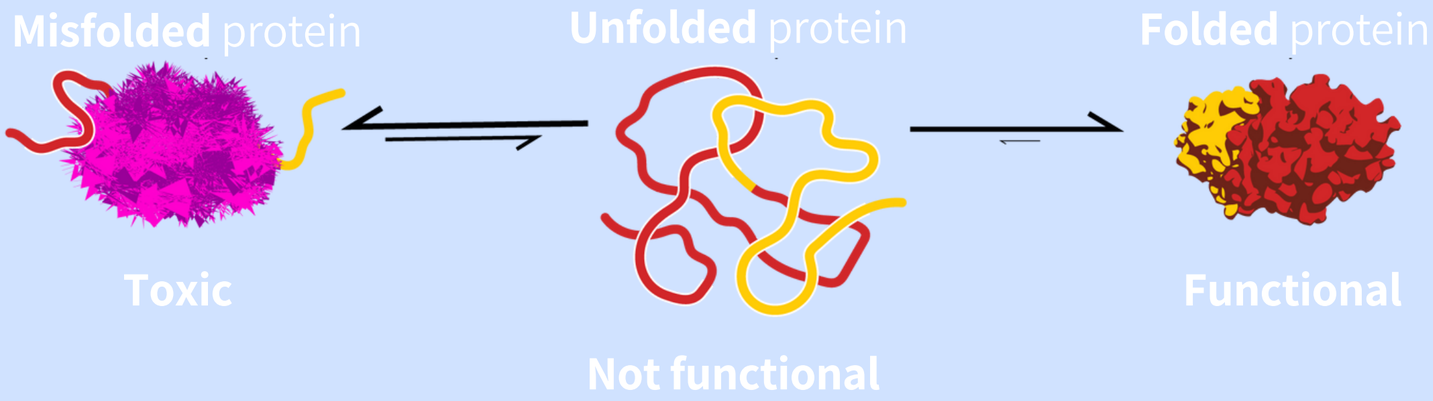
Protein Folding

Uniform Random Sampling using the HP-Model
and Randomly Generated Amino Acids

Protein Folding 101

Why do proteins fold?

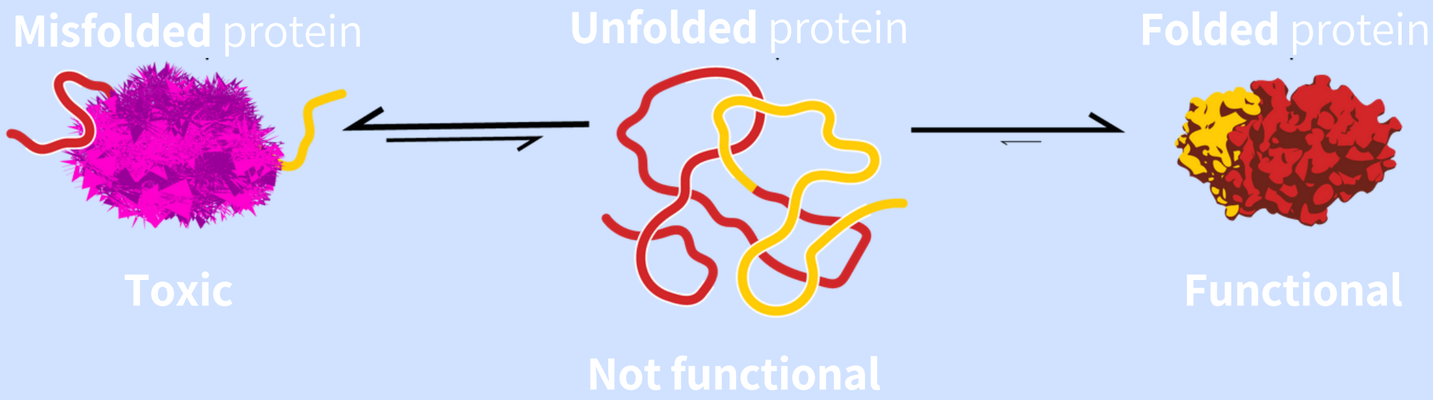
Folding is the process of taking on functional structure.



https://sites.krieger.jhu.edu

Why we fold proteins?

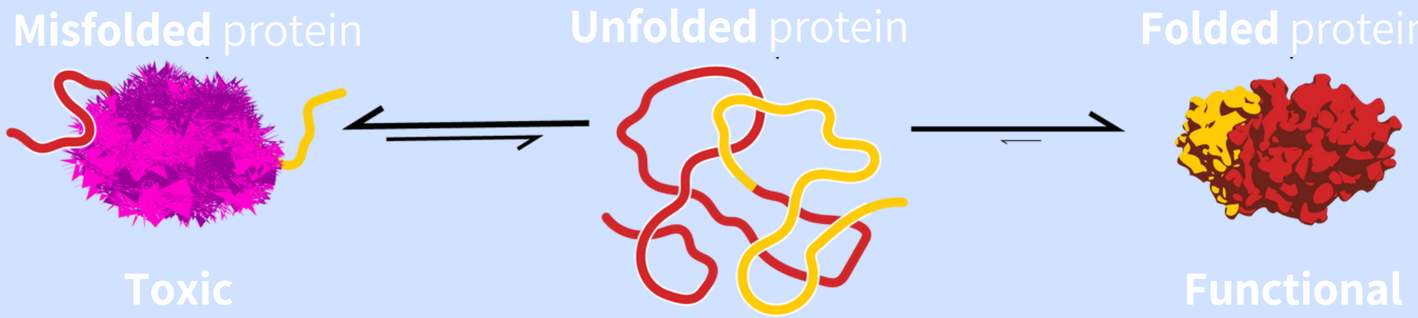
To predict the functional structure of proteins.



https://sites.krieger.jhu.edu

Why we fold proteins?

To predict the functional structure of proteins.



Diseases

Alzheimer's

Parkinson's

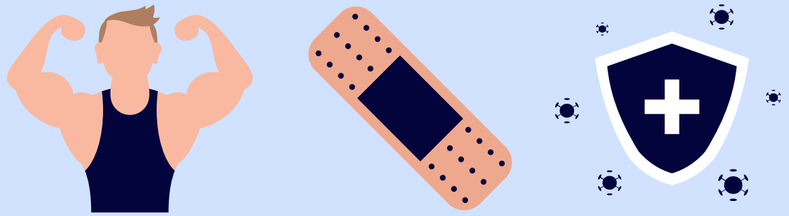
Diabetes

Functionalities

Muscle contraction

Skin healing

Immune system

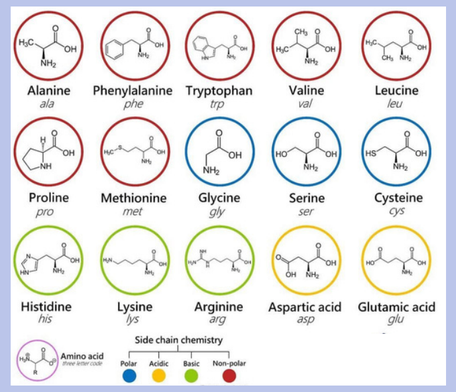


https://sites.krieger.jhu.edu

Protein and aminos

A Protein is a chain of amino acids.

20 different amino acids



Average protein 200-300 acids

Insulin - 51 amino acids

Chain A - 21 amino acids

GIVEQCCTSICSLYQLENYCN

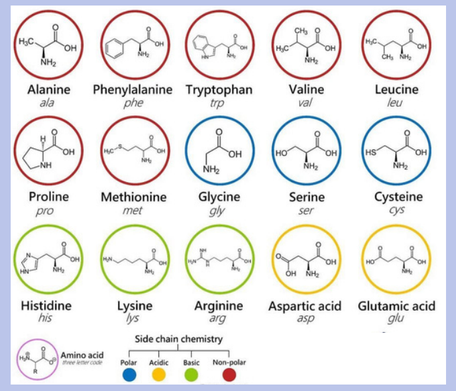
Chain B - 30 amino acids

FVNQHLCGSHLVEALYLVCGERGFFYTPKT

Protein and aminos

A Protein is a chain of amino acids.

20 different amino acids



Average protein
200-300 acids

Insulin - 51 amino acids

Chain A - 21 amino acids
GIVEQCCTSICSLYQLENYCN

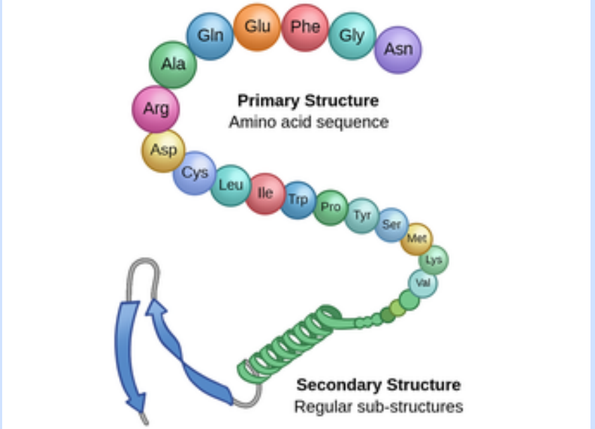
Chain B - 30 amino acids
FVNQHLCGSHLVEALYLVCGERGFFYTPKT

Many Proteins > 2.000 amino acids
CGSHLVEALYLVCGERGFFYTPKTFVNQHLCGSHLVEALYLVCGERGFFYTPKTCGSHLVEALYLVCGERGFFYTPKTCGSHLVEALYLVCGERGFFYTPKTFVNQHL

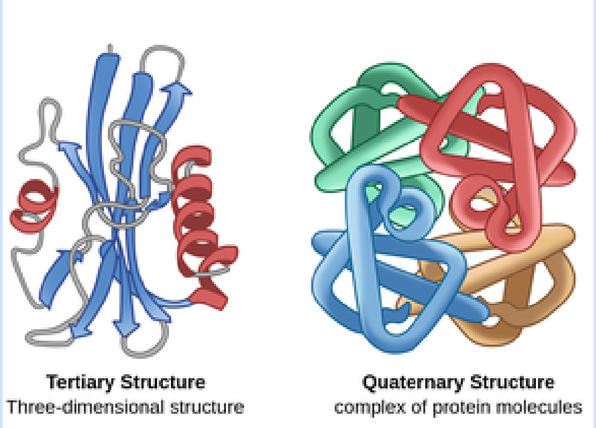
Protein Structure

Proteins can be described and folded in terms of structural levels.

