Interoperability: A Simple Word, Multi-Billion Dollar Problem. A Case Study of the PDB and MDAnalysis

Karen Bekhazi[1,2,3], John Ong[1,2,4], Richard Gowers, Ph.D.[2], Hugo MacDermott-Opeskin, Ph.D.[2]

(1) Station1 (Lawrence, MA), (2) MDAnalysis (Austin, TX), (3) Smith College (Northampton, MA), (4) University of Utah (Salt Lake City, UT)

Background

- MDAnalysis is an open-source Python library capable of analyzing molecular dynamics(MD) simulations [1].
  - Creating a common accessible interface to access raw simulation data of any format
- PDB(Protein Data Bank) - An international repository of protein and molecular data
- MD simulations are **pivotal** in designing new approaches to **medicine, drug delivery, and biofuels**. [2]

Literature Review:

- PDB founded in 1971[3] by Berman et al., stores **208,000+ of structure files**[4] and data.
- 2019 - Formal recognition of PDBx/mmCIF as the master file format of the PDB. [5]
- RCSB PDB(a branch of the PDB) accounts for **1 million annual users** and **$5.5 billion** in use value, alone [6]
- PDBx/mmCIF, despite the machine-readability, lack homogeneity between similar objects, making **parsing and interpretation difficult** [7].

Research Questions:

- How can users and the PDB design for more efficient and reliable data representations of diverse organic structures?
- Why might researchers prefer more varied methods of representing organic matter over a standardized format?
- Can we use precedents in linguistics, stemmatics, and biosemiotics to understand how different file formats impact stakeholders' perceptions of molecular dynamics simulations and the Protein Data Bank?

Hypothesis:

- Various PDB formats are motivated by scientific techniques and discoveries
- Design cannot be limited by current capabilities
- Need to design flexibility for an endearing design

```python
import MDAnalysis
from MDAnalysis.tests.datafiles import PSF, DCD
import numpy.linalg

u = MDAnalysis.Universe(PSF,DCD)
nterm = u.select_atoms('segid 4AKE and name N')[0]
cterm = u.select_atoms('segid 4AKE and name C')[-1]

bb = u.select_atoms('protein and backbone')

for ts in u.trajectory:
    r = cterm.position - nterm.position
    d = numpy.linalg.norm(r)
    rgyr = bb.radius_of_gyration()
    print("frame = {0}: d = {1} A, Rgyr = {2} A".format(
            ts.frame, d, rgyr))
```

Figure 1: An application of MDAnalysis on a trajectory file

Materials and Methods:

- Compiled literature examining the interplay between the PDB and its users using keywords such as molecular dynamics, PDB, computational chemistry, languages, stemmatics, and linguistics. The research papers were collected from e-libraries such as Elsevier, Science Direct, Lens, and the MIT libraries.
- Interviewed users, such as Dr. Beckstein [8], and PDB board members, including Dr. Berman [9], Dr. Vallat [9], and Dr. Peisach [10, 11].

Results & Discussion

- **Experimental methods, knowledge of proteins, data processing methods** [12] and **user bias/error** [13] are major factors that influence the "dialects" of PDB files
  - Dr. Berman: In 2011, there were 27 different "dialects" of PDB files

