANAGEMENT & CURATION.

SBI CORE FACILITY'S DATA PROBLEMS & ISSUES

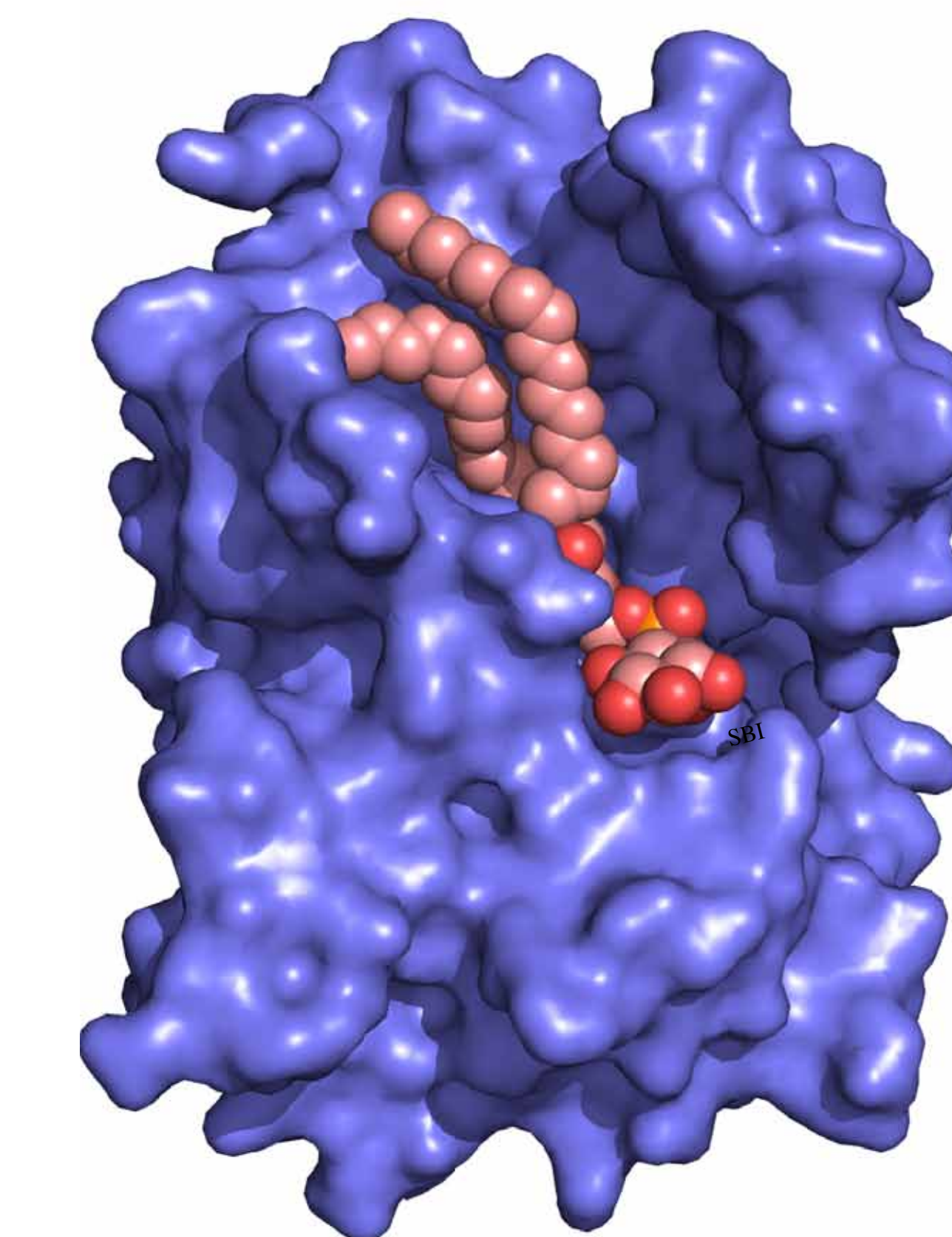
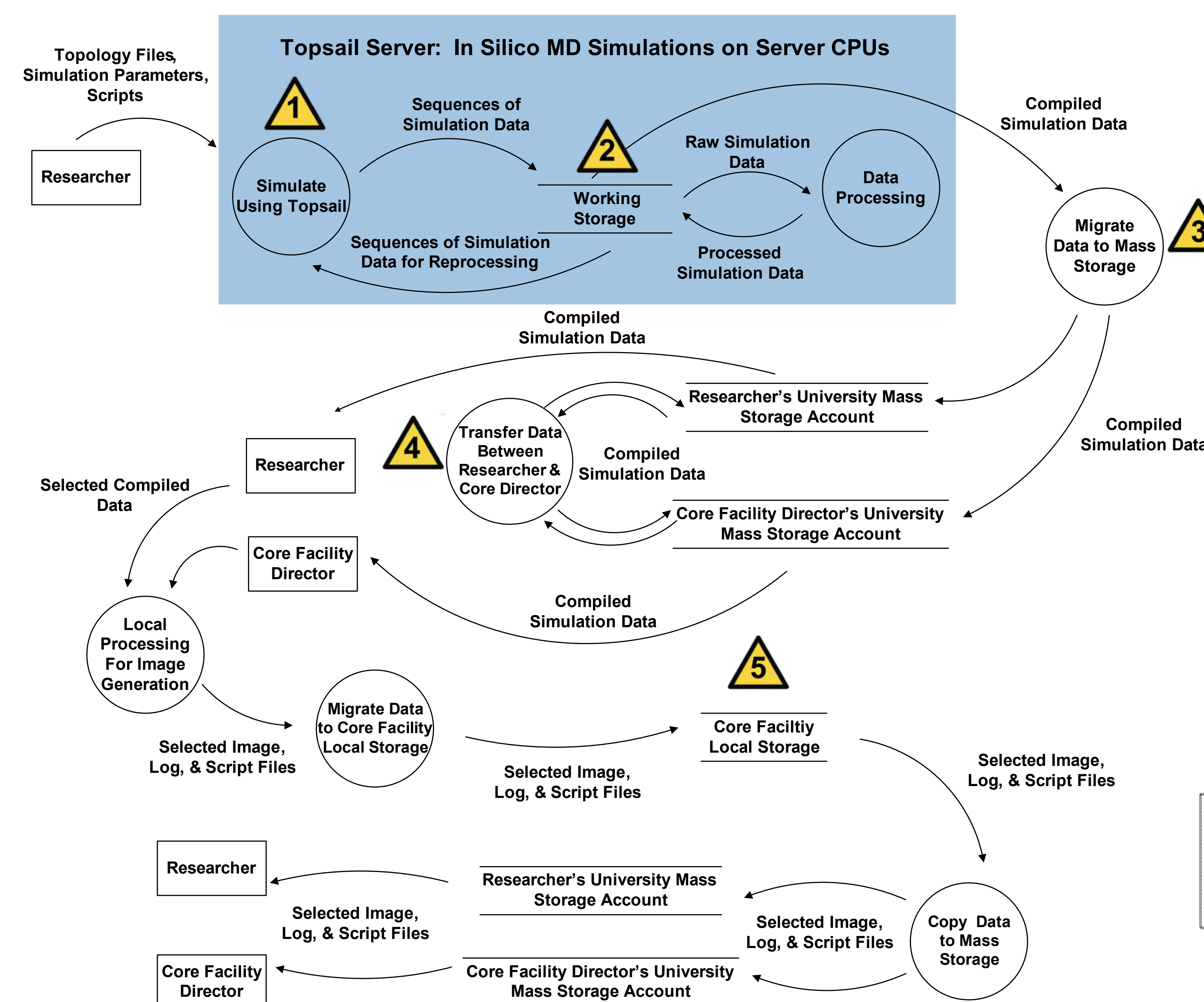
- USERS DO NOT FOLLOW DATA ORGANIZATION OR FILE NAMING CONVENTIONS.
- CORE FACILITY IS A UNIVERSITY-WIDE RESOURCE (NO DIRECT DEPARTMENTAL AFFILIATION/SUPPORT).
- CORE FACILITY IS OPERATING AT FULL CAPACITY.
- SBI DIRECTOR COORDINATES 10-15 PROJECTS AT ANY GIVEN TIME.
- OVERVIEW OF SBI CORE FACILITY'S DATA:

