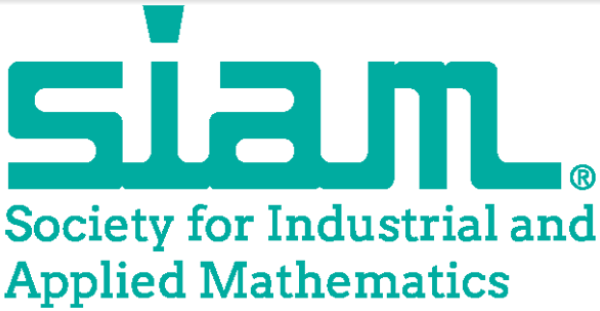


Machine-learning a virus assembly fitness landscape

SIAM Algebraic geometry, data science and fundamental physics
Bern, July 12, 2019

Pierre-Philippe Dechant
work with Y-H He and R Twarock

Pro Vice Chancellor's Office, York St John University
York Cross-disciplinary Centre for Systems Analysis, University of York
Department of Mathematics, University of York

Rationale

- Input vector: Genotype/Phenotype of length 12 (packaging signal strengths in 3 bands)
- Output vector: Assembly efficiency (out of 2000 possible capsids)
- Black box: Molecular dynamics simulations (computationally very costly)

Rationale

- Input vector: Genotype/Phenotype of length 12 (packaging signal strengths in 3 bands)
- Output vector: Assembly efficiency (out of 2000 possible capsids)

- Black box: Machine learning via a neural network

Rationale

	Genome	Fitness
0	111111111111	200
1	111111111112	1393
2	111111111113	1869
3	111111111121	1597
4	111111111122	1896
5	111111111123	1960
6	111111111131	1875
7	111111111132	1959
8	111111111133	1961
9	111111111211	1639
10	111111111212	1683
11	111111111213	1895
12	111111111221	1848
13	111111111222	1904
14	111111111223	1964
15	111111111231	1904
16	111111111232	1949
17	111111111233	1959
18	111111111311	1852
19	111111111312	1858

$3^{12} \sim \frac{1}{2}$ Million data points

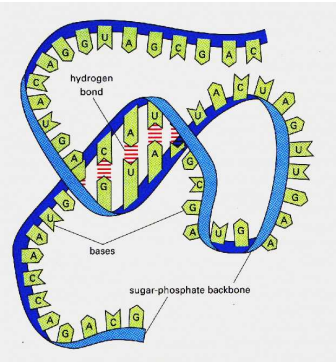
Overview

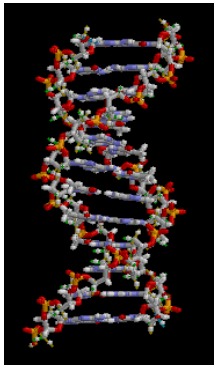
1.
   - Virus structure and assembly
   - Toy model and evolutionary fitness landscape

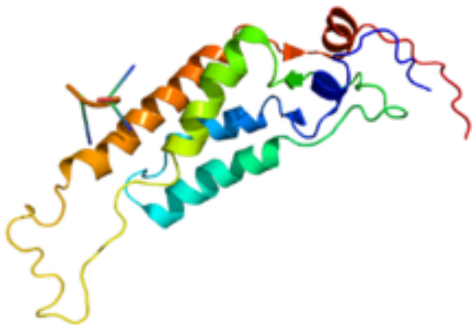
2.
   - Neural networks
   - Predictions

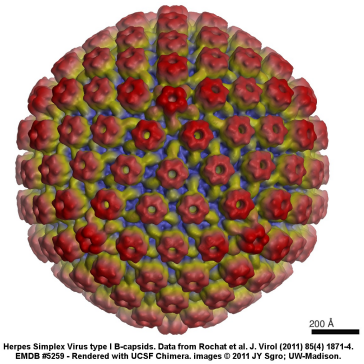
What is a Virus?

- Piece of genetic information in the form of RNA or DNA
- Protected by a protein shell: capsid made of geometric protein building block