1D and 2D



Multidimensional

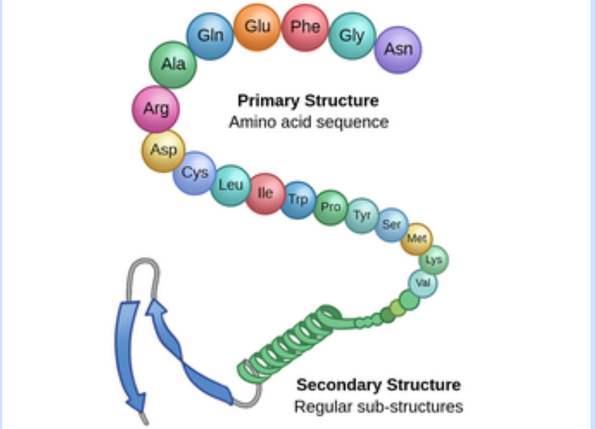


https://theory.labster.com/protein-structure/

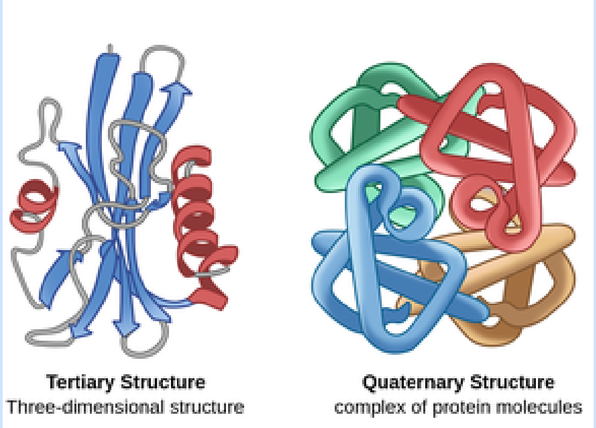
Protein Structure

Proteins can be described and folded in terms of structural levels.

1D and 2D



Multidimensional



https://theory.labster.com/protein-structure/

Approaches

Protein folding papers

Biochemistry **1985**, *24*, 1501–1509 1501

Koide, T., Odani, S., & Ono, T. (1982) *FEBS Lett. 141*, 222–224.
Laemmli, U. K. (1970) *Nature (London) 227*, 680–685.
Leung, L. L. K., Harpel, P. C., Nachman, R. L., & Rabellino, E. M. (1983) *Blood 62*, 1016–1021.
Leung, L. L. K., Nachman, R. L., & Harpel, P. C. (1984) *J. Clin. Invest. 73*, 5–12.
Levine, R. L., & Federici, M. M. (1982) *Biochemistry 21*, 2600–2606.
Lijnen, H. R., Hoylaerts, M., & Collen, D. (1980) *J. Biol. Chem. 255*, 10214–10222.
Lijnen, H. R., Hoylaerts, M., & Collen, D. (1983a) *J. Biol. Chem. 258*, 3803–3808.
Lijnen, H. R., Rylatt, D. B., & Collen, D. (1983b) *Biochim. Biophys. Acta 742*, 109–115.
Miles, E. W. (1977) *Methods Enzymol. 47*, 431–442.
Morgan, W. T. (1978) *Biochim. Biophys. Acta 535*, 319–333.
Morgan, W. T. (1981) *Biochemistry 20*, 1054–1061.
Morgan, W. T., & Muller-Eberhard, U. (1972) *J. Biol. Chem. 247*, 7181–7187.
Morgan, W. T., & Muller-Eberhard, U. (1976) *Arch. Biochem. Biophys. 176*, 431–441.
Morgan, W. T., & Smith, A. (1984) *Protides Biol. Fluids 31*, 23–26.
Morris, J. P., Blatt, S., Powell, J. R., Strickland, D. K., & Castellino, F. S. (1981) *Biochemistry 20*, 4811–4816.
Pajot, P. (1976) *Eur. J. Biochem. 63*, 263–269.
Rylatt, D. B., Sia, D. Y., Mundy, J. P., & Parish, C. R. (1981) *Eur. J. Biochem. 119*, 641–646.
Segrest, J. P., & Jackson, R. L. (1972) *Methods Enzymol. 28*, 54–63.
Skoza, L., & Mohos, S. (1976) *Biochem. J. 159*, 457–462.
Soininen, R., & Ellfolk, N. (1973) *Acta Chem. Scand. 27*, 35–46.
Spencer, R. L., & Wold, F. (1969) *Anal. Biochem. 32*, 185–190.

Theory for the Folding and Stability of Globular Proteins†

Ken A. Dill

Departments of Pharmaceutical Chemistry and Pharmacy, University of California, San Francisco, California 94143

Received May 3, 1984

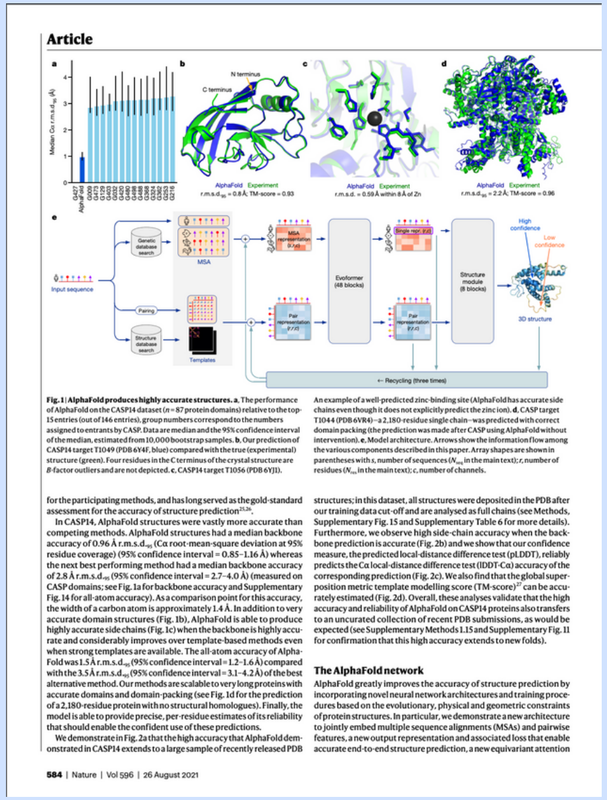
ABSTRACT: Using lattice statistical mechanics, we develop theory to account for the folding of a heteropolymer molecule such as a protein to the globular and soluble state. Folding is assumed to be driven by the association of solvophobic monomers to avoid solvent and opposed by the chain configurational entropy. Theory predicts a phase transition as a function of temperature or solvent character. Molecules that are too short or too long or that have too few solvophobic residues are predicted not to fold. Globular molecules should have a largely solvophobic core, but there is an entropic tendency for some residues to be "out of place", particularly in small molecules. For long chains, molecules comprised of globular domains are predicted to be thermodynamically more stable than spherical molecules. The number of accessible conformations in the globular state is calculated to be an exceedingly small fraction of the number available to the random coil. Previous estimates of this number, which have motivated kinetic theories of folding, err by many tens of orders of magnitude.

Few heteropolymers are both globular and soluble. Proteins are the principal exception. Globularity and solubility cannot be achieved with any random sequence of monomers; certain principles of structure and function must be obeyed (Kauzmann, 1959; Fisher, 1964; Flory, 1969; Tanford, 1968; Brandts, 1968; Edsall, 1968; Edsall & McKenzie, 1983; Lifschitz, 1968; Volkenstein, 1970; Richards, 1977; Klapper, 1971, 1973). A molecule with too many solvophilic residues will prefer solvation to globularity. Molecules with too many solvophobic residues will aggregate, as occurs with oil in water. In addition, globularity requires that the chain can pack well in the condensed state. Typical globular proteins have densities approaching those of crystalline hydrocarbons and amino acids and compressibilities a factor of 20 smaller than liquid hydrocarbons, nearly equal to those of some metals, and they contain less than 3 vol % of internal water or cavities (Richards, 1974, 1977; Klapper, 1971, 1973; Chothia, 1975; Kuntz & Kauzmann, 1974; Connolly, 1981; Gavish et al., 1983; Sturtevant, 1977; Nemethy et al., 1981). The importance of packing also follows from the fact that evolution conserves residue size and shape (Schultz & Schirmer, 1979). But high density comes at a high price; enormous configurational entropy must be overcome to achieve it. The fact that most enzymes are condensed suggests that catalytic function may require the high density state. It is a reasonable hypothesis that this is due to the requirement that the atoms of the active site have relatively invariant spatial positions during a significant fraction of the time required to attract and hold the substrate for the catalytic act. In this regard, the primary molecular mechanism for maintaining relative spatial invariance, i.e., for reducing the amplitude of out-of-phase internal thermal motion, is that of steric constraint, which is achieved, as in the solid state, through high density packing. In the solid state, incident thermal energy may be distributed in modes of motion whose spatial wavelengths are larger than the size of the active site, and thus, this thermal energy may be exchanged with the protein through relatively nondisruptive rigid body motions of the active site.

For those sequences which satisfy the requirements above, it follows that relatively few spatial conformations are available in the globular state, their number being limited by (i) the

† Acknowledgment is made to the donors of the Petroleum Research Fund, administered by the American Chemical Society, for partial support of this work and to the National Institute of General Medical Sciences.

0006-2960/85/0424-1501$01.50/0



Article

Fig. 1 | AlphaFold produces highly accurate structures. a, The performance of AlphaFold on the CASP14 dataset ($n = 87$ protein domains) relative to the top-15 entries (out of 146 entries). group numbers correspond to the numbers assigned to entrants by CASP. Data are median and the 95% confidence interval of the median, estimated from 10,000 bootstrap samples. **b**, Our prediction of CASP14 target T1049 (PDB 6Y4F, blue) compared with the true (experimental) structure (green). Four residues in the C terminus of the crystal structure are *B*-factor outliers and are not depicted. **c**, CASP14 target T1056 (PDB 6Y1J). An example of a well-predicted zinc-binding site (AlphaFold has accurate side chains even though it does not explicitly predict the zinc ion). **d**, CASP target T1044 (PDB 6VR4)—a 2,180-residue single chain—was predicted with correct domain packing (the prediction was made after CASP using AlphaFold without intervention). **e**, Model architecture. Arrows show the information flow among the various components described in this paper. Array shapes are shown in parentheses with *s*, number of sequences (N_{seq} in the main text); *r*, number of residues (N_{res} in the main text); *c*, number of channels.

for the participating methods, and has long served as the gold-standard assessment for the accuracy of structure prediction[21,26].