Technique	Typical Project Duration	Typical Data Size	Storage environment(s)
X-Ray Crystallography	1-2 Weeks	1.0GB - 2.0 GB	Local Storage (at Core Facility and Researchers' Labs)
NMR Spectroscopy	1-2 Weeks	0.5GB - 1.0 GB	Local Storage (at Core Facility and Researchers' Labs)
Homology Modeling	1-2 Weeks	100MB - 1.0 GB	Local Storage (at Core Facility and Researchers' Labs)
Molecular Dynamics	6 Mos - 1+ Years	1.0GB - 1.0+ TB	Combination of Working Storage on Topsail, Researchers' Mass Storage, Core Facility Director's Mass Storage, & Core Facility Local Storage

- MULTIPLE TYPES OF TECHNIQUES, ANALYSES, AND VARIOUS DATA TYPES.
- DATA FILE SIZE AND LOCAL STORAGE LIMITATIONS.
- UNIVERSITY ID-ASSOCIATED STORAGE SPACE PROHIBITS EFFICIENT DATA TRANSFER FROM ONE USER TO ANOTHER.

EXAMPLE DATA FLOW FOR MOLECULAR DYNAMICS (MD) SIMULATIONS

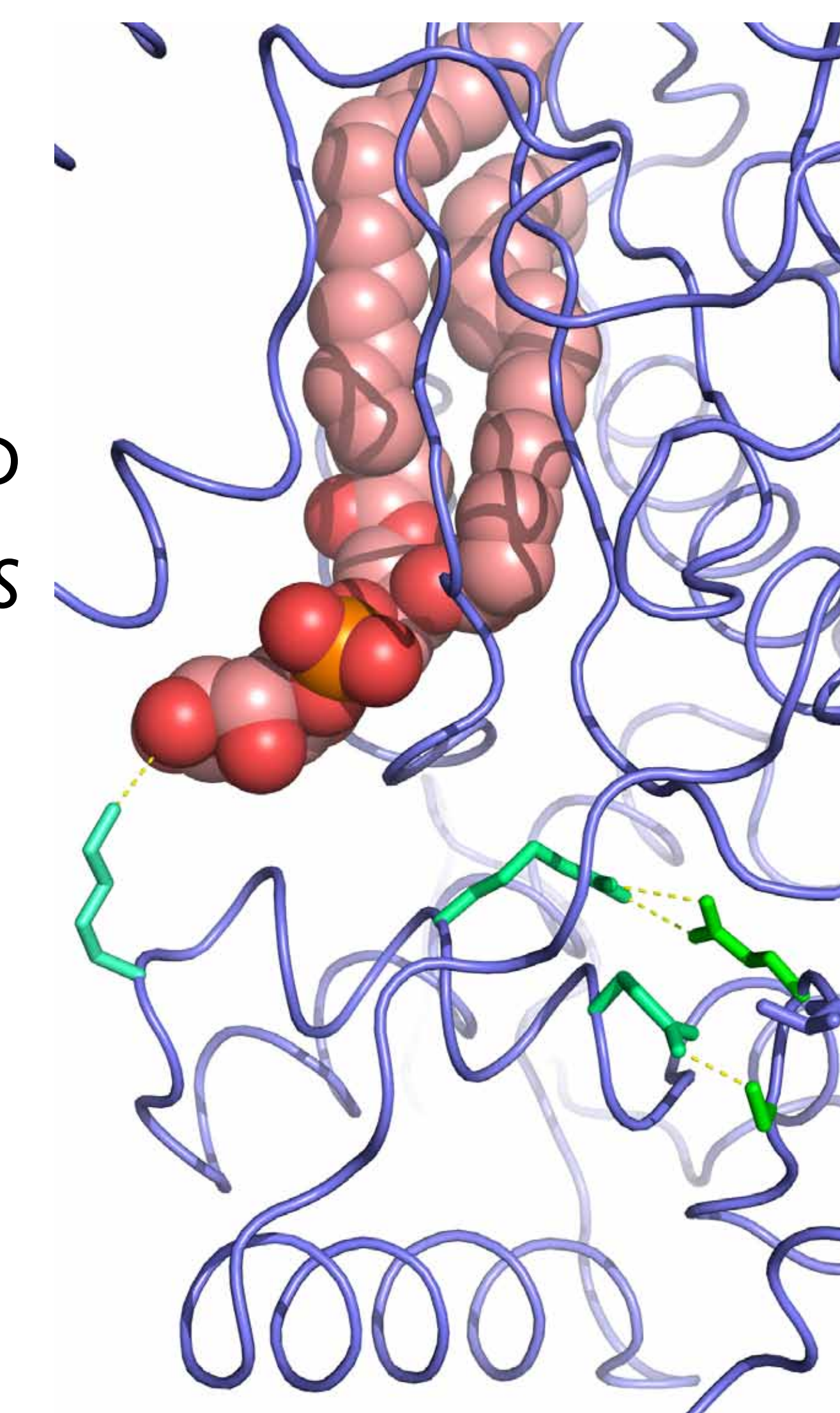
( = IDENTIFIED DATA MANAGEMENT CONCERN)



PROTEIN STRUCTURE
DETERMINED BY X-RAY
CRYSTALLOGRAPHY

REGION OF PROTEIN ANALYZED
DURING MOLECULAR DYNAMICS
SIMULATIONS

MOLECULAR DYNAMICS IS A FORM OF COMPUTER SIMULATION IN WHICH ATOMS AND MOLECULES ARE ALLOWED TO INTER-ACT FOR A PERIOD OF TIME BY APPROXIMATIONS OF KNOWN PHYSICS, GIVING A VIEW OF THE MOTION OF THE ATOMS.



