Automating the Fragmentation of Proteins

Gabriela Campbell

Qi Li, Ryan Richards, Theresa Windus

July 11, 2025

Background

- Many-body expansion
  - Problems
  - Optimization
- Computational complexity of energy calculations
  - Polynomial vs. Exponential
- ML approaches
  - Need for training data
- Breaking proteins up by amino acid

$$E_{\text{tot}} \approx E^{(n)}_{\text{eb-MBE}} = \sum_{I} E^{(1)}_{I} + \sum_{I<J} \Delta E^{(2)}_{IJ} + \sum_{I<J<K} \Delta E^{(3)}_{IJK} + \cdots + \sum \Delta E^{(n)}_{IJK\ldots}.$$

Figure: Energy-based many-body expansion [2]

Methodology

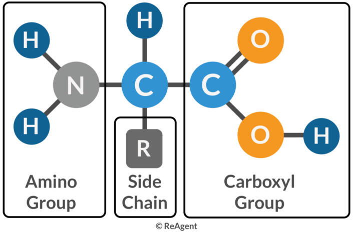
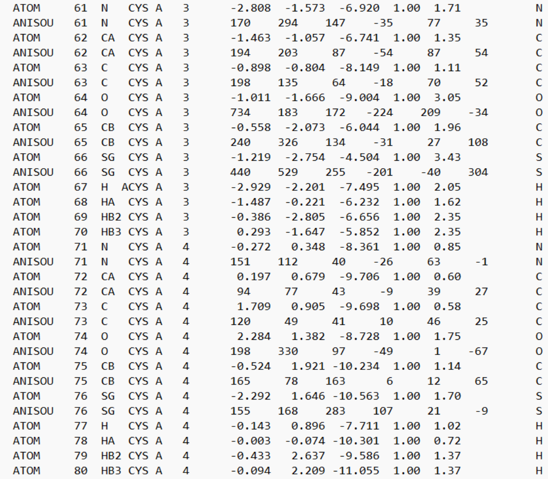


Figure: General amino acid structure [1]

- Segment by amino acid
- Write as .xyz file
- Establish bonds
- Test correctness of created molecule
- Cap atoms where bonds are missing
- Submit to testing