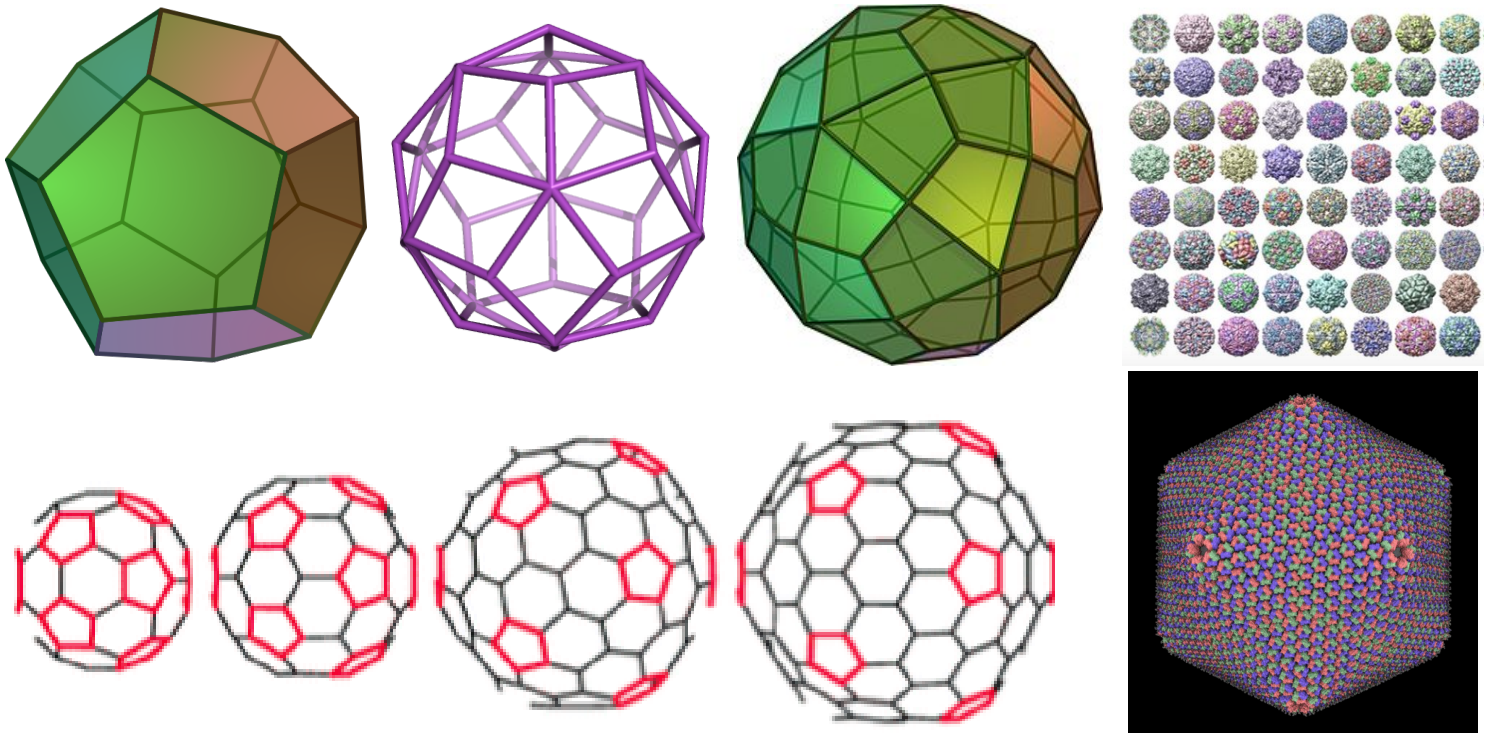






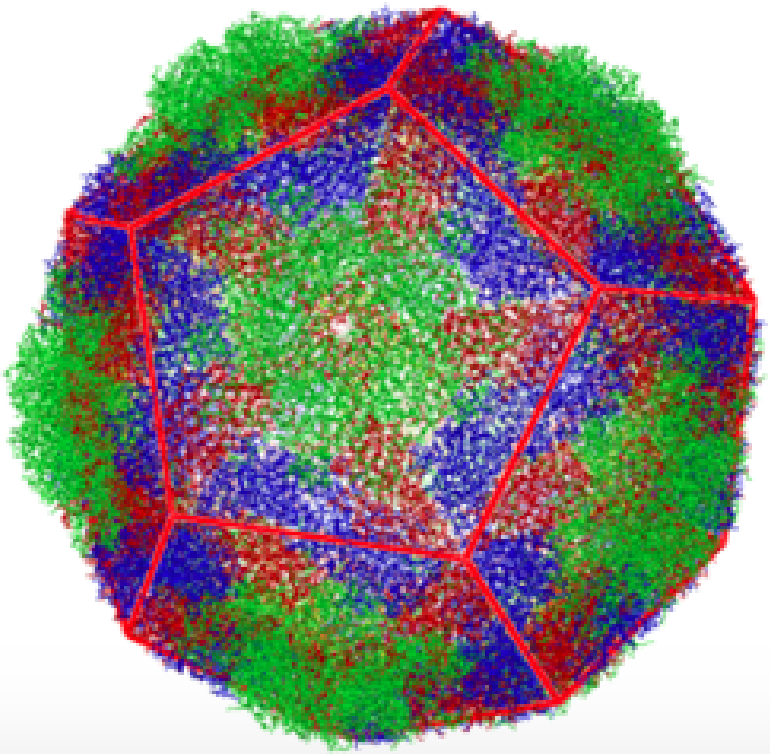


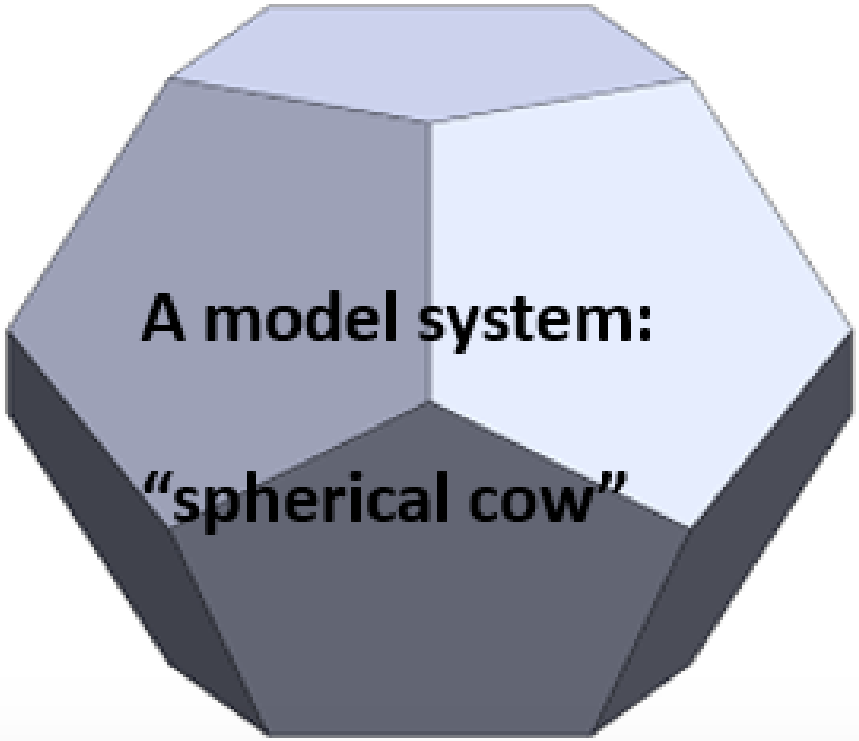
Most viruses are icosahedral



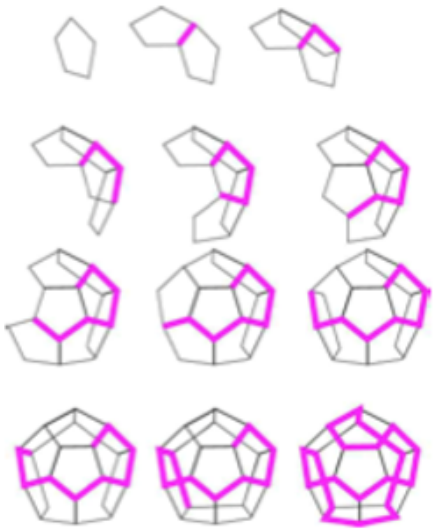
- Highly developed structure theory
- Nucleic acid component thought to be disordered