RESULTS AND DISCOVERIES TO DATE

- CREATING A DISTRIBUTED SYSTEM FOR MANAGING DATA AND REPLICATING IT IN A COMMON STORAGE AREA COULD OVERCOME DATA LOSS/DATA TRANSFER ISSUES.
- IMPLEMENTING DATA MANAGEMENT/DATA CURATION METHODS FOR THE SBI CORE FACILITY IS A FIRST STEP IN FULFILLING UNC'S RESEARCH COMPUTING GOALS.
- CREATING A STABLE AND EFFICIENT DATA ENVIRONMENT FOR THE SBI CORE FACILITY AND OTHER RESEARCH GROUPS REQUIRES COLLABORATION AMONG RESEARCH SCIENTISTS, LIBRARIANS AND INFORMATION PROFESSIONALS, AND SUPPORT FROM IT/COMPUTING SPECIALISTS.
- APPLIED FOR UNIVERSITY CANCER RESEARCH FUND GRANT TO SUPPORT DATA MANAGEMENT SYSTEM DEVELOPMENT.

POTENTIAL ROLES FOR UNC HSL

- CREATE/DEVELOP STANDARDIZED DOMAIN-SPECIFIC ONTOLOGIES AND ORGANIZATIONAL SCHEMES.
- DEVELOP PROTOCOLS FOR DATA STORAGE AND ARCHIVING.
- FACILITATE COMMUNICATION AMONG COLLABORATORS AND ADVOCATE FOR INSTITUTIONAL BEST PRACTICES.
- LEVERAGE REPUTATION AND RELATIONSHIPS WITH THE RESEARCH COMMUNITY TO IDENTIFY GROUPS WHO COULD BENEFIT FROM DATA MANAGEMENT ASSISTANCE.
- PROVIDE PROOF-OF-CONCEPT TO UNC ADMINISTRATORS TO PROMOTE EXPANSION OF DATA MANAGEMENT PRACTICES UNIVERSITY-WIDE.

-  **1** USERS DO NOT FOLLOW COMMON NAMING PRACTICES OR FILE ORGANIZATION SCHEMES.
➤ CONDUCT DATA AUDIT. ESTABLISH FILE NAMING, ORGANIZATION SCHEMES, PROTOCOLS, & DETERMINE APPROPRIATE METADATA.
-  **2** WORKING STORAGE IS ONLY TEMPORARY AND FILES ARE DELETED AFTER 21 DAYS OF INACTIVITY.
➤ CONSULT WITH CAMPUS IT RESEARCH COMPUTING SERVICES UNIT TO IDENTIFY BEST PERMANENT STORAGE SOLUTION.
-  **3** MIGRATING DATA TO MASS STORAGE ASSOCIATES IT TO A SINGLE INDIVIDUAL; THERE IS NO SHARED ACCESS.
➤ CONSULT WITH DATA INTENSIVE CYBER ENVIRONMENTS (DICE) GROUP TO IMPLEMENT IRODS FEDERATED DATA GRID.
-  **4** TRANSFERRING DATA BETWEEN USERS IS TIME CONSUMING AND REQUIRES CONSCIENTIOUS PLANNING & COORDINATION.
➤ ESTABLISH DATA TRANSFER PROTOCOLS FOR SBI CORE FACILITY DATA. AUTOMATE DATA TRANSFERS VIA IRODS DATA GRID.
-  **5** CORE FACILITY LOCAL STORAGE IS LIMITED: ARCHIVING SPECIFIC AND IMPORTANT DATA FOR LONG-TERM STORAGE IS NOT POSSIBLE.
➤ AUTOMATE DATA ARCHIVING WITH IRODS DATA GRID AS APPROPRIATE.

¹Health Sciences Library (HSL), UNC-Chapel Hill; ²School of Information & Library Science (SILS), UNC-Chapel Hill; ³R. L. Juliano Structural Bioinformatics (SBI) Core Facility, UNC-Chapel Hill