In CASP14, AlphaFold structures were vastly more accurate than competing methods. AlphaFold structures had a median backbone accuracy of 0.96 Å r.m.s.d_{95} (Cα root mean-square deviation at 95% residue coverage) (95% confidence interval = 0.85–1.16 Å) whereas the next best performing method had a median backbone accuracy of 2.8 Å r.m.s.d_{95} (95% confidence interval = 2.7–4.0 Å) (measured on CASP domains; see Fig. 1a for backbone accuracy and Supplementary Fig. 14 for all-atom accuracy). As a comparison point for this accuracy, the width of a carbon atom is approximately 1.4 Å. In addition to very accurate domain structures (Fig. 1b), AlphaFold is able to produce highly accurate side chains (Fig. 1c) when the backbone is highly accurate and considerably improves over template-based methods even when strong templates are available. The all-atom accuracy of AlphaFold was 1.5 Å r.m.s.d_{95} (95% confidence interval = 1.2–1.6 Å) compared with the 3.5 Å r.m.s.d_{95} (95% confidence interval = 3.1–4.2 Å) of the best alternative method. Our methods are scalable to very long proteins with accurate domains and domain-packing (see Fig. 1d for the prediction of a 2,180-residue protein with no structural homologues). Finally, the model is able to provide precise, per-residue estimates of its reliability that should enable the confident use of these predictions.

We demonstrate in Fig. 2a that the high accuracy that AlphaFold demonstrated in CASP14 extends to a large sample of recently released PDB structures; in this dataset, all structures were deposited in the PDB after our training data cut-off and are analysed as full chains (see Methods, Supplementary Fig. 15 and Supplementary Table 6 for more details). Furthermore, we observe high side-chain accuracy when the backbone prediction is accurate (Fig. 2b) and we show that our confidence measure, the predicted local-distance difference test (pLDDT), reliably predicts the Cα local-distance difference test (lDDT-Cα) accuracy of the corresponding prediction (Fig. 2c). We also find that the global superposition metric template modelling score (TM-score)[27] can be accurately estimated (Fig. 2d). Overall, these analyses validate that the high accuracy and reliability of AlphaFold on CASP14 proteins also transfers to an uncurated collection of recent PDB submissions, as would be expected (see Supplementary Methods 1.15 and Supplementary Fig. 11 for confirmation that this high accuracy extends to new folds).

The AlphaFold network

AlphaFold greatly improves the accuracy of structure prediction by incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures. In particular, we demonstrate a new architecture to jointly embed multiple sequence alignments (MSAs) and pairwise features, a new output representation and associated loss that enable accurate end-to-end structure prediction, a new equivariant attention

HP-model, 1985

AlphaFold, 2023

https://doi.org/10.1038/s41586-021-03819-2

https://doi.org/10.1021/bi00327a032

HP-model

Used to predict the most stable conformation of a protein.

Primary Structure (1D)

Description: Primary structure refers to the linear sequence of amino acids that make up the protein.

Importance: Sequence determines the protein’s type and function. Even a small change in the sequence can significantly affect the protein’s properties.

Input

Secondary Structure (2D)

Description: Folding amino acid chain into regular, repeating structures stabilized by hydrogen bonds.

Importance: Structures provide a level of organization and stability within the protein.

Output

Optimization Algorithms

Tertiary Structure (3D)

Description: Folding and arrangement of the secondary structural elements, stabilized by various types of bonds and interactions.

Importance: Vital for the protein's function, as it positions the protein’s active sites and other functional domains in specific orientations that are necessary for activity.

Quaternary Structure (4D)

Description: Arrangement of multiple subunits into a functional protein complex. Not all proteins have a quaternary structure.

Importance: Mediates the interaction of proteins with other molecules, including other proteins

AlphaFold

AlphaFold is used to predict the 3D of proteins.

Primary Structure (1D)

Description: Primary structure refers to the linear sequence of amino acids that make up the protein.

Importance: Sequence determines the protein’s type and function. Even a small change in the sequence can significantly affect the protein’s properties.

Input

Neural network

Secondary Structure (2D)

Description: Folding amino acid chain into regular, repeating structures stabilized by hydrogen bonds.

Importance: Structures provide a level of organization and stability within the protein.

Substructure

Neural network

Tertiary Structure (3D)

Description: Folding and arrangement of the secondary structural elements, stabilized by various types of bonds and interactions.

Importance: Vital for the protein's function, as it positions the protein’s in specific orientations that are necessary for activity.

Output

Quaternary Structure (4D)

Description: Arrangement of multiple subunits into a functional protein complex. Not all proteins have a quaternary structure.

Importance: Mediates the interaction of proteins with other molecules, including other proteins

AlphaFold VS HP-model

Advantages and disadvantages

HP-model

Pros

Simplicity
Computational effiency
Educational Tool

Cons

Over Simplification
Limited Predictive Power

Theory for the Folding and Stability of Globular Proteins*

Ken A. Dill

Departments of Pharmaceutical Chemistry and Pharmacy, University of California, San Francisco, California 94143

Received May 2, 1984

AlphaFold

Pros

High Accuracy
Complexity Handling

Cons

Computational expensive
Dependent on training data
Black box

Article

https://doi.org/10.1038/s41586-021-03819-2
https://doi.org/10.1021/bi00327a032

AlphaFold VS HP-model

Advantages and disadvantages

HP-model

Pros

Simplicity
Computational effiency
Educational Tool

Cons

Over Simplification
Limited Predictive Power

Theory for the Folding and Stability of Globular Proteins*

Ken A. Dill

Departments of Pharmaceutical Chemistry and Pharmacy, University of California, San Francisco, California 94143

Received May 2, 1984

AlphaFold

Pros

High Accuracy
Complexity Handling

Cons

Computational expensive
Dependent on training data
Black box

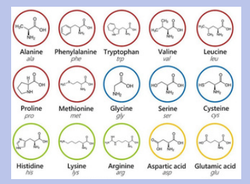
Article

https://doi.org/10.1038/s41586-021-03819-2
https://doi.org/10.1021/bi00327a032

HP-model

A binary perspective

20 to 2 acids

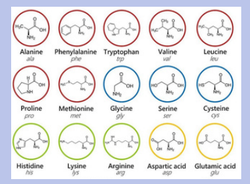


H P

HP-model

A binary perspective

20 to 2 acids



H P

Insulin - 51 amino acids

Chain A - 21 amino acids

GIVEQCCTSICSLYQLENYCN

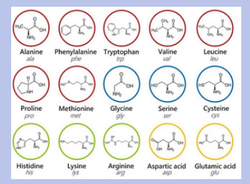
Chain B - 30 amino acids

FVNQHLCGSHLVEALYLVCGERGFFYTPKT

HP-model

A binary perspective

20 to 2 acids



H P

Insulin - 51 amino acids

Chain A - 21 amino acids
GIVEQCCTSICSLYQLENYCN

Chain B - 30 amino acids
FVNQHLCGSHLVEALYLVCGERGFFYTPKT

Insulin - 51 amino acids

Chain A - 21 amino acids
HPPPHHHPPPHHHPPPHP

Chain B - 30 amino acids
HPHHPPPHPPPPHHPPPHHHPPPHHHPP

Simplification

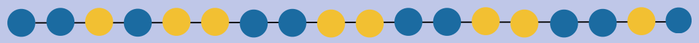
HP-model

From 1D to 2D

1D representation

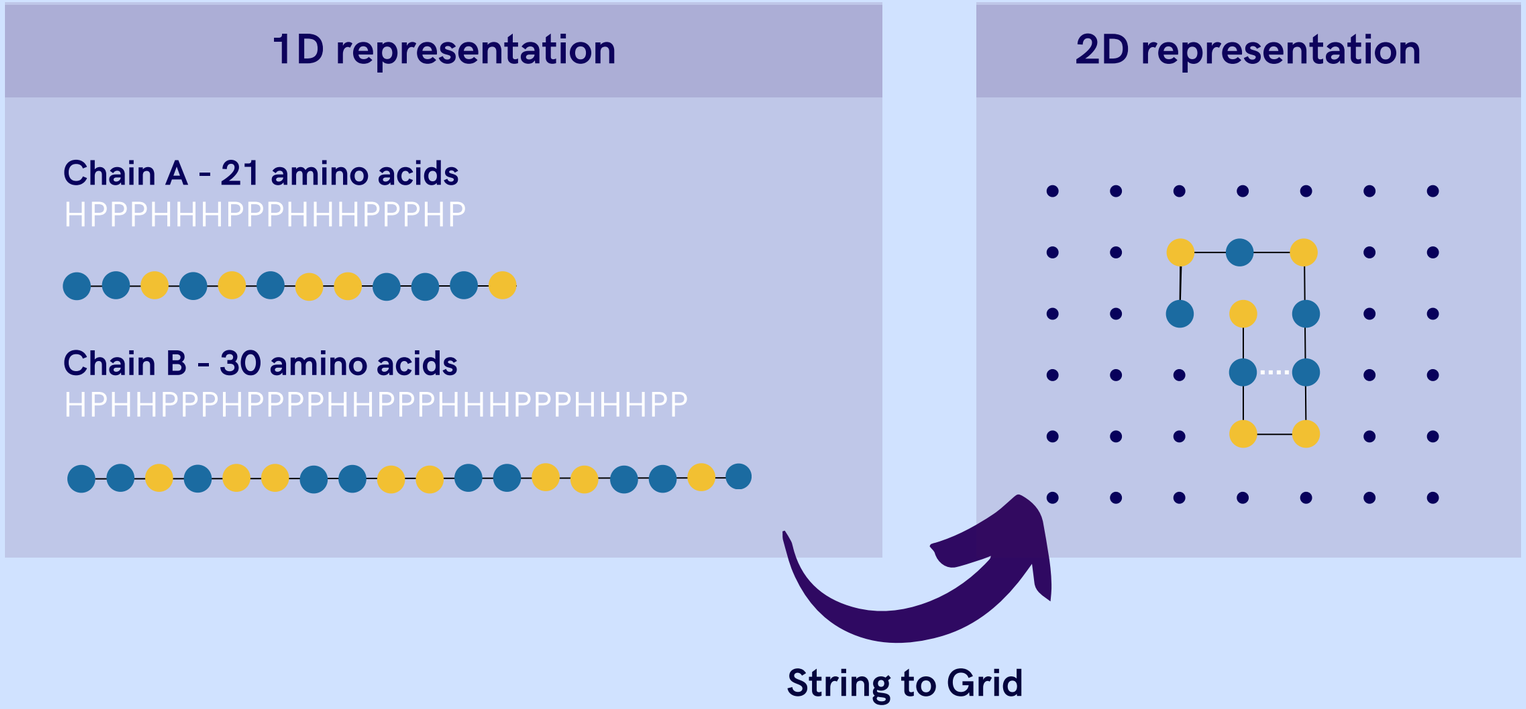
Chain A - 21 amino acids
HPPPHHHPPPHHHPPPHP