Simplest model: a dodecahedron

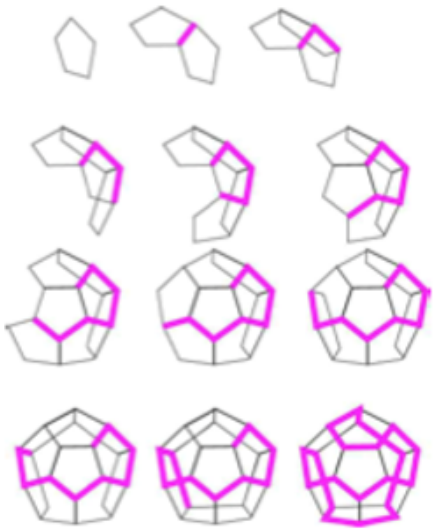


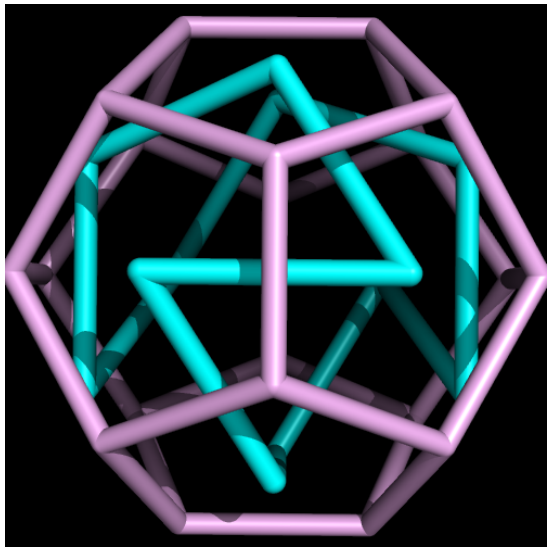


Assembly and thermodynamics – Hamiltonian paths

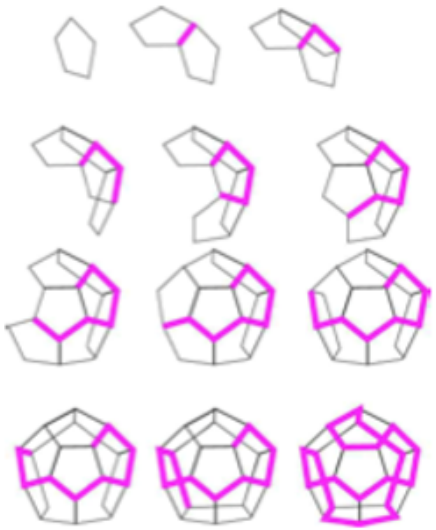


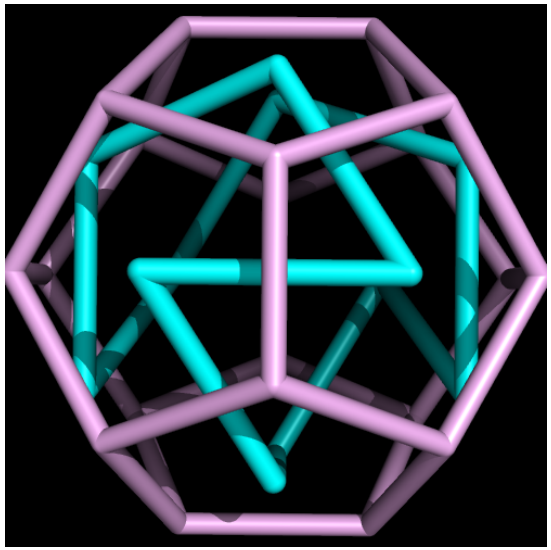
Assembly and thermodynamics – Hamiltonian paths

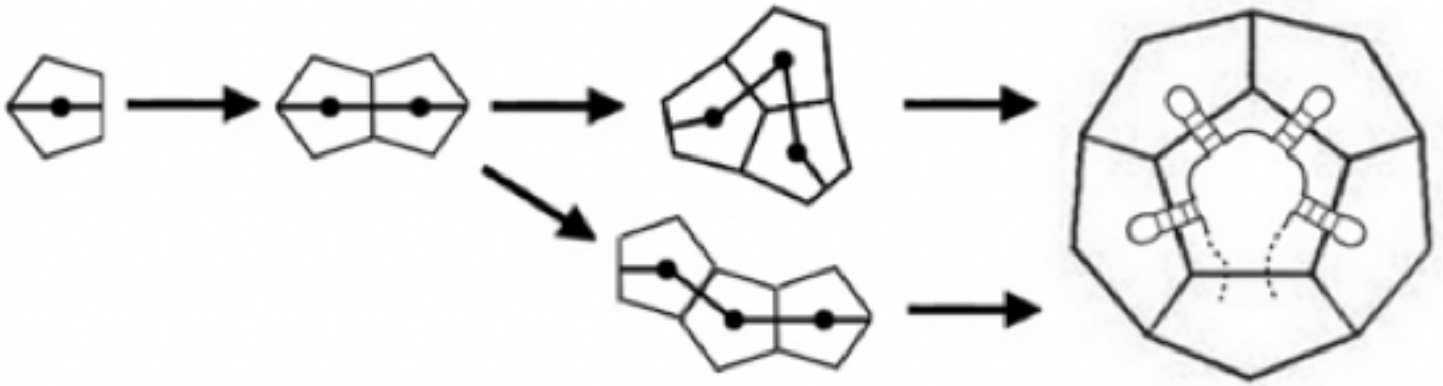




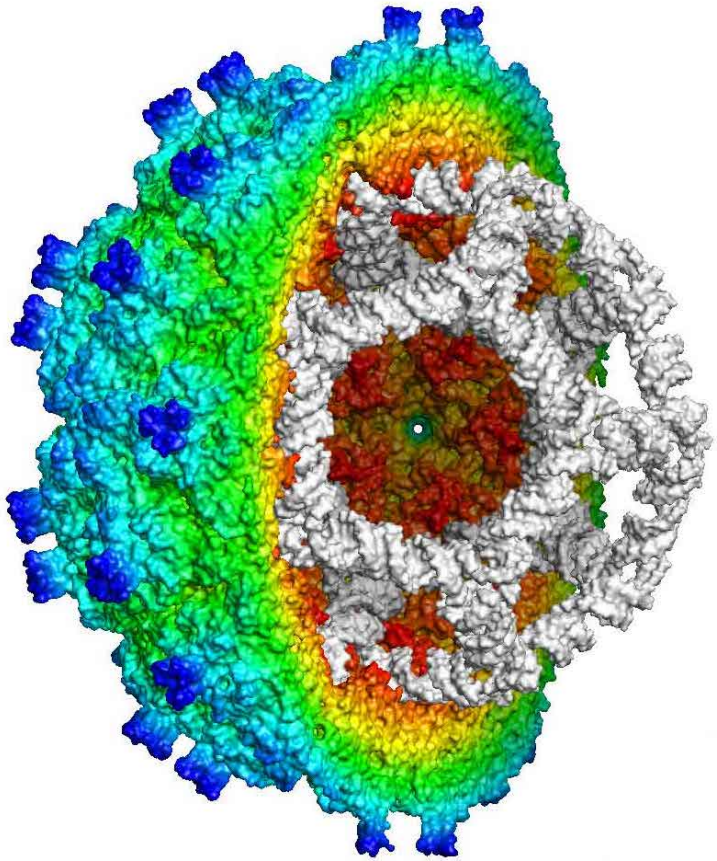
Assembly and thermodynamics – Hamiltonian paths