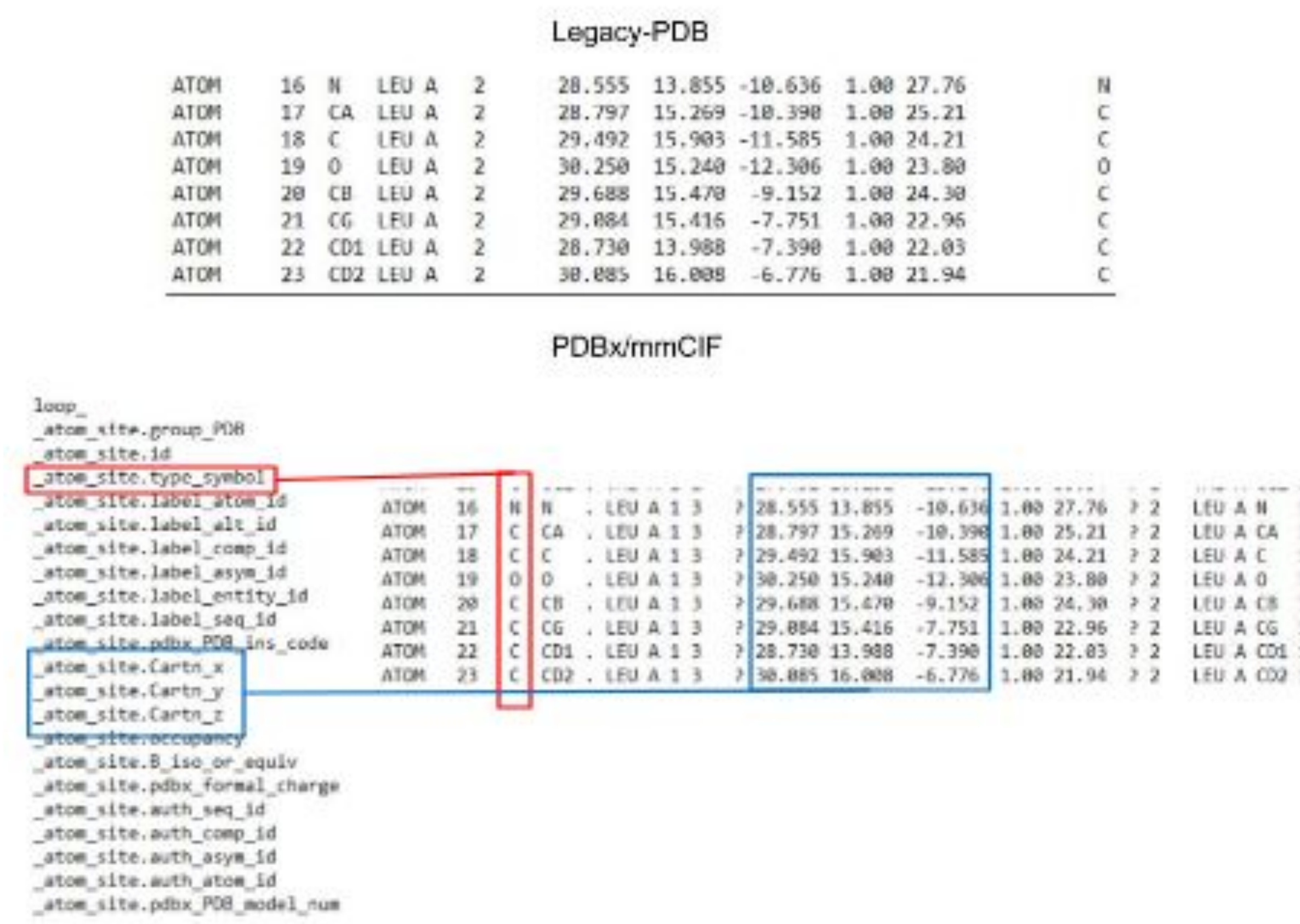

Figure 2: The difference between the legacy PDB and PDBx/mmCIF file format. Smith, R.D et. al, 1997)

- The PDB adheres to a lateral approach, **equally** valuing each stakeholder and lets **scientific advancement** determine its actions.