Chain B - 30 amino acids
HPHHPPPHPPPPHHPPPHHHPPPHHHPP



HP-model

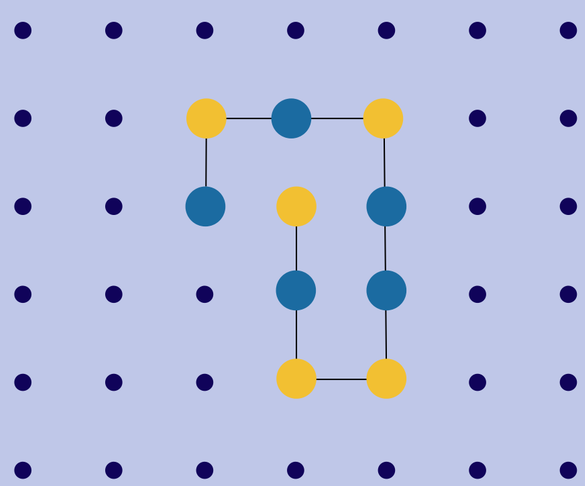
From 1D to 2D



HP-model

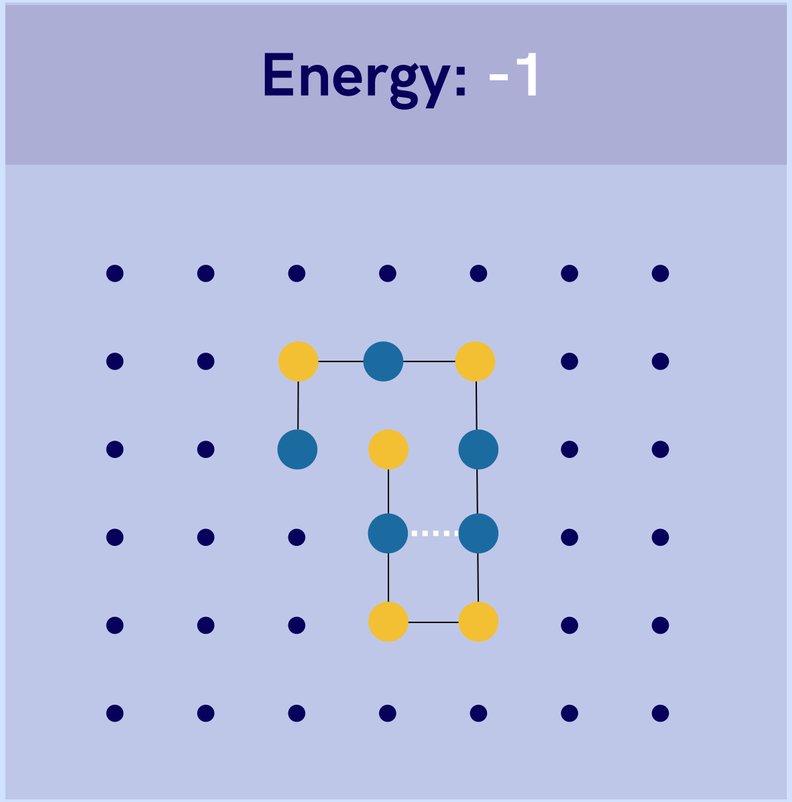
Calculating the stability with H-bonds

Energy:



HP-model

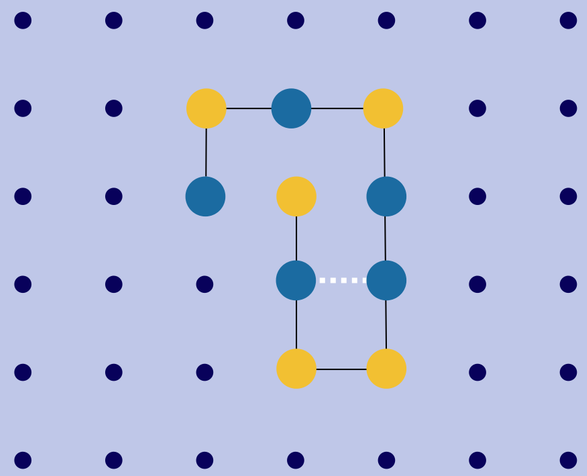
Calculating the stability with H-bonds



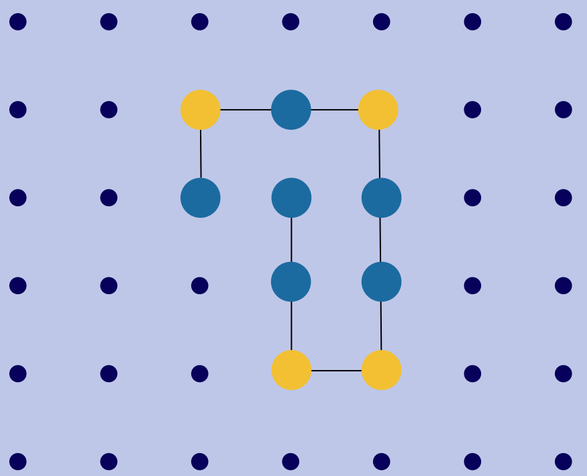
HP-model

Calculating the stability with H-bonds

Energy: -1



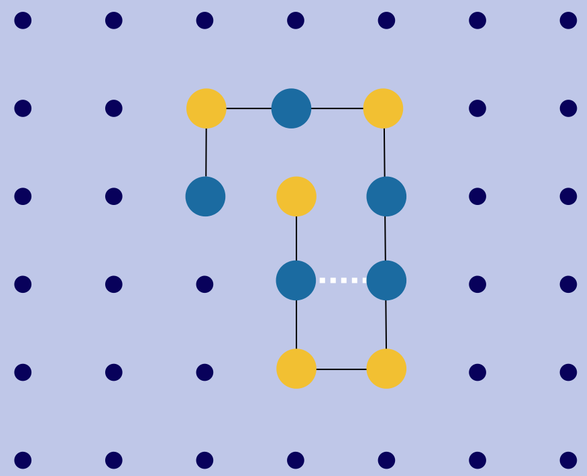
Energy:



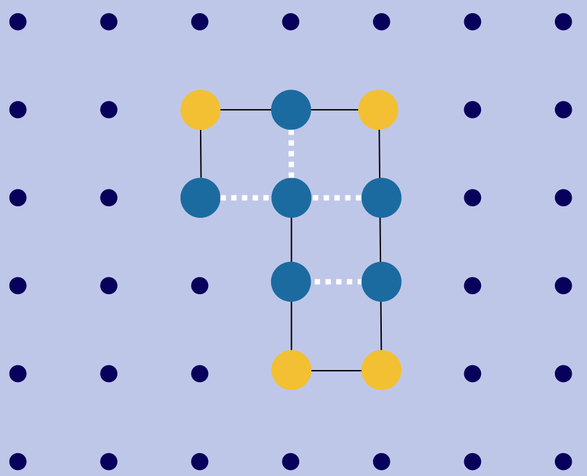
HP-model

Calculating the stability with H-bonds

Energy: -1



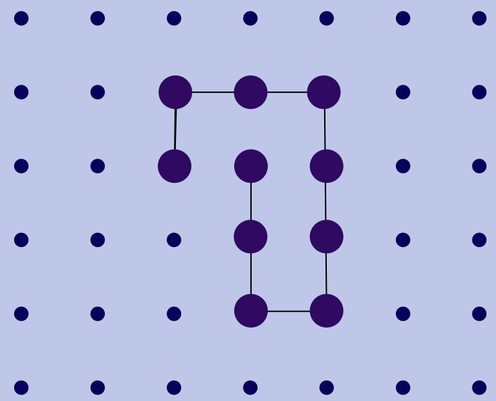
Energy: -4



Protein Folding, NP-hard?

Finding most stable conformation optimal fold is considered NP-hard

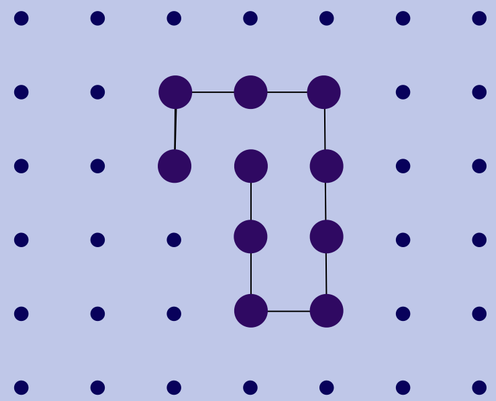
Length: 12
Conformations: 59.049



Protein Folding, NP-hard?

Finding most stable conformation optimal fold is considered NP-hard

Length: 12
Conformations: 59.049



Protein Length	Number of solutions
5	27
10	6561
15	1.594.323

The problem scales exponentially with the length of the protein: 1 x 1 x 3(n- 2)

Scope

Uniform Random Sampling

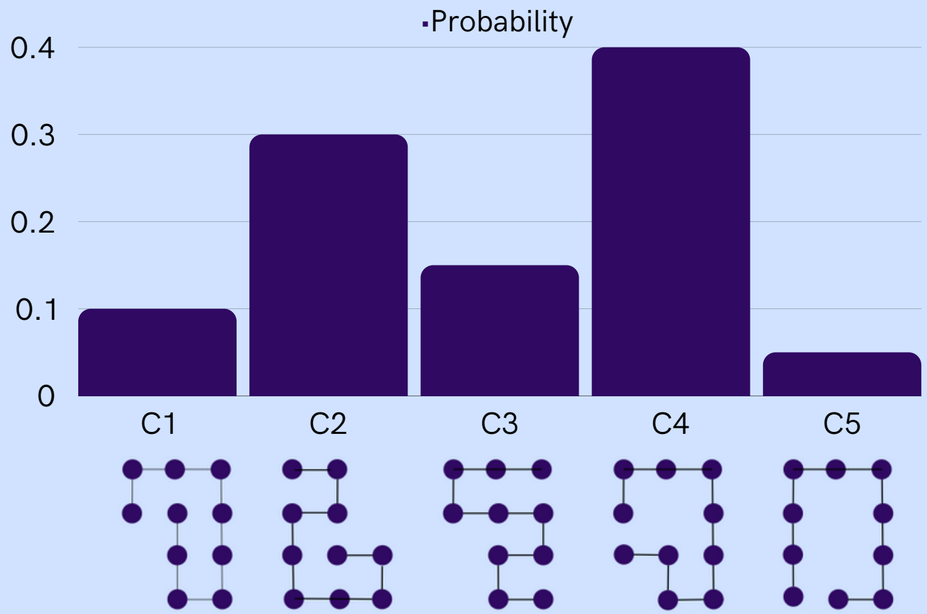
Can we ensure the **uniform randomness** in initial populations before folding proteins (e.g. with Genetic algorithms)?