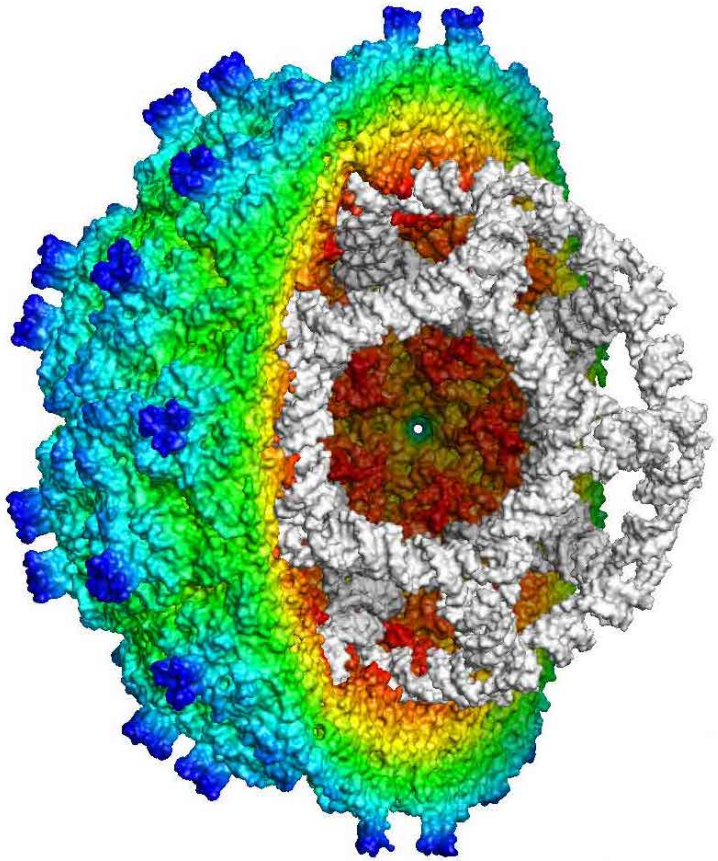


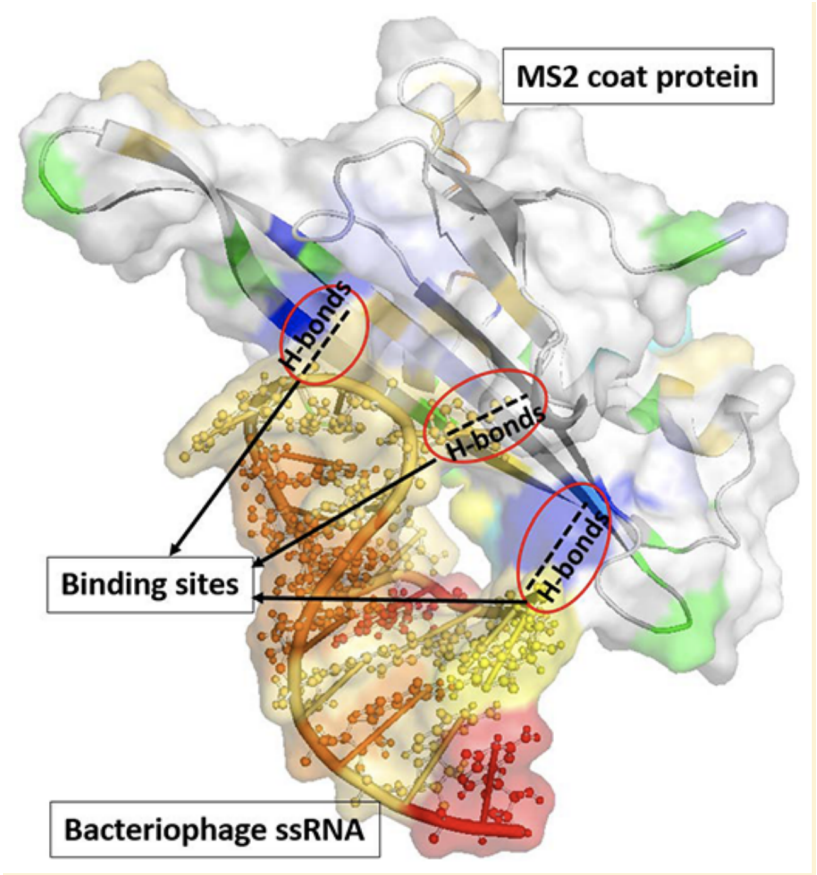
3D distribution: RNA-CP contacts



There are specific interactions between RNA and coat protein (CP) given by icosahedral symmetry axes