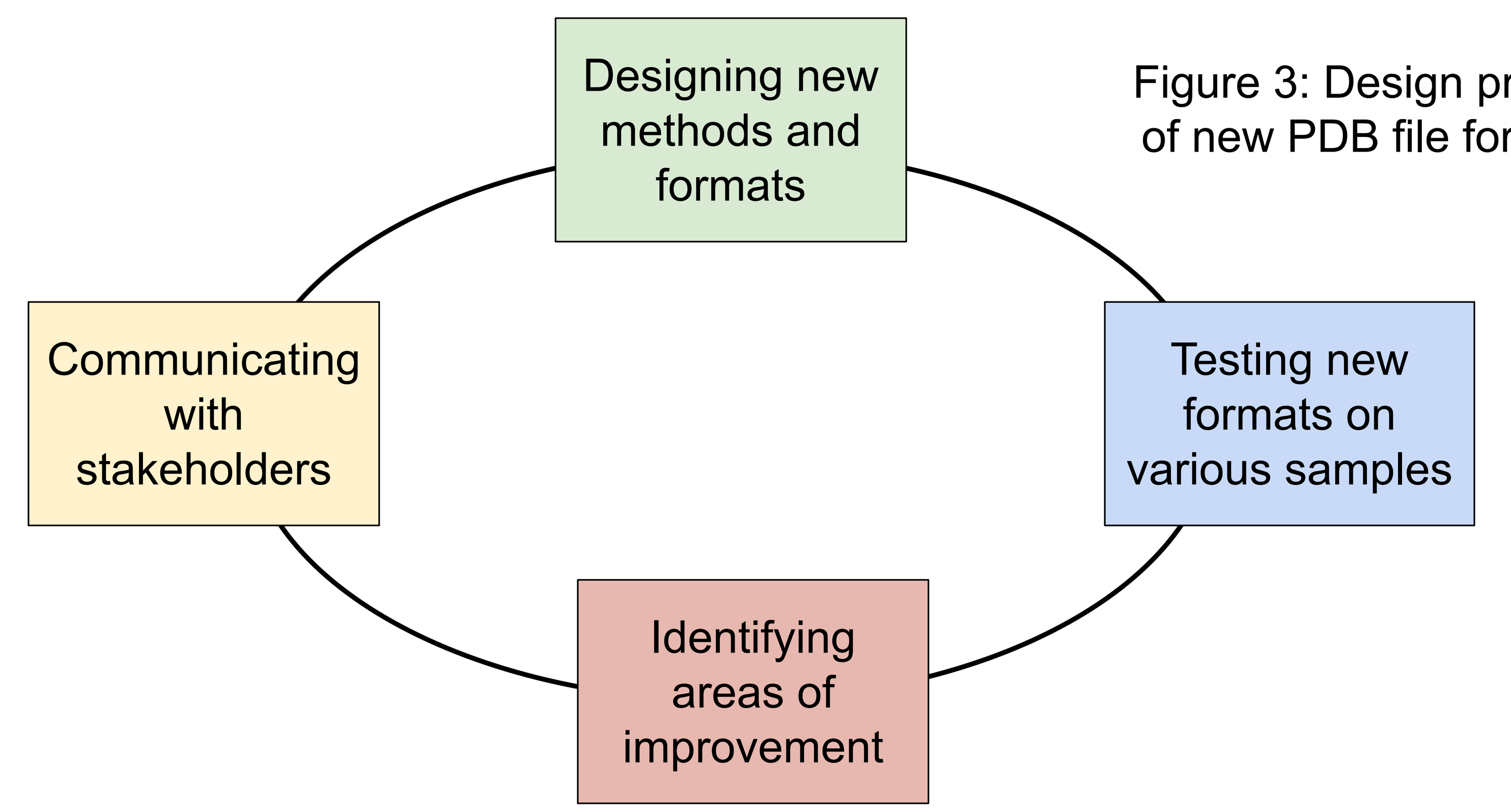

Figure 3: Design process of new PDB file formats.

- Accurate results need accurate data; missteps may cost years of research.
  - MDAnalysis stakeholders work in fields that need accurate data.

