Ab-initio synthesis of amino-acids

N Sowmya Manojna [*] **Sahana Gangadharan** [**]

[*] *Indian Institute of Technology, Madras,*
(e-mail: sowmyamanojna@smail.iitm.ac.in)
[**] *Indian Institute of Technology, Madras,*
(e-mail: be17b038@smail.iitm.ac.in)

Abstract: The Urey-Miller experiment, demonstrated for the first time, the prebiotic synthesis of amino acids. Although this result has been reproduced multiple times experimentally, there have been very few in-silico attempts of the same. Our study explores the role of standard Gibbs free energy change as a key parameter in the in-silico synthesis of Glycine in a Miller-like experiment. A network theoretic approach is used to model the experiment and simulated annealing is used to scan the reaction space. Our model has identified key intermediates of glycine synthesis such as formaldehyde, aminoacetonitrile and produced glycine and alanine through the Strecker amino acid synthesis reaction. Our results show that the standard Gibbs free energy of Glycine (-664.58 KJ) was the lowest among all the compounds formed and hence substantiating our hypothesis.

Keywords: Network biology, Amino acid synthesis, Simulated annealing, Miller's experiment, Gibbs free energy, Strecker amino acid synthesis, Radon Graph Generation

1. INTRODUCTION

About seventy years ago, Stanley L. Miller and Harold C. Urey demonstrated the first evidence for the ab-initio synthesis of life. Their path-breaking experiment simulated conditions that are very similar to primitive Earth - atmosphere consisting of reducing gas mixture such as methane, ammonia and carbon dioxide, water as the solvent, acidic conditions and high electric current discharges accounting for lightning. Miller demonstrated that Glycine, α-Alanine, β-Alanine and Aspartic acid were formed from prebiotic Earth conditions (Miller (1953)). Miller's reaction mechanism (Pietrucci and Saitta (2015)) proved that hydrogen cyanide, formamide, aldehydes and ketones are the key precursors in amino acid synthesis. Similar results were observed in other experiments with varied initial gas mixture. The in-silico approach by Saitta and Saija (Saitta and Saija (2014)), using electric field potential as a key parameter has identified formaldehyde and formic acid as the key intermediates in the formation of Glycine.

Since Glycine is the simplest of the 20 naturally occurring α-amino acids, we decided to study its formation in our model. As stable compounds have highly negative standard Gibbs free energy change, we used this as the key parameter to evaluate the compounds formed in our in-silico model. In order to test our hypothesis, we used the thermodynamic data from the Reaction Mechanism Generator (RMG) web server, which is developed jointly by Richard H. West's research group at Northeastern University and William H. Green's research group at MIT.(Gao et al. (2016))

2. METHODS

2.1 Thermodynamic data collection

All thermodynamic data used in this study was obtained from the RMG web server. The data available in RMG is segregated in the form of libraries and the data available in all libraries are represented in one of the two formats - NASA and Group Additivity.

In the NASA format, coefficients of the NASA thermodynamic equations - heat capacity at constant pressure, standard enthalpy change and standard entropy change, for two different temperature regimes were available. The Group additivity format comprises of high charts of heat capacity at constant pressure, standard enthalpy change and standard entropy change as a function of temperature.

As no index of compounds was available in the database and as parsing data from highcharts using Python was tough, data from the NASA format was alone parsed. Since the temperature ranges varied across different libraries, the coefficients and their respective temperature ranges were parsed from the web servers and stored in a CSV format. The data parsed was segregated based on the source library and the compound's label within the library.

2.2 Initial network setup

A graph theory based approach has been used in this study, where the nodes represent the atoms and edges represent the covalent bonds between the atoms. The number of bonds that an atom can form is constrained by the valency of the atom. In order to accommodate multiple bonds between two atoms, we used the `networkx MultiGraph` object. Edge attributes were used to indicate the number of bonds between two atoms.

In order to encapsulate all the above mentioned conditions, an `Atom` class was created. The name of the atom, its atomic symbol and its valency are attributes of the class. Instances of the `Atom` class were used as nodes in our network. Member functions of the class were used to access the maxi-

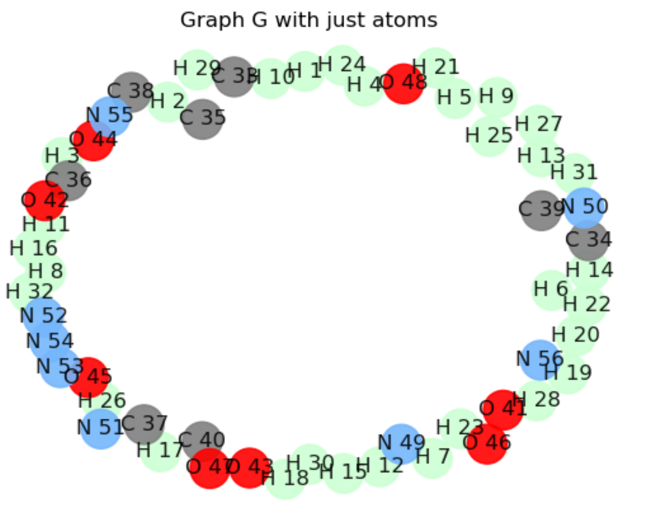


Fig. 1. Initial network with just the nodes - Carbon, Oxygen, Hydrogen and Nitrogen