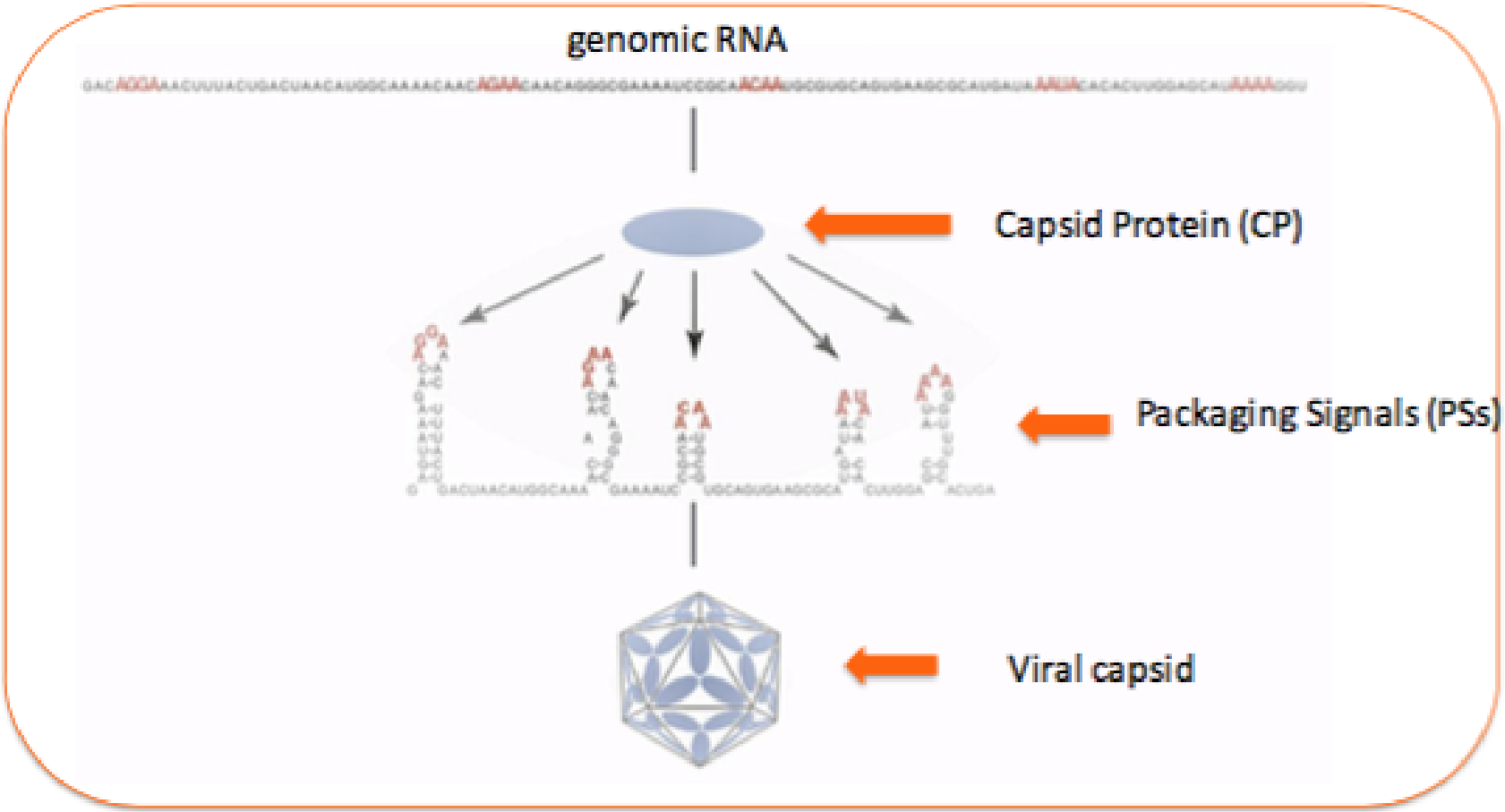
3D distribution: RNA-CP contacts



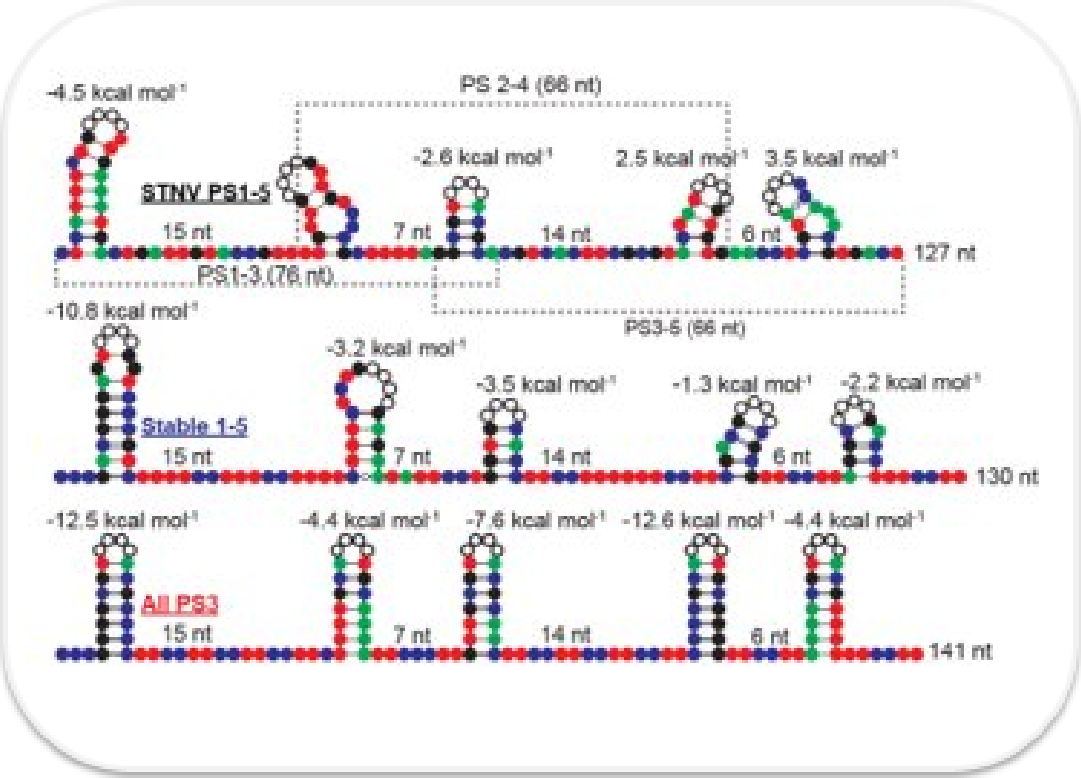


There are specific interactions between RNA and coat protein (CP) given by icosahedral symmetry axes

Packaging signal-mediated assembly