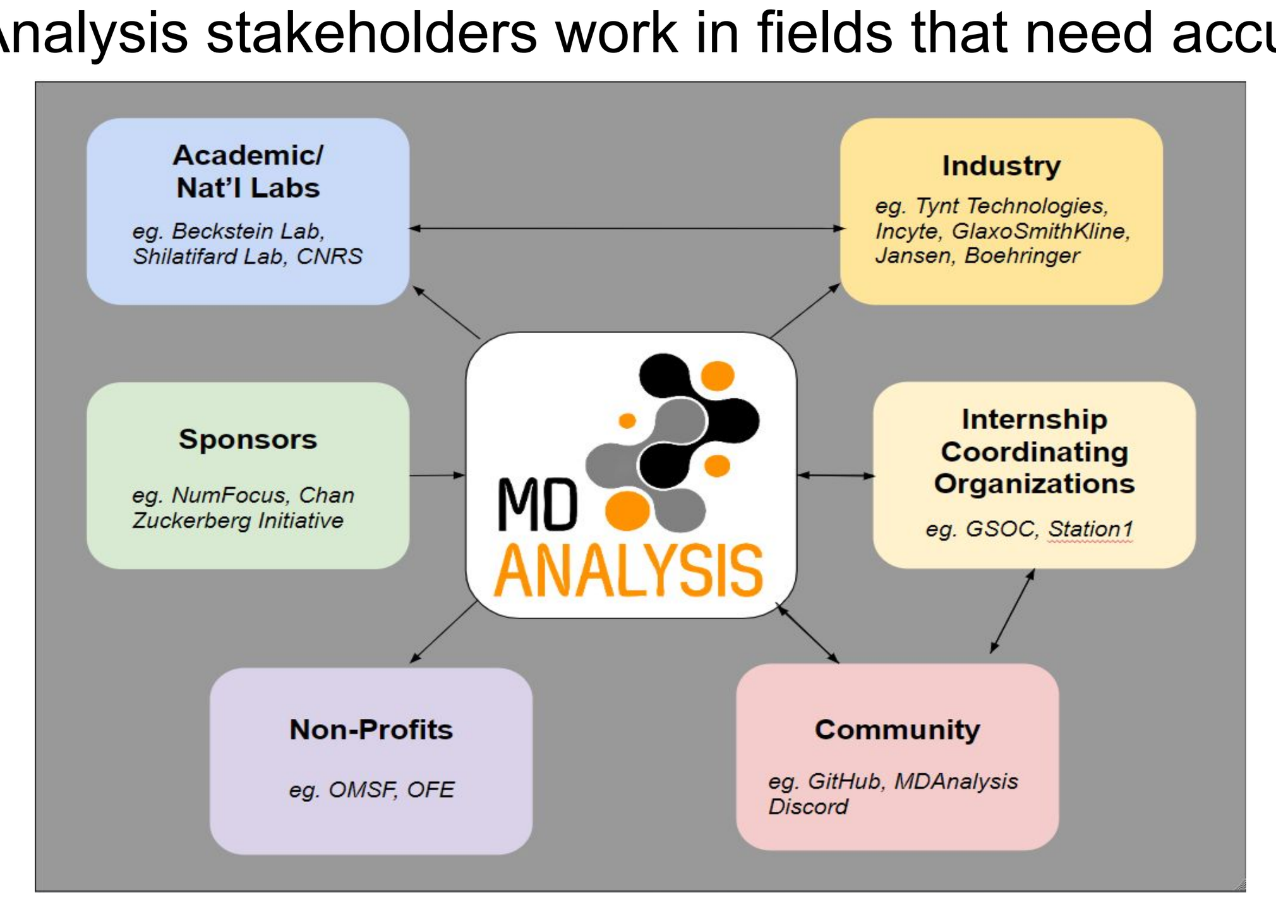

Figure 4: MDAnalysis's stakeholders map.