What is the impact of **different sampling techniques** on the quality and diversity of solutions?

Uniform Random Sampling

What does uniform random mean?

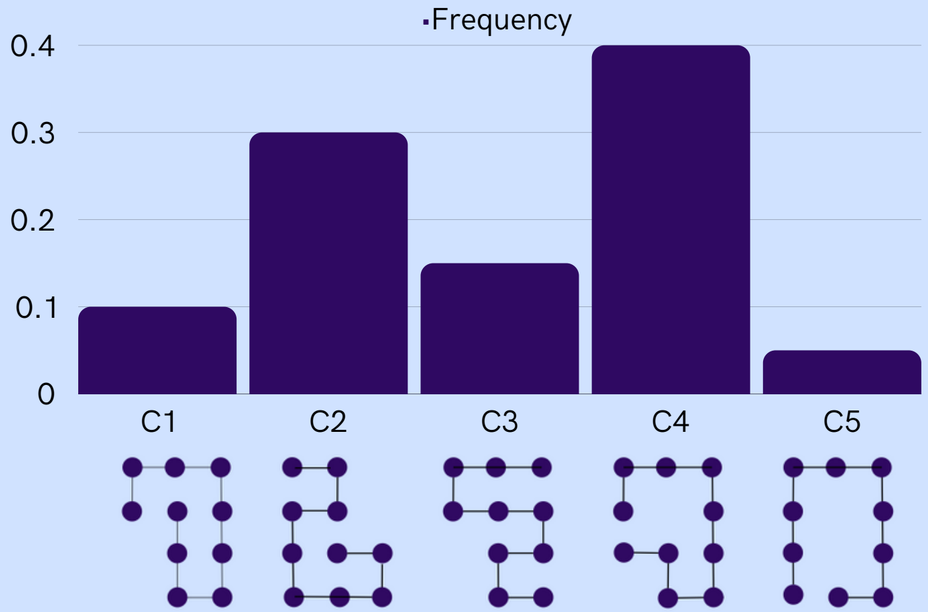
❌ Not Uniform Random



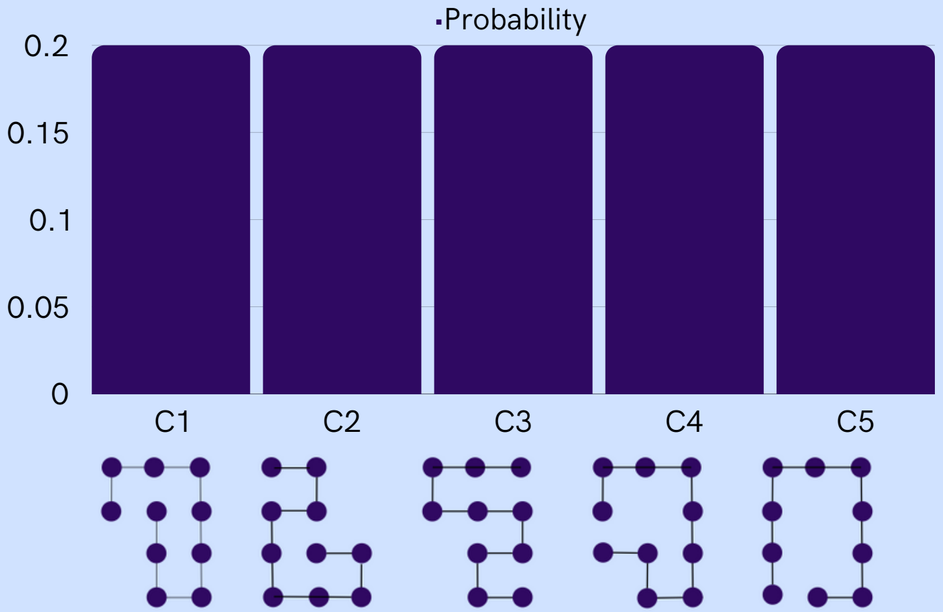
Uniform Random Sampling

What does uniform random mean?

Not Uniform Random



Uniform Random



Uniform Random Samling

Why do we want uniform random samples?

Limited Diversity

- **Problem:** Non-uniform random sampling leads to a population that is not diverse.
- **Result:** The algorithm converges to a local optimum. It doesn’t have the diversity needed to explore other potentially superior solutions.

Uniform Random Samling

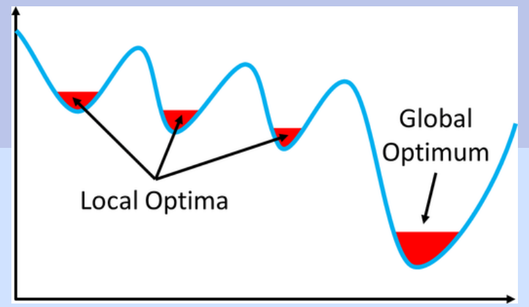
Why do we want uniform random samples?

Limited Diversity

- **Problem:** Non-uniform random sampling leads to a population that is not diverse.
- **Result:** The algorithm converges to a local optimum. It doesn't have the diversity needed to explore other potentially superior solutions.

Premature Convergence

- **Problem**:The algorithm settles on a suboptimal solution too quickly.
- **Result**: It doesn't explore enough of the solution space to find the global optimum.



Experiment

Experiment

6 Different sampling methods

*Work by Jansen et al. (2023)

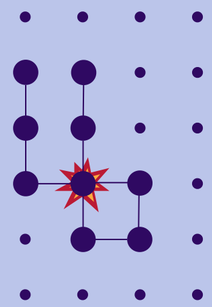
Experiment

6 Different sampling methods

0*

Collision Sampling

Counts the number of **collisinos**

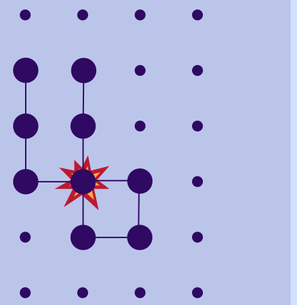
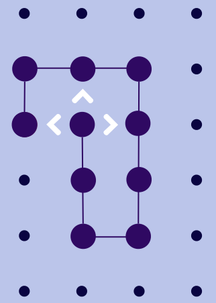


Lengths	**{5, 10, 15 ... 200}**
Placement	**Stochastic**
Backtracking	**---**

*Work by Jansen et al. (2023)

Experiment

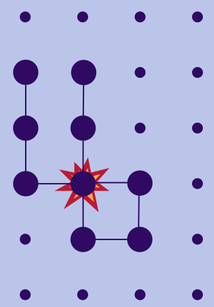
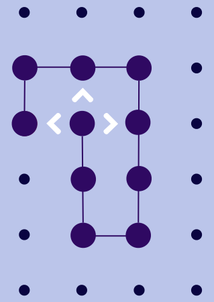
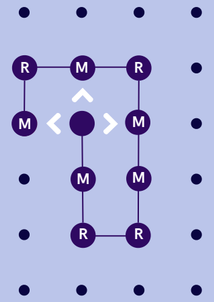
6 Different sampling methods

	0*	1
	Collision Sampling Counts the number of **collisinos**	**Stochastic Break Sampling** **Break** when no valid neighbour
		
Lengths	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}
Placement	Stochastic	Stochastic
Backtracking	---	---

*Work by Jansen et al. (2023)

Experiment

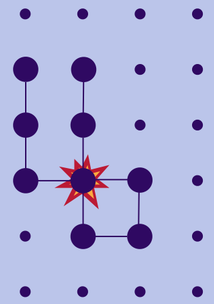
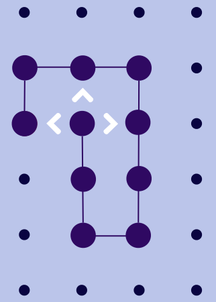
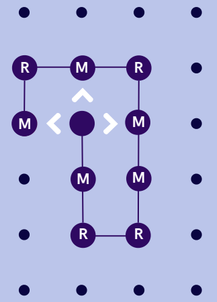
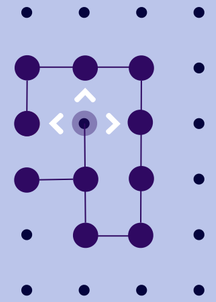
6 Different sampling methods

	0*	1	2
	Collision Sampling Counts the number of **collisinos**	**Stochastic Break Sampling** **Break** when no valid neighbour	**Deterministic Break Sampling** **Break** when no valid neighbour
			
Lengths	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}
Placement	Stochastic	Stochastic	Determistic
Backtracking	---	---	---

*Work by Jansen et al. (2023)

Experiment

6 Different sampling methods

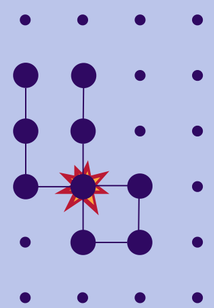
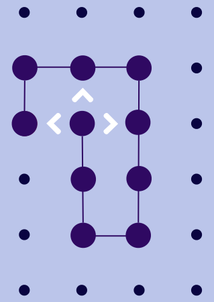
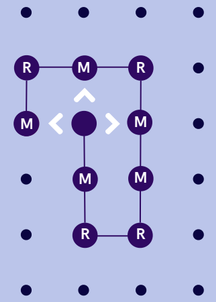
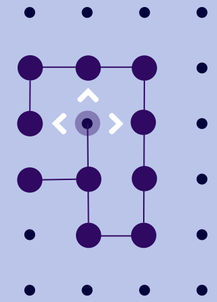
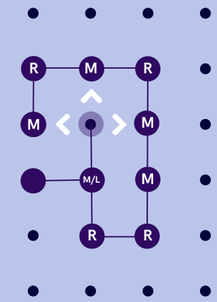
	0*	1	2	3
	Collision Sampling Counts the number of **collisinos**	**Stochastic Break Sampling** **Break** when no valid neighbour	**Deterministic Break Sampling** **Break** when no valid neighbour	**Stochastic Recursive Sampling** **Backtrack** when no valid neighbour
				
Lengths	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}	{5, 10, 15 ... 100}
Placement	Stochastic	Stochastic	Determistic	Stochastic
Backtracking	---	---	---	Stochastic

M R L M R L
L M R L M R
R L M R L M

*Work by Jansen et al. (2023)