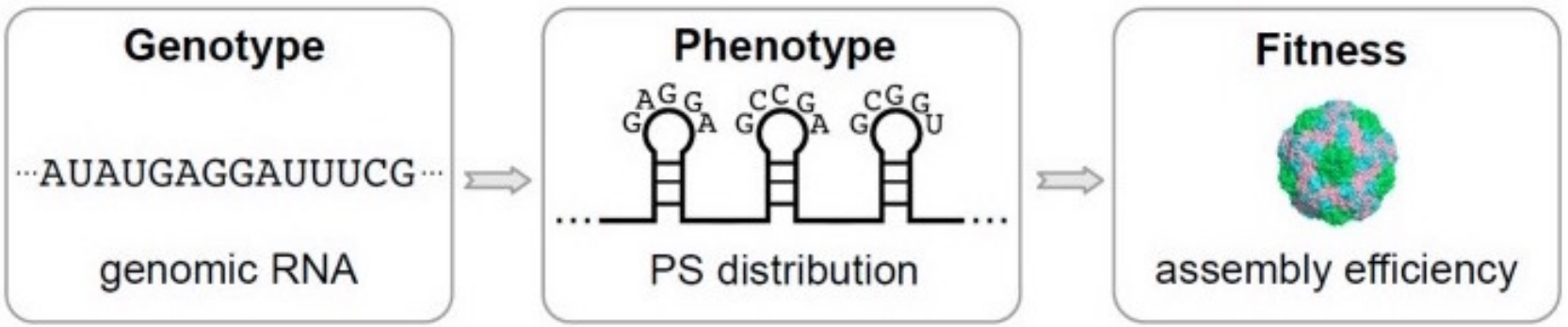


Engineering Packaging Signals to make VLPs

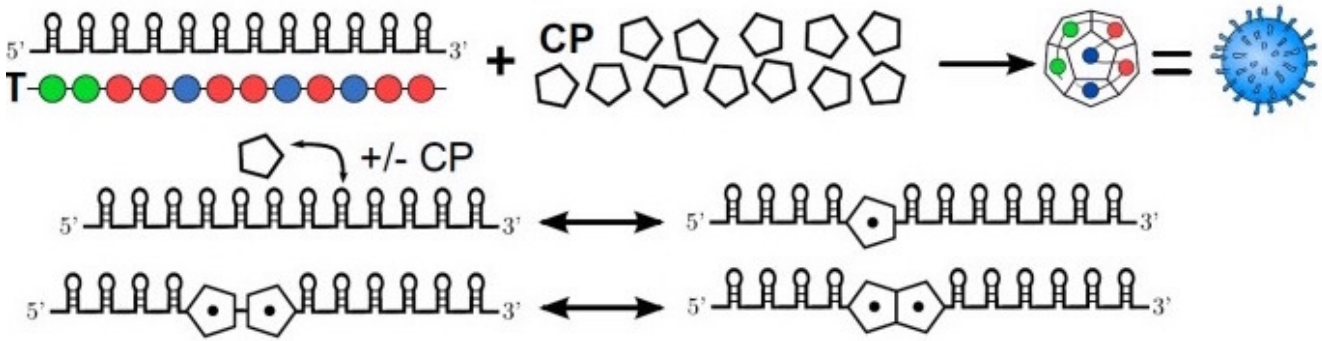


Virus-like particles with improved PS sequences assemble twice as efficiently. Potential applications to vaccines or drug delivery.