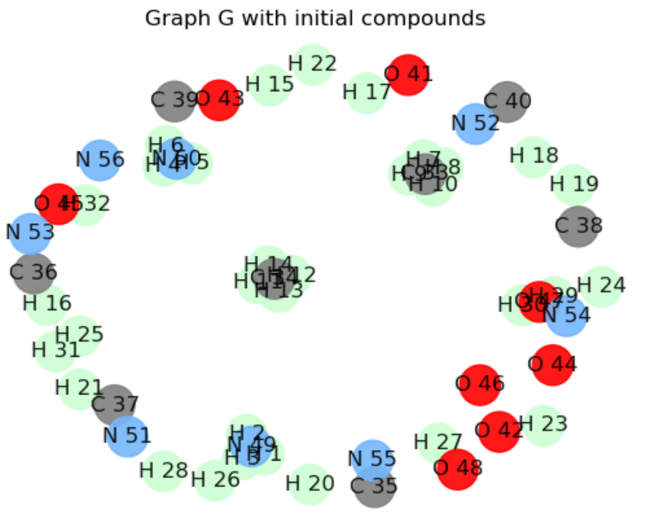


Fig. 2. Initial network with the initial compounds formed - Methane, Ammonia and Water

mum valency of an atom and the valency of the atom at any instant, given the network setup. An external function `fix_valencies` which takes graph as input was used to update the valencies of all the nodes in a graph.

The initial nodes in our network are - carbon, oxygen, nitrogen and hydrogen. Based on the concentration calculations from Miller's original paper (Miller (1953)) and other in-silico approaches (Saitta and Saija (2014)), we began the simulation with H:C:O:N ratio of 4:1:1:1, with a scaling factor of 8. As Python doesn't support multiple nodes having the same name, we introduced tagged nodes - with their atomic symbol and an index (number notation). Hence, the first 32 nodes represent "H 1" to "H 32" the next 8 nodes represent "C 33" to "C 40" and so on.

Based on the possible reaction mechanisms proposed by Miller and J. Bada (Miller (1953), Bada (2013)), initial edges resulting in the formation of 2 methane, 2 ammonia and 1 water molecules were added to the network. After each edge addition or deletion, the `fix_valencies` function is called.

2.3 Gibbs free energy calculation

Functions were written to access the thermodynamic coefficients from the parsed CSV files (compound wise - given library, index number and across libraries) and calculate the standard enthalpy change, standard entropy change at any given temperature. Using the standard enthalpy change and standard entropy change, the standard Gibbs free energy difference for a compound was calculated.

Another graph `H` was created with an initial list of compounds. A dictionary with the subgraph of the connected components as the keys and their respective standard Gibbs free energy change as values was created. Whenever a compound is formed, it is compared for isomorphism with the connected components in graph `H`. If the compound is found to be non-isomorphic with the connected components in graph `H`, it is included in the dictionary and the standard Gibbs free energy change of the compound is taken as an input from the user.

2.4 Bond distribution (over iterations)

The maximum number of bonds that can be formed using all the nodes in the network was calculated. This was done by considering the sum of maximum valencies of all the nodes and reducing them by a factor of two (considering all single bonds). Taking the maximum number of bonds that can be formed and the iteration number as parameters, the `get_number_bonds` function, generates a Michaelis-Menten like curve to determine the number of bonds that can be formed at a given iteration.

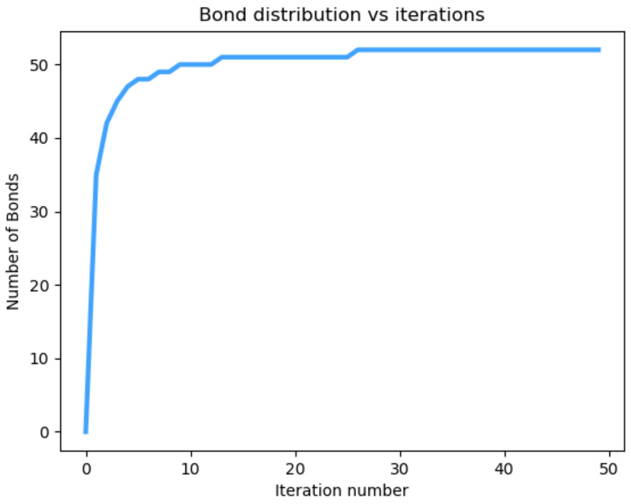


Fig. 3. Bond distribution over Iteration following a Michaelis-Menten like curve (Maximum bond size: 52)

2.5 Random Graph Generation

A random graph generator algorithm was used to generate connected components in our network, given the maximum size of the compound and the number of bonds to be formed. For each iteration, the bond distribution function (`get_number_bonds`) was called and new bonds were introduced in the network whenever the number of edges in the network was found to be lesser than the number returned from `get_number_bonds`. Optimization of bond formation is implemented in the network in the following manner:

(1) A random node is initially picked from all the nodes in the network.
(2) All nodes which have their valency satisfied were rejected and more nodes were sampled.
(3) Once a node with unsatisfied valency is picked, a second node is selected from the network, which also has an unsatisfied valency.
(4) After the two nodes are selected, a random integer, `max_num_edge`, ranging from 1 to the minimum of the unsatisfied valencies of the two nodes is picked.
(5) Between the two nodes picked, `max_num_edge` bonds are added.