Experiment

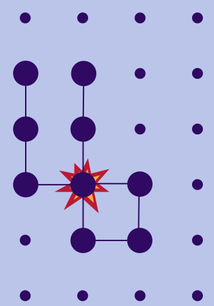
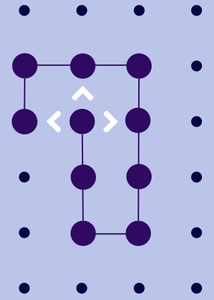
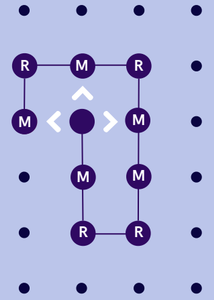
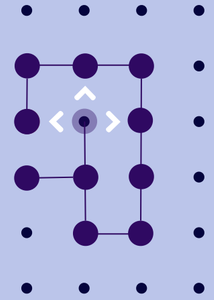
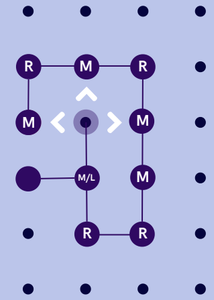
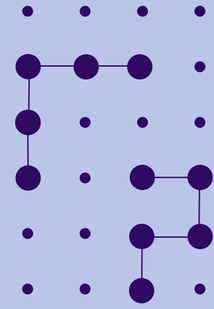
6 Different sampling methods

	0*	1	2	3	4
	Collision Sampling Counts the number of **collisinos**	**Stochastic Break Sampling** **Break** when no valid neighbour	**Deterministic Break Sampling** **Break** when no valid neighbour	**Stochastic Recursive Sampling** **Backtrack** when no valid neighbour	**Heuristic Recursive Sampling** **Backtrack** when no valid neighbour
					
Lengths	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}	{5, 10, 15 ... 100}	{5, 10, 15 ... 30}
Placement	Stochastic	Stochastic	Determistic	Stochastic	Stochastic
Backtracking	---	---	---	Stochastic	Heuristic
				$\begin{bmatrix} M & R & L & M & R & L \\ L & M & R & L & M & R \\ R & L & M & R & L & M \end{bmatrix}$	$\begin{bmatrix} L & L & L & L & L & L \\ M & M & M & M & M & M \\ R & R & R & R & R & R \end{bmatrix}$

*Work by Jansen et al. (2023)

Experiment

6 Different sampling methods

	0*	1	2	3	4	5
	Collision Sampling Counts the number of **collisinos**	**Stochastic Break Sampling** **Break** when no valid neighbour	**Deterministic Break Sampling** **Break** when no valid neighbour	**Stochastic Recursive Sampling** **Backtrack** when no valid neighbour	**Heuristic Recursive Sampling** **Backtrack** when no valid neighbour	**Solution Space Sampling** Find **all** conformations
						
Lengths	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}	{5, 10, 15 ... 200}	{5, 10, 15 ... 100}	{5, 10, 15 ... 30}	{5, 10, 15}
Placement	Stochastic	Stochastic	Determistic	Stochastic	Stochastic	---
Backtracking	---	---	---	Stochastic	Heuristic	---
				$\begin{bmatrix} M & R & L & M & R & L \\ L & M & R & L & M & R \\ R & L & M & R & L & M \end{bmatrix}$	$\begin{bmatrix} L & L & L & L & L & L \\ M & M & M & M & M & M \\ R & R & R & R & R & R \end{bmatrix}$	

*Work by Jansen et al. (2023)

Experiment

Sampling information and parameters

X = number of folded proteins per length

N = protein lengths ∈ {5, 10, 15... N}.

	Amino Acid Length	Num Hydrophobic	Num Polar	1D protein	2D protein	Amino Acids on Grid	Trimmed 2D protein	Shape 2D protein	Amino Acid Order	H-Bonds	H-Ratio	Recursions	Time Taken (s)
19995	100	29	71	['P', 'P', 'P', 'P', 'P', 'P', 'P', 'P', 'P', 'P', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' ''...	100	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' 'H...	(16, 18)	[('P', (100, 100)), ('P', (100, 101)), ('P', (...	4	0.04	6	7.152557e-07
19996	100	67	33	['H', 'P', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' ''...	100	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' 'H' 'P' 'H' 'H' '...	(12, 17)	[('H', (100, 100)), ('P', (101, 100)), ('H', (...	19	0.19	0	1.192093e-06
19997	100	80	20	['H', 'H', 'P', 'H', 'H', 'H', 'H', 'H', 'H', 'H', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' ''...	100	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' 'H' 'H' 'H' 'H' '...	(11, 23)	[('H', (100, 100)), ('H', (101, 100)), ('P', (...	24	0.24	0	0.000000e+00
19998	100	96	4	['H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' ''...	100	[['' '' '' '' '' 'H' 'H' '' '' '' '' '' '' '' '' '']...	(13, 14)	[('H', (100, 100)), ('H', (100, 99)), ('H', (1...	56	0.56	1	0.000000e+00
19999	100	55	45	['H', 'H', 'P', 'H', 'P', 'H', 'H', 'P', 'H', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' '' ''...	100	[['' '' '' '' '' '' 'P' 'H' '' '' '' '' '' '' '' '' '']\n ...	(24, 13)	[('H', (100, 100)), ('H', (101, 100)), ('P', (...	9	0.09	216331	0.000000e+00

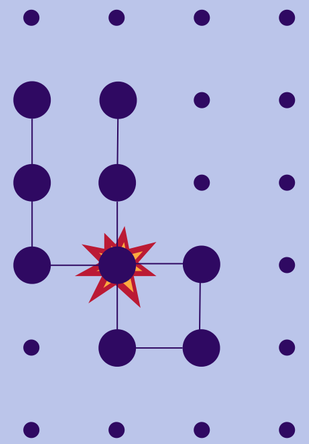
Results

Results

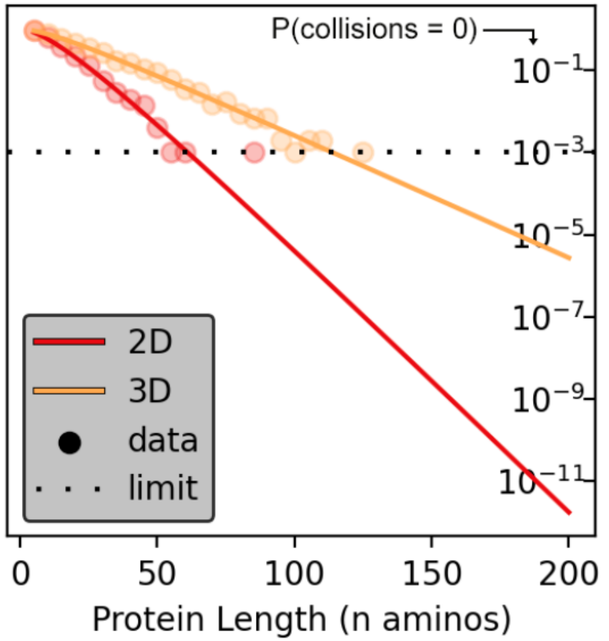
Experiment 0*

Collision Sampling

Counts the number of **collisinos**



Valid conformations per amino acid length



Collisions increase as protein get longer

Chance of zero-collision conformation drops exponentially in n.

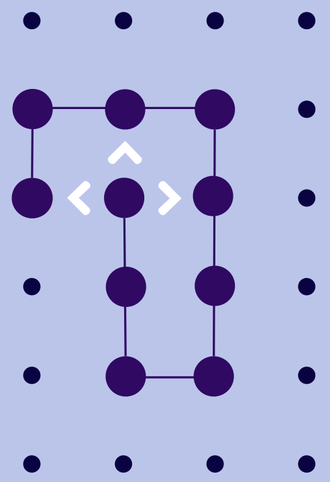
https://www.scitepress.org/publishedPapers/2023/122485l

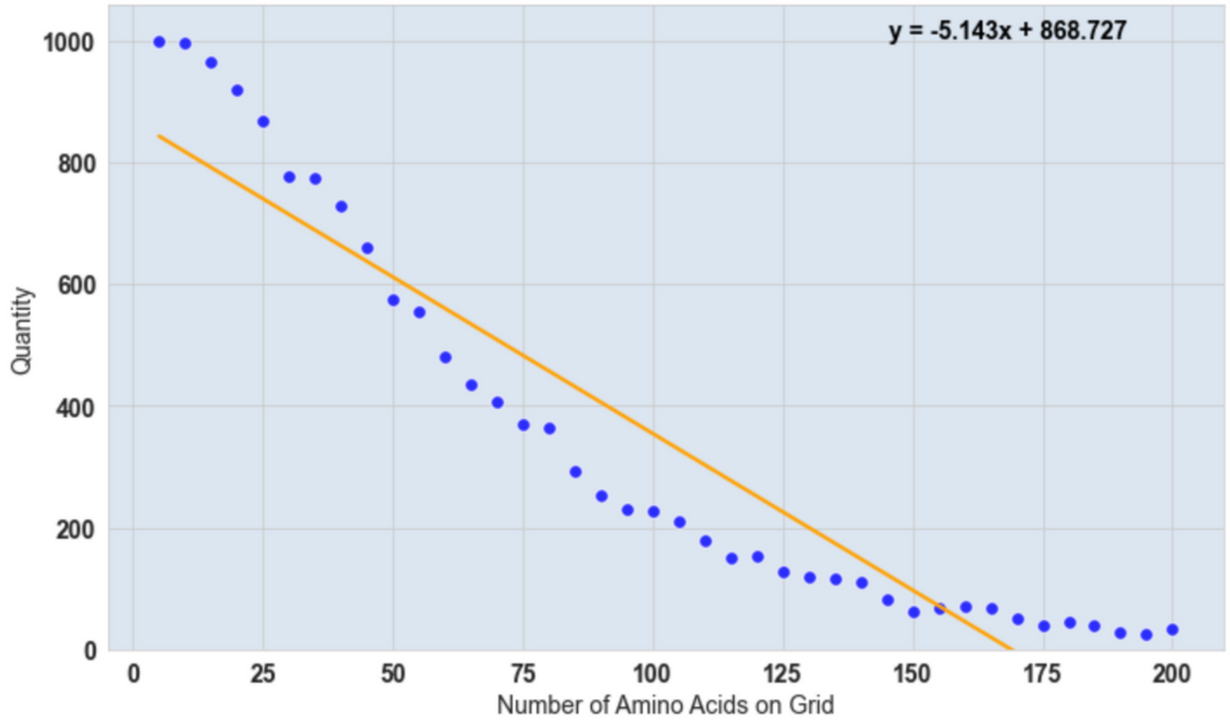
Results

Experiment 1

Stochastic Break Sampling

Break when no valid neighbour





Sampling for n < 20
all conformations are valid.