Social Impact Discussion

- Open-source information and software allows for community-based participatory research (ie. Folding@Home).
- Design for the Margins [14]: increase diversity of users and professionals (developers, etc.).
- Non-profit with **400,000+ downloads** - creates a **strong and sustainable community**.[15]

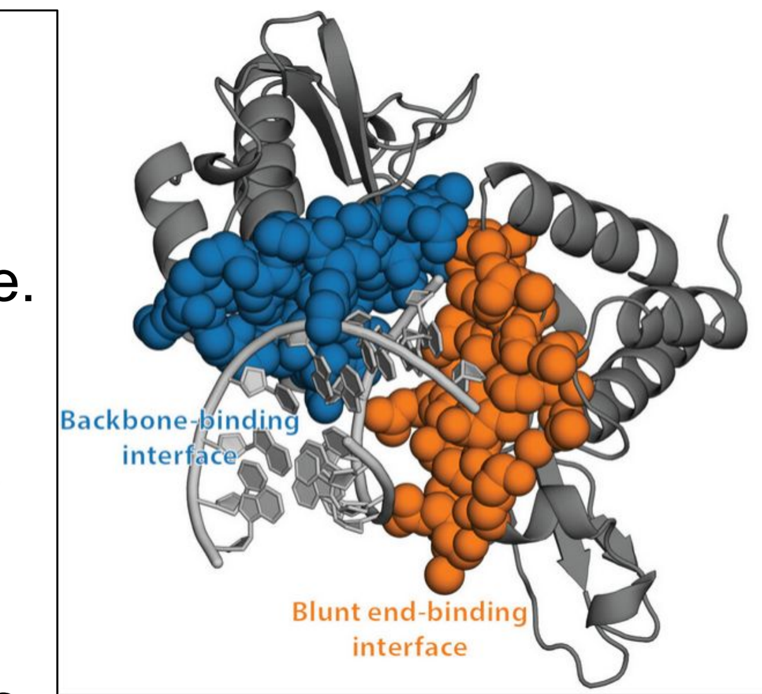

Figure 5: A Folding@Home result showing the value of CBPR via Ebola

Future Work

MDAnalysis can:

- Implement **resilient** approaches to utilizing PDB data aided by our findings

PDB can:

- Continue support for various file formats that **balance human and machine readability**
- Continue to invest in methods of better automation in submission techniques for efficient resource distribution
- Support raw files to expand methodology and integrity
- Continued **support** for **scientific flexibility**

PDB stakeholders can:

- Adhere to the definitions of the PDBx/mmCIF dictionary
- Continued commitment to academic **integrity** and iterative **experimentation**

Integrative Conclusions

These insights provide valuable information on the social implications of both repositories like the Protein Data Bank and Python libraries like MDAnalysis. Understanding the creation of different file formats and their impact on both professionals and other users helps broaden the accessibility of such tools and, ultimately, provides a deeper understanding of their cognitive implications.

Acknowledgements

We'd like to thank all Station1 instructors, Dr. Gowers, and Dr. MacDermott-Opeskin for their efforts this summer. We'd also like to thank Dr. Berman, Dr. Beckstein, Dr. Vallat, and Dr. Peisach for setting time aside to speak to us. Finally, we'd like to thank the MDAnalysis community and the GSOC interns for their continued support.

References

[1] Gowers, Richard, Max Linke, Jonathan Barnoud, Tyler Reddy, Manuel Melo, Sean Seyler, Jan Domański, et al. "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," 98–105. Austin, Texas, 2016. https://doi.org/10.25080/Majora-629e541a-00e

[2] Perilla, Juan R, Boon Chong Goh, C Keith Cassidy, Bo Liu, Rafael C Bernardi, Till Rudack, Hang Yu, Zhe Wu, and Klaus Schulten. "Molecular Dynamics Simulations of Large Macromolecular Complexes." *Current Opinion in Structural Biology* 31 (April 2015): 64–74. https://doi.org/10.1016/j.sbi.2015.03.007.