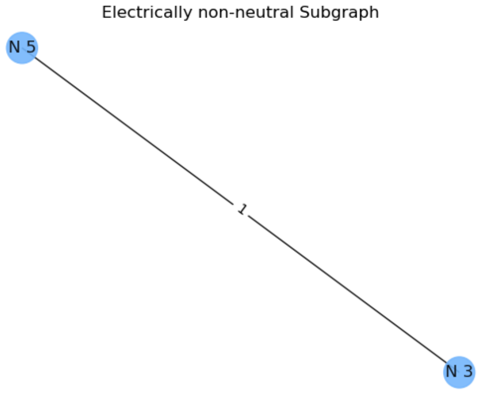
This procedure is repeated until there is only one connected component having size less than `max_size`.

2.6 Ensuring electrical neutrality

After analyzing the results from our random graph generator algorithm, we noticed that the valencies of the atoms in the connected components formed weren't always satisfied. As RMG largely supports only electrically neutral species, the connected components obtained from random graph generator were made electrically neutral.

The electrical neutrality of the compounds was ensured in the following manner:

(1) All connected components of size 2 and above were selected
(2) Atoms in the connected component that did not have a satisfied valency were identified and the valencies of the neighboring atoms were scanned. Atoms that had neighboring atoms with unsatisfied valency, were allowed to form multiple bonds with the neighboring atoms until the valency of the atom or its neighbor was satisfied.
(3) If the valency of the atom is still not satisfied, then free hydrogens in the network were added to the atom with a valency deficiency. In order to take all cases into consideration, hydrogen atoms were added to the network when needed.



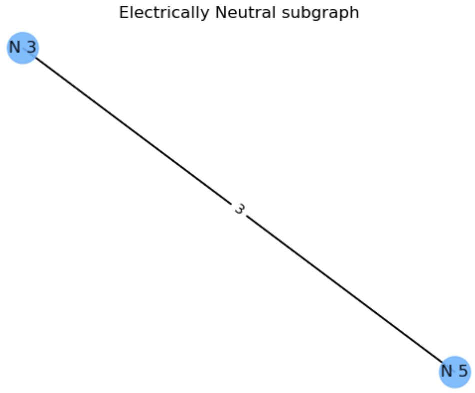


Fig. 4. Working example of ensuring electrical neutrality. 1 represents single bonds, 3 represents triple bonds (Multiple bonds between two atoms were created)

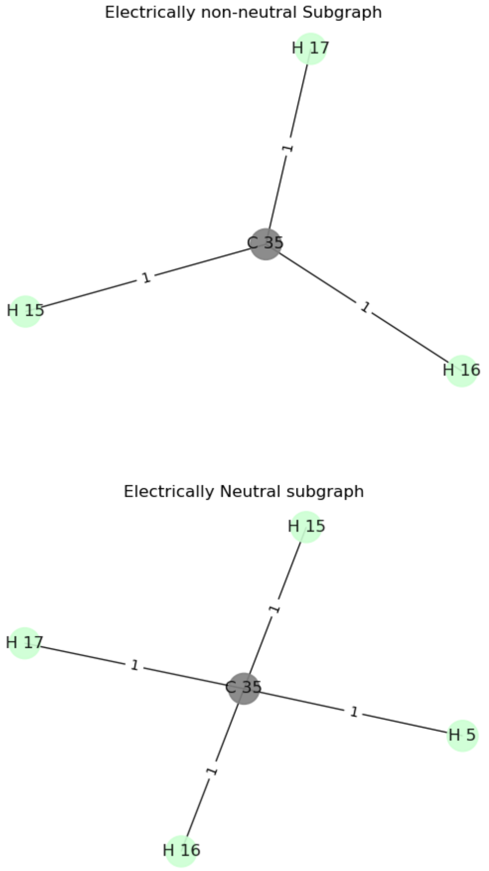

Electrically non-neutral Subgraph
H 17
1
C 35
H 15
1
1
H 16
Electrically Neutral subgraph
H 15
1
H 17
1
C 35
1
H 5
1
H 16

Fig. 5. Working example of ensuring electrical neutrality. Hydrogen atoms are added