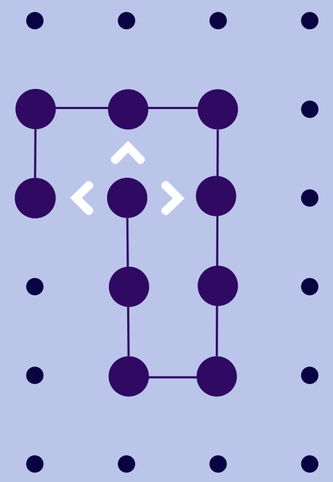
n=100 22,7% is valid
n=200 3.4% is valid

Results

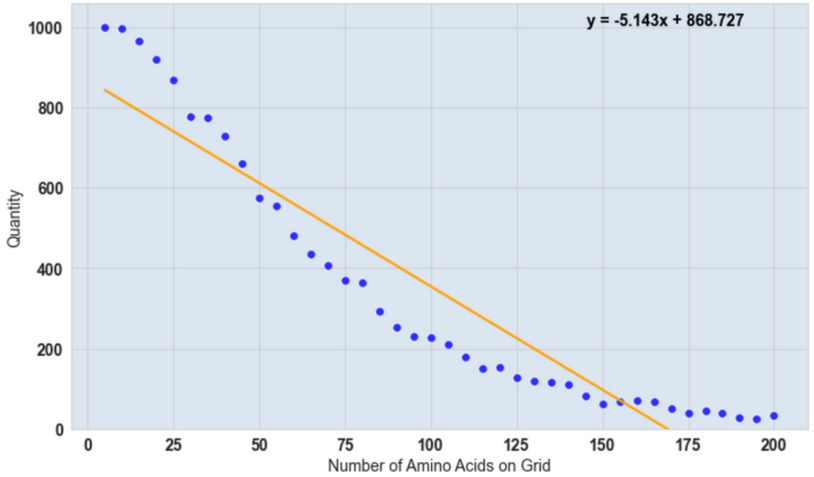
Experiment 1

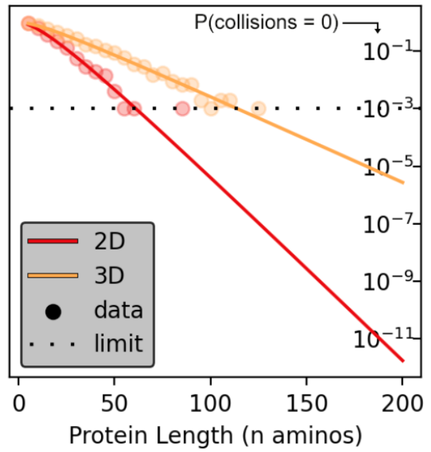
Stochastic Break Sampling

Break when no valid neighbour



Valid conformations per amino acid length





Experiments show similair results

Chance of zero-collision conformation drops exponentially in n.

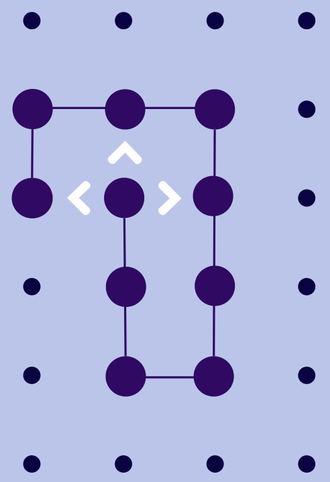
https://www.scitepress.org/publishedPapers/2023/122485l

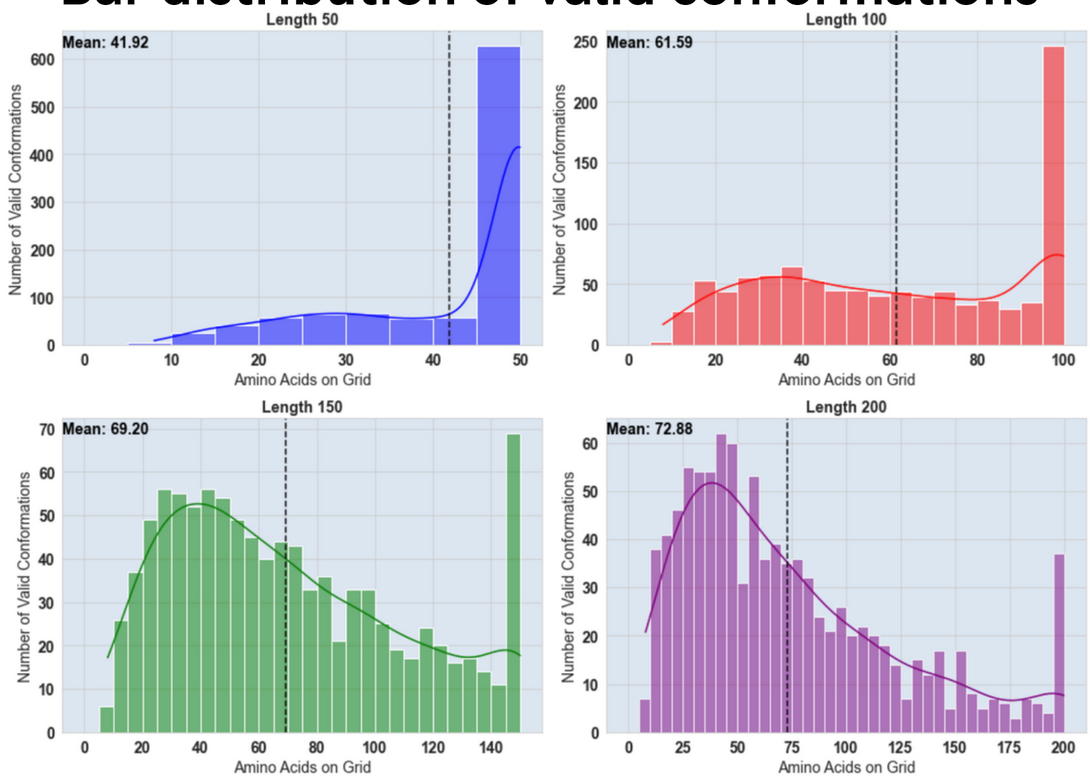
Results

Experiment 1

Stochastic Break Sampling

Break when no valid neighbour





Shorter proteins exhibit peak near maximums

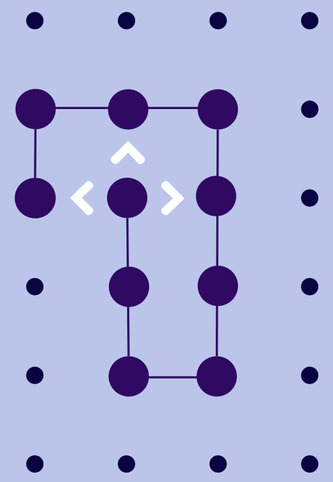
Longer proteins show a broader distribution

Results

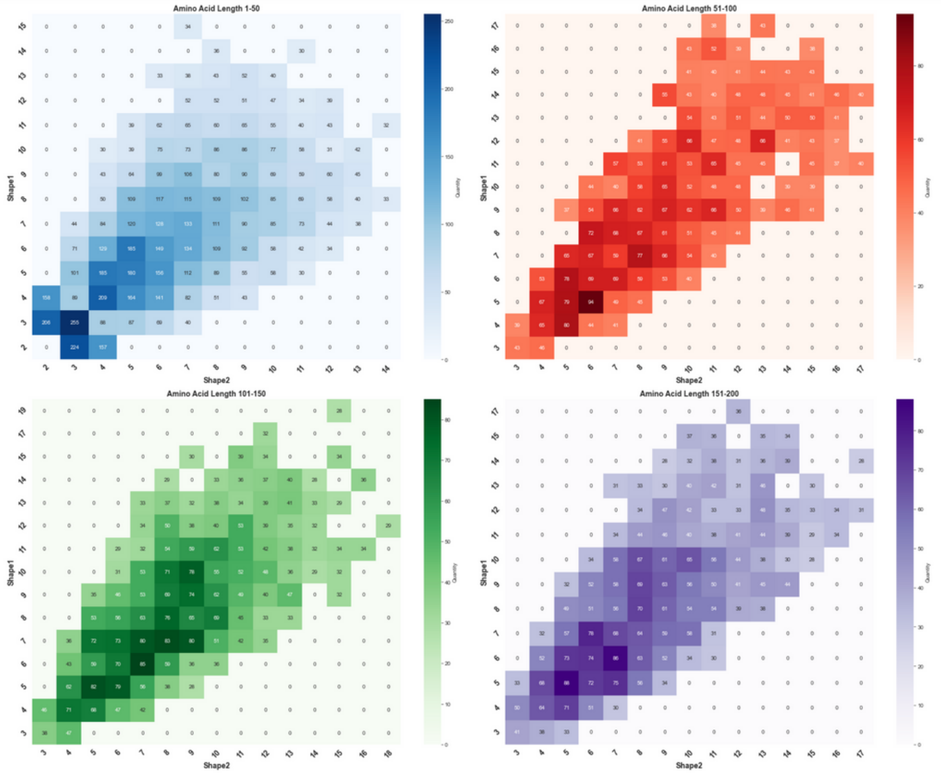
Experiment 1

Stochastic Break Sampling

Break when no valid neighbour



Heatmap distribution of shape combinations



Shorter proteins manifest concentrated shape patterns

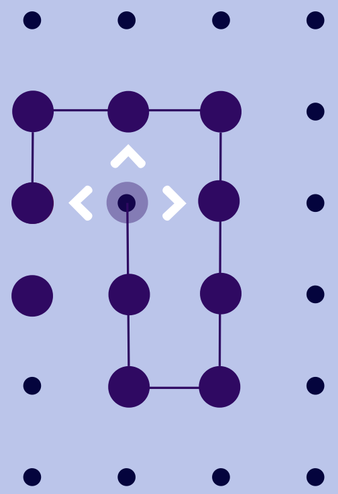
Longer proteins exhibit show more varied combinations