[3] Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. "The Protein Data Bank." *Nucleic Acids Research* 28, no. 1 (January 1, 2000): 235–42. https://doi.org/10.1093/nar/28.1.235.

[4] *PDB statistics: Overall growth of released structures per year*. RCSB PDB. (n.d.). https://www.rcsb.org/stats/growth/growth-released-structures

[5] Adams, P. D., P. V. Afonine, K. Baskaran, H. M. Berman, J. Berrisford, G. Bricogne, D. G. Brown, et al. "Announcing Mandatory Submission of PDBx/MmCIF Format Files for Crystallographic Depositions to the Protein Data Bank (PDB)." *Acta Crystallographica Section D: Structural Biology* 75, no. 4 (April 1, 2019): 451–54. https://doi.org/10.1107/S2059798319004522.

[6] Sullivan, Kevin P., Peggy Brennan-Tonetta, and Lucas J. Marxen. "Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank." *RCSB Protein Data Bank*, May 1, 2017. https://doi.org/10.2210/rcsb_pdb/pdb-econ-imp-2017.

[7] D'Arminio, Nancy, Deborah Giordano, Bernardina Scafuri, Angelo Facchiano, and Anna Marabotti. "Standardizing Macromolecular Structure Files: Further Efforts Are Needed." *Trends in Biochemical Sciences* 48, no. 7 (July 2023): 590–96. https://doi.org/10.1016/j.tibs.2023.03.002.

[8] Oliver R. Beckstein, interview by Karen Bekhazi, John Ong, July 28, 2023, interview 1, recording, https://drive.google.com/file/d/15K5RLI_5iqvVENyd-0qe1mNn4BIq5EeL/view?usp=drive_link

[9] Helen M. Berman, Brinda Vallat, interview by Karen Bekhazi, John Ong, July 28, 2023, interview 1, recording, https://drive.google.com/file/d/1KDaBj946ww8hWopcviHJuCFHEBIqYQ18/view?usp=drive_link

[10] Ezra Peisach, interview by Karen Bekhazi, John Ong, July 28, 2023, interview 1, recording, https://drive.google.com/file/d/1kLILG76PxfBbQ8FH3pTSy6wKEyzHkMt5/view?usp=drive_link

[11] Ezra Peisach, interview by Karen Bekhazi, John Ong, July 28, 2023, interview 2, recording https://drive.google.com/file/d/1nK9FesOoatYsJecH6UOgW6F40KFcQ4Ub/view?usp=drive_link

[12] Bhat, T. N., Philip Bourne, Zukang Feng, Gary Gilliland, Shri Jain, Veerasamy Ravichandran, Bohdan Schneider, et al. "The PDB Data Uniformity Project." *Nucleic Acids Research* 29, no. 1 (January 1, 2001): 214–18. https://doi.org/10.1093/nar/29.1.214.

[13] Dauter, Zbigniew, Alexander Wlodawer, Wladek Minor, Mariusz Jaskolski, and Bernhard Rupp. "Avoidable Errors in Deposited Macromolecular Structures: An Impediment to Efficient Data Mining." *IUCrJ* 1, no. Pt 3 (April 14, 2014): 179–93. https://doi.org/10.1107/S2052252514005442.

[14]"McDowell_MIT_Civic_Design_Principles_2018.Pdf - LA Docs." Accessed August 2, 2023.https://mynotebook.labarchives.com/doc/view/MjcwLjQwMDAwMDAwMDAwMDAzfDk1Mzk5Mi8yMDgvRW50cnlQYXJ0LzEwMjExNjY5MTh8Njg2LjQ=?nb_id=MTI0MDE4OS42fDk1Mzk5Mi85NTM5OTIvTm90ZWJvb2svNDEwMjk0MzQ2MnwzMTQ4MTczLjU5OTk5OTk5OTY%3D.

[15] *Mdanalysis*. Anaconda.org. (n.d.). https://anaconda.org/conda-forge/mdanalysis