2.7 Simulated Annealing

A simulated annealing approach was used to generate the initial reaction intermediates in our network. The sum total of the Gibbs free energy change of compounds in the network was used as the measure of "goodness" of a solution state. As the iteration increases, the probability of accepting a bad solution decreases. This probability is governed by the following equations:

$$p = \exp\Big(\frac{-(\Delta G^{\circ}_{tot,current} - \Delta G^{\circ}_{tot,best})}{T}\Big) \quad (1)$$

$$T = (0.995)^{iteration} T_{intial} \quad (2)$$

Where, ΔG°_{tot} represents the total standard free energy change, T represents the temperature and T_{intial} represents the initial temperature. In order to sample the solution space, the rearrange connected component function was used. A random variable (p), distributed uniformly between $[0, 1]$, was picked and the outcome of the random distribution was used to determine the intensity of connected component rearrangement. Based on the value of p, the extent of reshuffling was determined in the following manner:

- `p < 0.1`: All the connected components are completely reshuffled
- `0.1 < p < 0.6`: Four connected components are selected, two reshufflings are performed
- `0.6 < p < 0.8`: Two connected components are selected, one reshuffling is performed
- `p > 0.8`: No reshuffling is performed

2.8 Rearrange connected components

The `rearrange_connected_components` function is used to increase the sample space of the simulated annealing approach. Two nodes from two different connected components that have the same number of edges to another node are picked. The two connected components are split and the compounds are crossed over. The number of crossovers in the graph are largely determined by p from Simulated Annealing.

2.9 Consolidation of all the compounds obtained so far

A graph consisting of all the unique compounds formed so far was created. All connected components of size 5(8) and above were removed from the graph. As the computational ability available to us was limited, the above mentioned set of codes were run for different initial conditions and the final set of compounds of size 4(7) and less, were consolidated across runs. A total of 11(15) unique compounds were added to a new graph N for further analysis of Glycine(Alanine) formation.

2.10 Reaction simulation

As most of the compounds obtained after the consolidation were Lewis acids and bases, acid-base-like reactions were stimulated using hydrogen atoms. A constraint for the maximum size of compounds that can form was imposed. The in-silico reaction mechanism is simulated as follows:

(1) A pair of connected components were selected from the graph sequentially.
(2) All the hydrogen atoms in both the connected components were identified. And if any of the components didn't have hydrogen atoms, it was prevented from being iterated again.
(3) If the connected component had more than one hydrogen atom, the subsequent reactions was carried out sequentially for all the hydrogen atoms. The hydrogen atoms picked from the two connected components were named `hydro1` and `hydro2`.
(4) The neighbor of the hydrogen atoms `hydro1` and `hydro2`, from the connected components, was identified and named `neighbor1` and `neighbor2`.
(5) The bonds between the (`neighbor1`, `hydro1`) and (`neighbor2`, `hydro2`) were broken and bonds between (`neighbor1`, `neighbor2`) were created.
(6) Atoms that have neighbors that are not exclusively hydrogen atoms were identified. If the number of atoms having a non hydrogen neighbor is greater than one, then the entire list is looped over sequentially to obtain all possible combinations. The atom picked was named `non_h_neighbor`.
(7) Edges between the `hydro1` atom `non_h_neighbor` are added in the graph and an edge between `non_h_neighbor` and one of its initial neighbors (`init_neigh`) is removed from the graph to satisfy the valency conditions. A bond between `init_neigh` and `hydro2` is added to the network.
(8) All new compounds are added to a dictionary as the key and the standard Gibbs free energy value is taken as the value.

Fig. 6A. Reactant 1: Formaldehyde

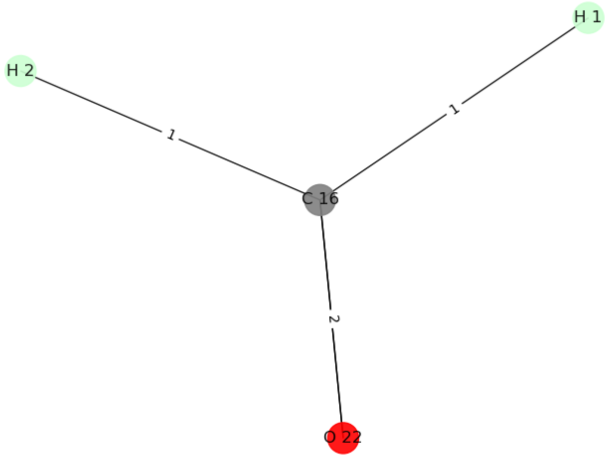


Fig. 6B. Reactant 2: Hydrogen Cyanide

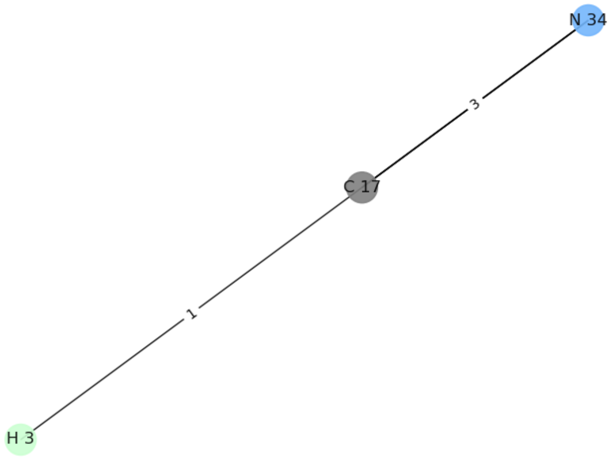


Fig. 6C. Hydrogen selected: H 1(`hydro1`) and H 3(`hydro2`. Bond formed between C16(`neigh1`) and C17(`neigh2`): Glyoxylonitrile