Results

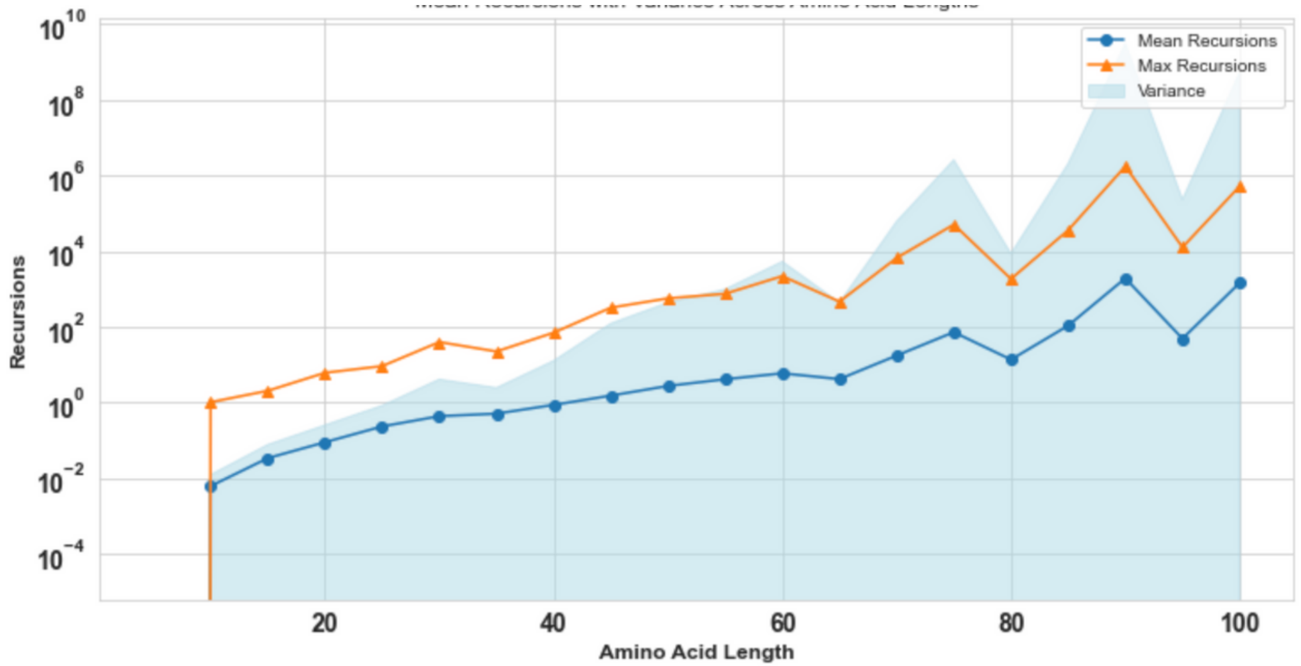
Experiment 3

Stochastic Recursive Sampling

Backtrack when no valid neighbour



Distribution of mean and max recursion



Short proteins have few recursions

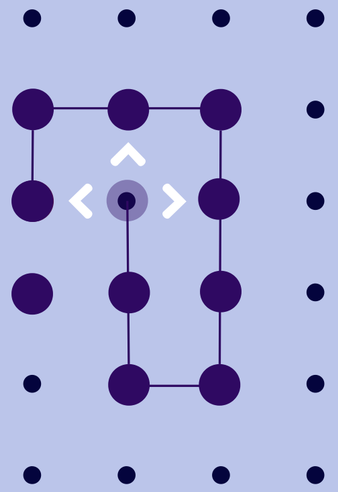
Longer proteins longer show exponentially more recursions

Results

Experiment 3

Stochastic Recursive Sampling

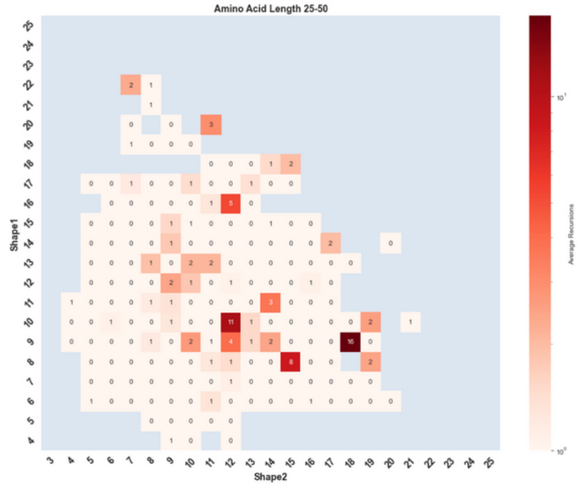
Backtrack when no valid neighbour

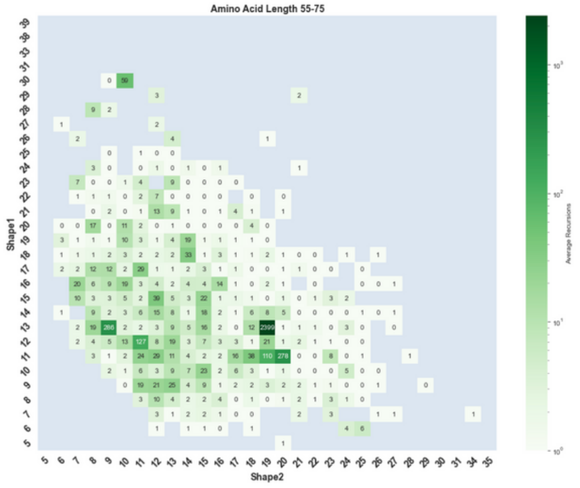


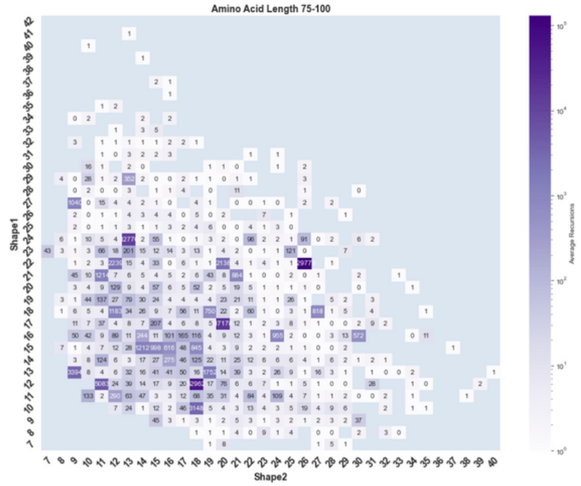
Heatmap distribution of mean recursion per shape

Average recursion increases when a protein gets longer

Longer proteins are more sensitive to outliers





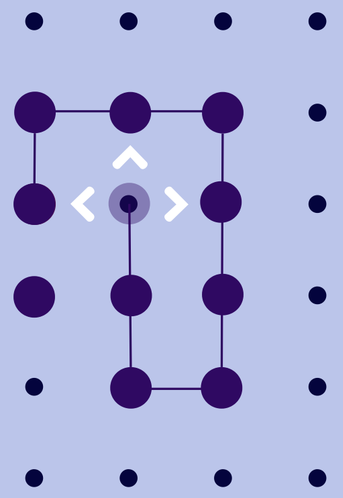


Results

Experiment 5

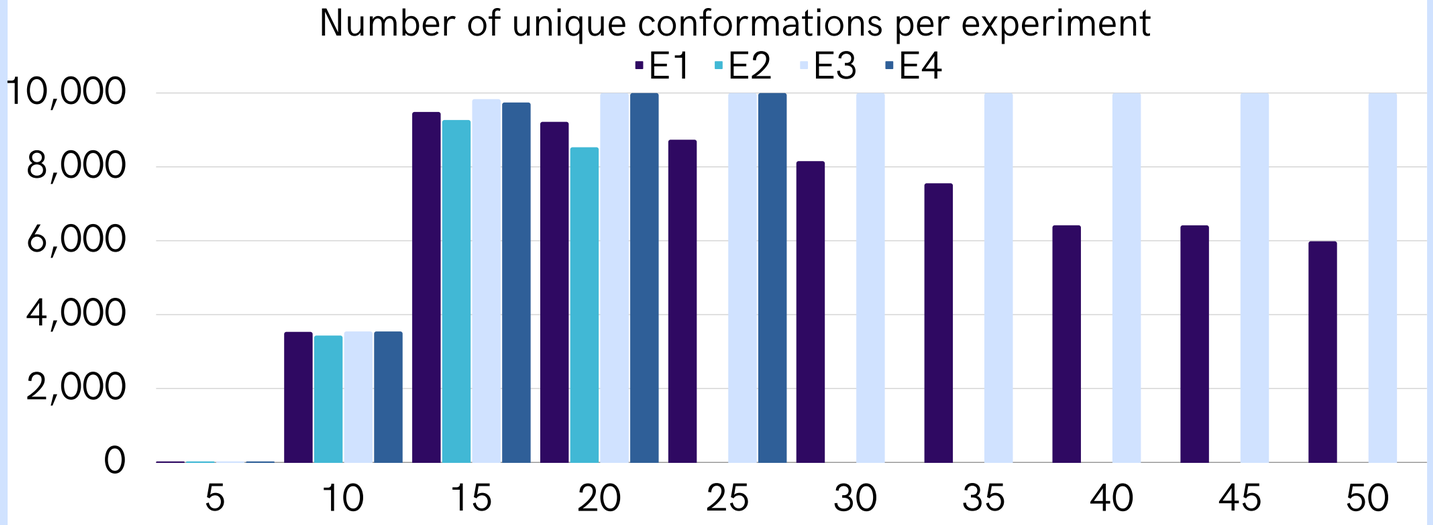
Stochastic Recursive Sampling

Backtrack when no valid neighbour



Benchmarking against sampling space

n	Sn	Sv	Su
5	27	25	13
10	6561	4067	2034
15	1,594,323	593,611	296,806



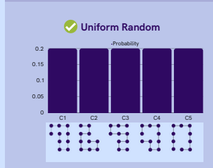
Diversity of sample space is explored for short proteins

For longer proteins the unique conformations drop for non-backtracking

Conclusion

What do the results say?

Can we ensure the uniform randomness in initial populations before folding proteins?



For **short proteins**, it seems possible. However, the average protein is 200-300 acids. No real life implications.

For **longer proteins**, Probably only with backtracking, but this has many limitations and is hard(er) to verify.

- Scaling
- Recursion depth

What is the impact of different sampling techniques on the quality and diversity of solutions?

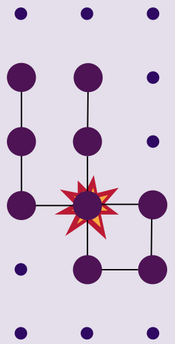
Break sampling seems to work for for shorter proteins. Details have to be worked out further.

Recursive sampling seems to improve the quality (valid samples) and diversity (unique samples) of the initial population landscape. Details have to be worked out further.

Thanks!

Collision Sampling

Counts the number of **collisinos**



1

Stochastic Break Sampling

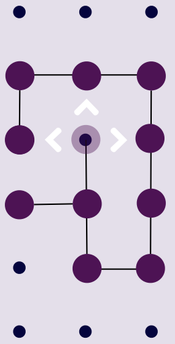
Break when no valid neighbour

Deterministic Break Sampling

Break when no valid neighbour

Stochastic Recursive

Backtrack when no valid



4