Example PDB File Layout

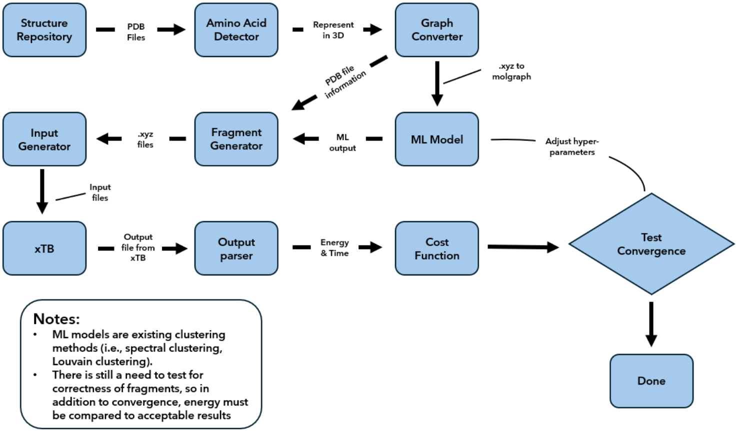


```
ATOM     61  N   CYS A   3      -2.808  -1.573  -6.920  1.00  1.71           N
ANISOU   61  N   CYS A   3      170    294    147    -35     77     35       N
ATOM     62  CA  CYS A   3      -1.463  -1.057  -6.741  1.00  1.35           C
ANISOU   62  CA  CYS A   3      194    203     87    -54     87     54       C
ATOM     63  C   CYS A   3      -0.898  -0.804  -8.149  1.00  1.11           C
ANISOU   63  C   CYS A   3      198    135     64    -18     70     52       C
ATOM     64  O   CYS A   3      -1.011  -1.666  -9.004  1.00  3.05           O
ANISOU   64  O   CYS A   3      734    183    172   -224    209    -34       O
ATOM     65  CB  CYS A   3      -0.558  -2.073  -6.044  1.00  1.96           C
ANISOU   65  CB  CYS A   3      240    326    134    -31     27    108       C
ATOM     66  SG  CYS A   3      -1.219  -2.754  -4.504  1.00  3.43           S
ANISOU   66  SG  CYS A   3      440    529    255   -201    -40    304       S
ATOM     67  H  ACYS A   3      -2.929  -2.201  -7.495  1.00  2.05           H
ATOM     68  HA  CYS A   3      -1.487  -0.221  -6.232  1.00  1.62           H
ATOM     69  HB2 CYS A   3      -0.386  -2.805  -6.656  1.00  2.35           H
ATOM     70  HB3 CYS A   3       0.293  -1.647  -5.852  1.00  2.35           H
ATOM     71  N   CYS A   4      -0.272   0.348  -8.361  1.00  0.85           N
ANISOU   71  N   CYS A   4      151    112     40    -26     63     -1       N
ATOM     72  CA  CYS A   4       0.197   0.679  -9.706  1.00  0.60           C
ANISOU   72  CA  CYS A   4       94     77     43     -9     39     27       C
ATOM     73  C   CYS A   4       1.709   0.905  -9.698  1.00  0.58           C
ANISOU   73  C   CYS A   4      120     49     41     10     46     25       C
ATOM     74  O   CYS A   4       2.284   1.382  -8.728  1.00  1.75           O
ANISOU   74  O   CYS A   4      198    330     97    -49      1    -67       O
ATOM     75  CB  CYS A   4      -0.524   1.921 -10.234  1.00  1.14           C
ANISOU   75  CB  CYS A   4      165     78    163      6     12     65       C
ATOM     76  SG  CYS A   4      -2.292   1.646 -10.563  1.00  1.70           S
ANISOU   76  SG  CYS A   4      155    168    283    107     21     -9       S
ATOM     77  H   CYS A   4      -0.143   0.896  -7.711  1.00  1.02           H
ATOM     78  HA  CYS A   4      -0.003  -0.074 -10.301  1.00  0.72           H
ATOM     79  HB2 CYS A   4      -0.433   2.637  -9.586  1.00  1.37           H
ATOM     80  HB3 CYS A   4      -0.094   2.209 -11.055  1.00  1.37           H
```

Applications and Additional Functionality

- Removing the need for user input
- Testing for correctness in generated .xyz files
- Identifies double and triple bonds within a molecule
- Allows for the generation of large quantities of files

Pipeline



Notes:

- ML models are existing clustering methods (i.e., spectral clustering, Louvain clustering).
- There is still a need to test for correctness of fragments, so in addition to convergence, energy must be compared to acceptable results

Figure: ML Pipeline for Automating Fragmentation of Proteins

SIMCODES Commentary

- Professional development
  - Application of CS, DS, and MA practices
  - Collaboration with a team
- Structure
- Encountering complexities
  - Chemistry's resistance to algorithms
  - Using unfamiliar software like Jupyter notebooks
  - Designing code with modularity in mind

Acknowledgments

This material is based on work supported by the National Science Foundation under Grant Number 2348724. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Thank you to Iowa State University and AMES Lab for hosting this program, and to all mentors and staff I've worked with on this project.

References

[1] Jessica Clifton. *What Are Amino Acids*. 2021. URL: https://www.reagent.co.uk/blog/what-are-amino-acids/.

[2] Stefanie Schürmann et al. “Accurate quantum-chemical fragmentation calculations for ion–water clusters with the density-based many-body expansion”. In: *Phys. Chem. Chem. Phys.* 25 (1 2023), pp. 736–748. DOI: 10.1039/D2CP04539G. URL: http://dx.doi.org/10.1039/D2CP04539G.

Resources

PDB converter repository:

- https://github.com/SIMCODES-ISU/Campbell_Repo

xTB gitHub:

- https://github.com/grimme-lab/xtb

xyz2graph gitHub:

- https://github.com/zotko/xyz2graph/tree/main