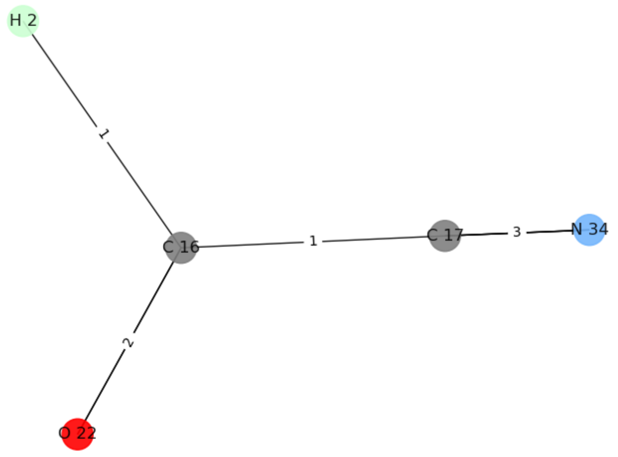


Fig. 6D. Hydrogens added back to the compound with C 17(`non_h_neighbor`) and N 34 (`inti_neigh`): Iminoacetaldehyde

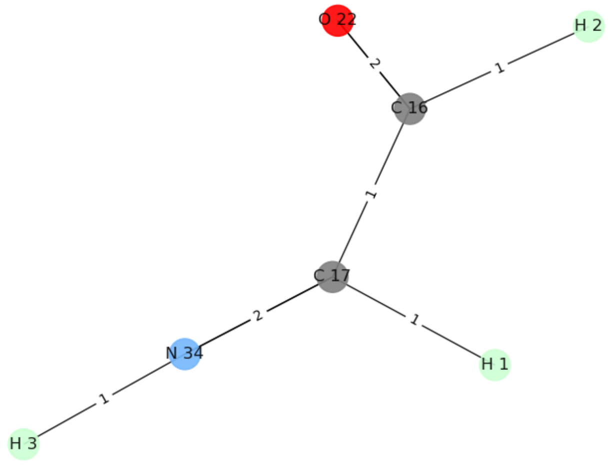


Fig. 6. Working example of the reaction simulation mediated by Hydrogen atoms

Once all possible combinations of reactions are performed, the resulting dictionary is passed as the output. As this function was largely dependent on the order of reaction between the connected components, upto twenty possible combinations of the order were tried and the results were consolidated.

In order to check if the compounds returned from this function existed, the compounds were searched using their structure on ChemSpider, database managed by the Royal Society of Chemistry (Pence and Williams (2010)). In cases where the compound returned did not exist, a penalty of 100 KJ was awarded as the Gibbs free energy to the compound. All entries in the dictionary were screened and compounds that have a standard Gibbs free energy change greater than 10KJ were removed from the dictionary for further analysis.

2.11 Strecker amino acid synthesis

Since hydrogen cyanide, formaldehyde and aminoacetonitrile were present among the list of compounds obtained so far, we decided to perform the Strecker's amino acids synthesis reaction for all aldehydes and ketones. The Strecker's amino acid synthesis reaction is also a part of the Urey-Miller experiment, which finally led to the synthesis of amino acids. This is implemented in our model by scanning the list of compounds obtained for aldehydes and ketones. Compounds were also scanned for presence of aminonitriles. The ammonia and hydrogen cyanide used in this reaction were added as additional nodes in the network. This was done to ensure that all aldehydes and ketones in the initial list would take part in the Strecker's reaction. These additional nodes introduced in this step were numbered from 100.The strecker amino acids synthesis is as follows:

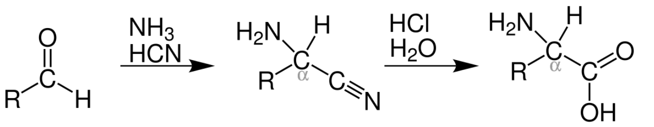


Fig. 7. Strecker amino acid synthesis mechanism (GerrietB (2016))

Fig. 8 A. Initial reactant: Formaldehyde

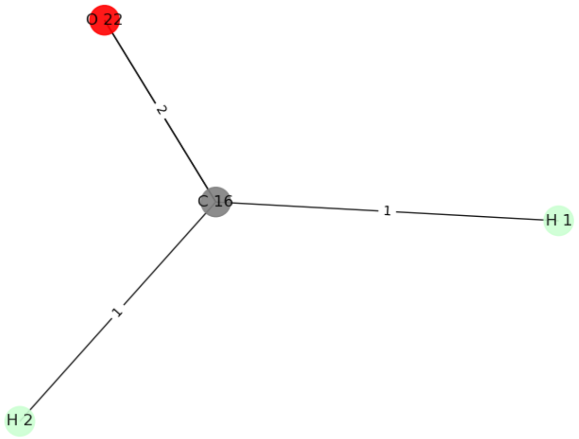
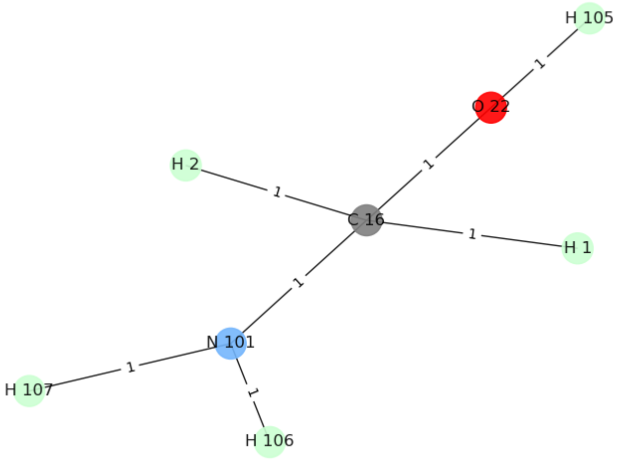


Fig. 8 B. Addition of ammonia: Aminomethanol



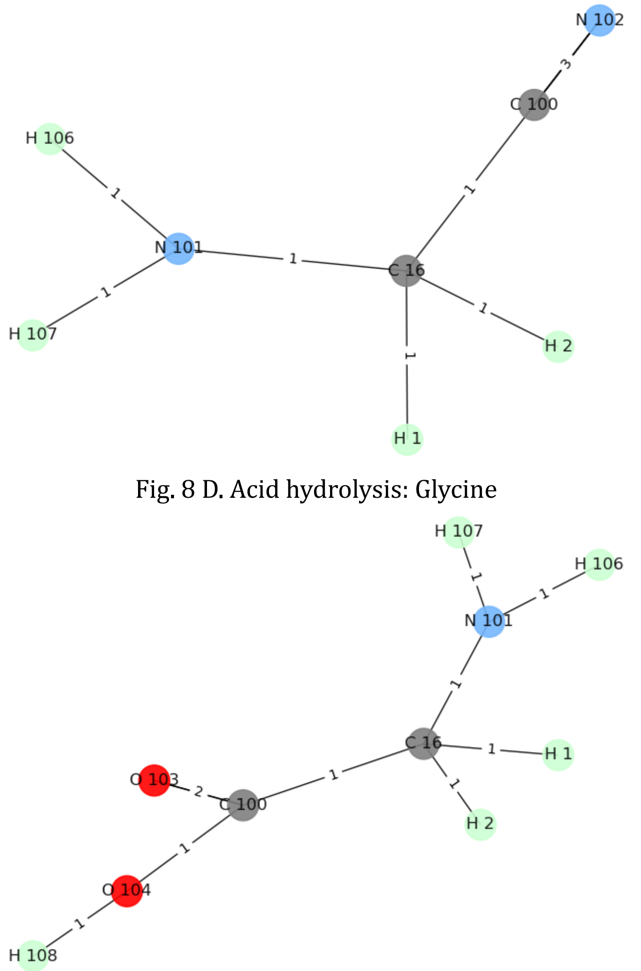


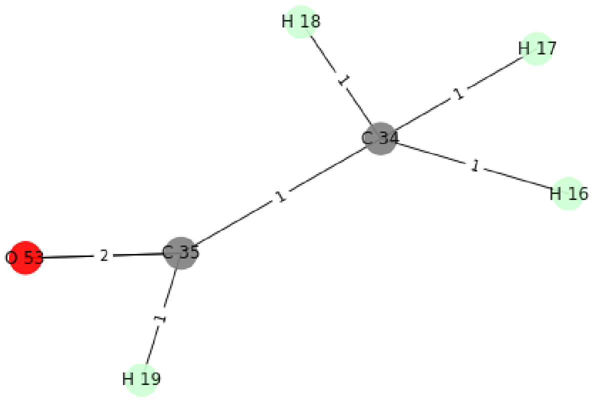
Fig. 8 C. Addition of hydrogen cyanide: Aminoacetonitrile

Fig. 8 D. Acid hydrolysis: Glycine

Fig. 8. Working example of Strecker reaction - Glycine formation

Acid hydrolysis on these compounds was performed in-silico by adding an amino and carboxylic acid group or by converting the cyanide group into a carboxylic acid group.

Fig. 9 A. Initial reactant: Acetaldehyde



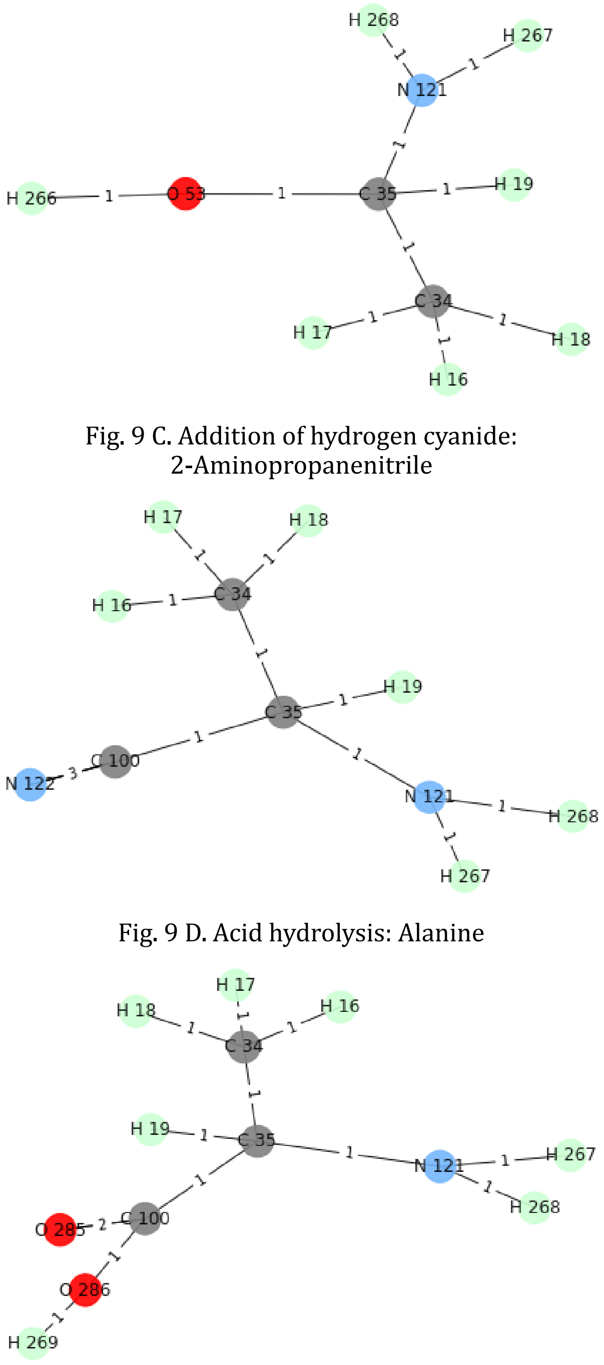


Fig. 9 B. Addition of ammonia: Ethylamine

Fig. 9 C. Addition of hydrogen cyanide: 2-Aminopropanenitrile

Fig. 9 D. Acid hydrolysis: Alanine

Fig. 9. Working example of Strecker reaction - Alanine formation

3. RESULTS AND DISCUSSIONS

Random graph generator was used to generate compounds from the initial set of atoms, with a size limitation of 2 and 3. The resultant compounds were used as inputs for the simulated annealing function. The simulated annealing function was used to scan the reaction space and return a set of compounds that have low Gibbs free energy. The output from simulated annealing was screened for compounds that have a size lesser than 4 for Glycine and 7 for Alanine. These compounds were used as inputs to the acid-base reaction function and the enhanced output was then used as an input for the Strecker amino acid synthesis reaction.Throughout our model, the compounds generated were analyzed, filtered and eliminated based on their standard Gibbs free energy.

From analysis of the final set of products formed, we found that formaldehyde, aminoacetonitrile and hydrogen cyanide were the key intermediates for the formation of glycine. This can be observed from the numbering of the atoms in Fig. 8. The Gibbs free energy distribution over reaction progress is as follows:

Table 1. Gibbs free energy distribution of intermediates involved in Glycine synthesis

Compound	Gibbs free energy change
Formaldehyde	-269.5222798
Methylamine	-193.3923271
Aminoacetonitrile	1.575047801
Glycine	-664.5829247

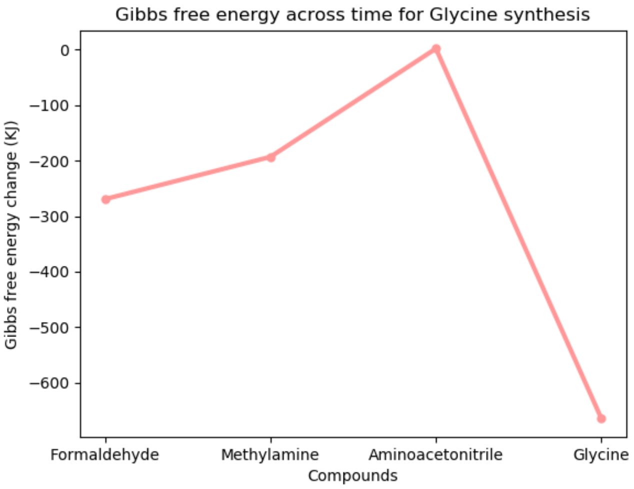


Fig. 10. Gibbs free energy across time for Glycine synthesis

Table 2. Gibbs free energy distribution of intermediates involved in Alanine synthesis

Compound	Gibbs free energy change
Acetaldehyde	-368.2453749
Ethylamine	-263.9192457
2-Aminopropanenitrile	-6.190248565965583
Alanine	-623.24864

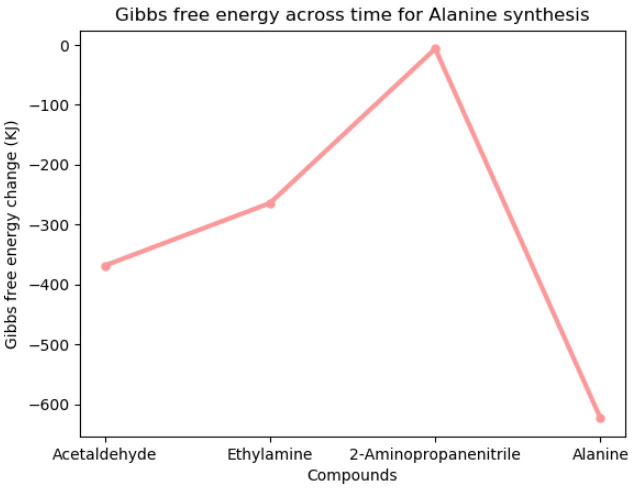


Fig. 11. Gibbs free energy across time for Alanine synthesis

From Fig. 10 and Fig. 11, we find that our model was able to identify the characteristic peak and drop in the Gibbs free energy change that is generally associated with all chemical reactions.

The formation of glycine ($\Delta G^\circ : -664.5829247KJ$) in this study with the standard Gibbs free energy change as the key parameter strengthened our hypothesis that highly negative free energy of amino acids in prebiotic Earth conditions could have caused their accumulation and eventual build-up. The results obtained from this study were ratified by several previous studies.

- All codes written as a part of the project are available on GitHub: https://github.com/sowmyamanojna/Computational-Systems-Biology-project.
- The Gibbs free energy difference of all compounds formed in our model is available here: https://bit.ly/2BSSLqr.
- The images of all compounds formed in our project is available here: https://bit.ly/2Zwo8Pu.

Table 3. Gibbs free energy distribution of other key compounds formed in our model

Compound	Gibbs free energy change
Formamide	-374.5075468
Nitrosamine	-107.61248
Hydroxylamine	-215.860156
Glycolonitrile	-261.5047571
Iminoacetaldehyde	-7.380497132
Nitroxyl / Azazone	-55.40205694
Hydrogen peroxide	-309.1934413

4. CHALLENGES FACED

The unavailability of an index for charged and lone pair compounds, did not permit us to simulate detailed reactions involving electron exchanges or incomplete compounds. This limitation forced us to ensure electrical neutrality of all the compounds formed. A simulated annealing approach with narrow acceptance rate for configurations with high Gibbs free energy was imposed in the consequent steps.

The other challenge that we faced in the project was the acid-base reaction simulation. A total of 11 unique compounds were sent as input to the acid-base reaction. The order in which these compounds were sent was randomized. As the function depends on the order of compounds, a total of 110 trails should have been done. However, as each trail resulted in approximately 10 new graphs, we weren't computationally equipped to handle 110 runs. The attempt to run the same, resulted in high memory utilization which led to our laptop computers hanging. Hence, we restricted our study to just 20 runs. This project could be made better by addressing the factors listed above.

5. ACKNOWLEDGMENTS

We would like to thank our Professor Karthik Raman for giving us this opportunity to work on many interesting and open-ended problems as a part of our Computational Systems Biology course. We owe our knowledge gained to our professor who has made the subject very interesting.

REFERENCES

Bada, J.L. (2013). New insights into prebiotic chemistry from stanley miller's spark discharge experiments. *Chem. Soc. Rev.*, 42, 2186–2196. doi:10.1039/C3CS35433D.

Gao, C.W., Allen, J.W., Green, W.H., and West, R.H. (2016). Reaction mechanism generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications*, 203, 212 – 225. doi:https://doi.org/10.1016/j.cpc.2016.02.013.

GerrietB (2016). *Strecker amino acid synthesis*. URL `https://en.wikipedia.org/wiki/Strecker_amino_acid_synthesis`.

Miller, S.L. (1953). A production of amino acids under possible primitive earth conditions. *Science*, 117(3046), 528–529. doi:10.1126/science.117.3046.528.

Pence, H.E. and Williams, A. (2010). Chemspider: An online chemical information resource. *Journal of Chemical Education*, 87(11), 1123–1124. doi:10.1021/ed100697w. URL `https://doi.org/10.1021/ed100697w`.

Pietrucci, F. and Saitta, A.M. (2015). Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios. *Proceedings of the National Academy of Sciences*, 112(49), 15030–15035. doi:10.1073/pnas.1512486112. URL `https://www.pnas.org/content/112/49/15030`.

Saitta, A.M. and Saija, F. (2014). Miller experiments in atomistic computer simulations. *Proceedings of the National Academy of Sciences*, 111(38), 13768–13773. doi:10.1073/pnas.1402894111.