Genotype – Phenotype – Fitness map

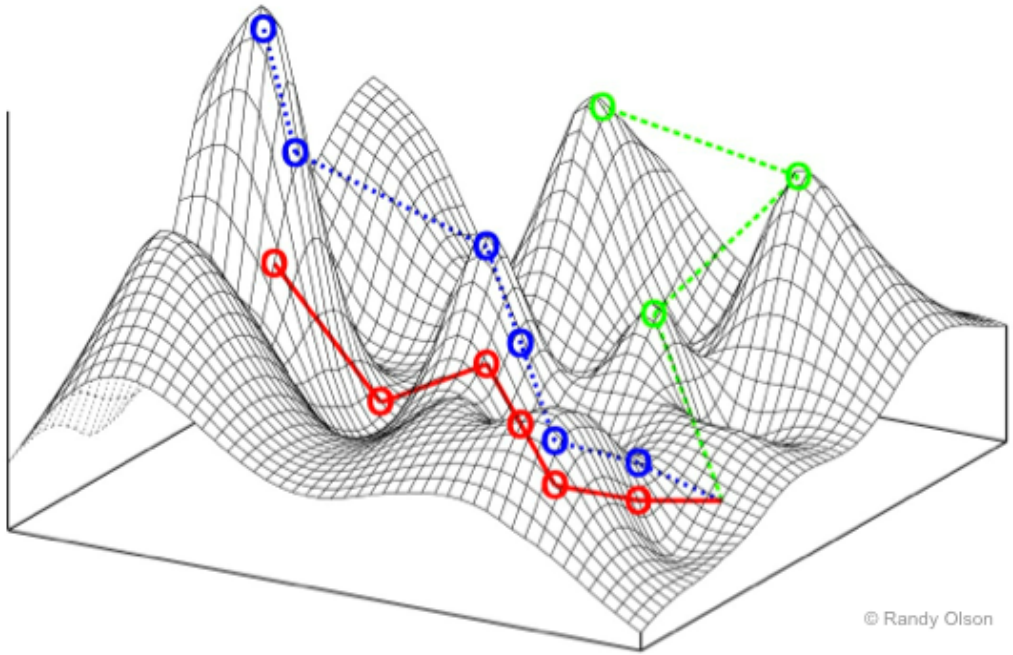


Simplest model: the dodecahedron



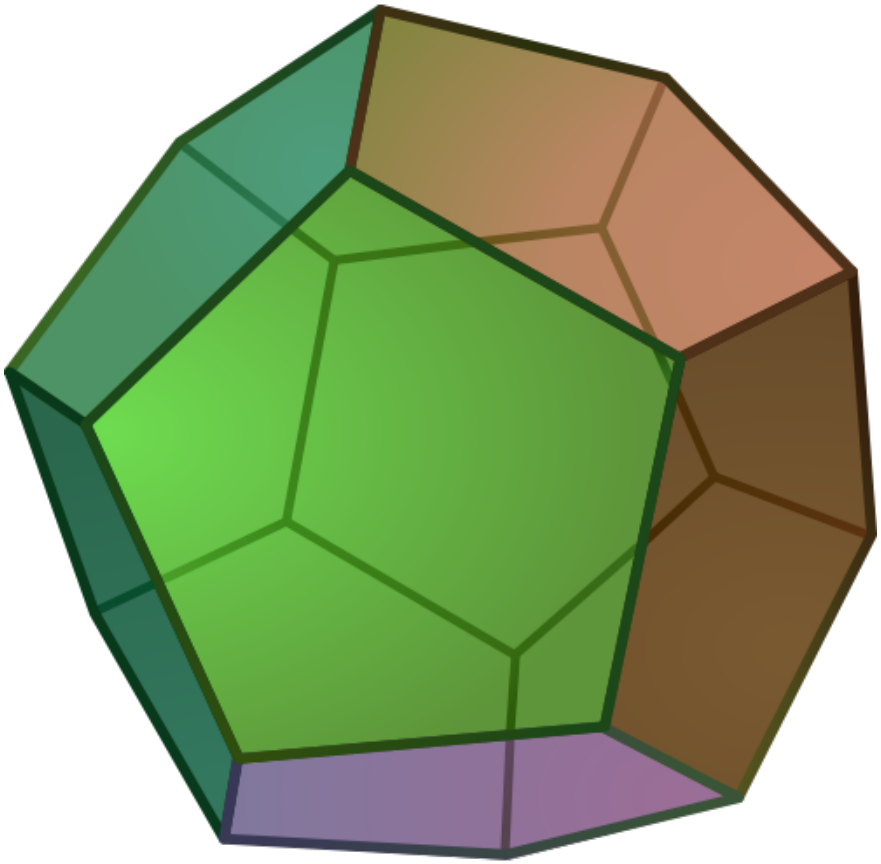
- 12 PSs in 3 bands (strong/intermediate/weak, 12/8/4, 3/2/1, green/blue/red)
- Molecular dynamics simulation: stochastically select one possible reaction at a time
- Enough resources for 2000 virus capsids

Fitness Landscape



Generally messy (many contributions) and difficult to quantify.
Here capture the assembly contribution for the phenotype space of 3^{12} points with (stochastic) assembly efficiency (< 2000).

Fundamental Physics



Genotype–fitness map

	Genome	Fitness
0	111111111111	200
1	111111111112	1393
2	111111111113	1869
3	111111111121	1597
4	111111111122	1896
5	111111111123	1960
6	111111111131	1875
7	111111111132	1959
8	111111111133	1961
9	111111111211	1639
10	111111111212	1683
11	111111111213	1895
12	111111111221	1848
13	111111111222	1904
14	111111111223	1964
15	111111111231	1904
16	111111111232	1949
17	111111111233	1959
18	111111111311	1852
19	111111111312	1858

$3^{12} \sim \frac{1}{2}$ Million data points

Overview

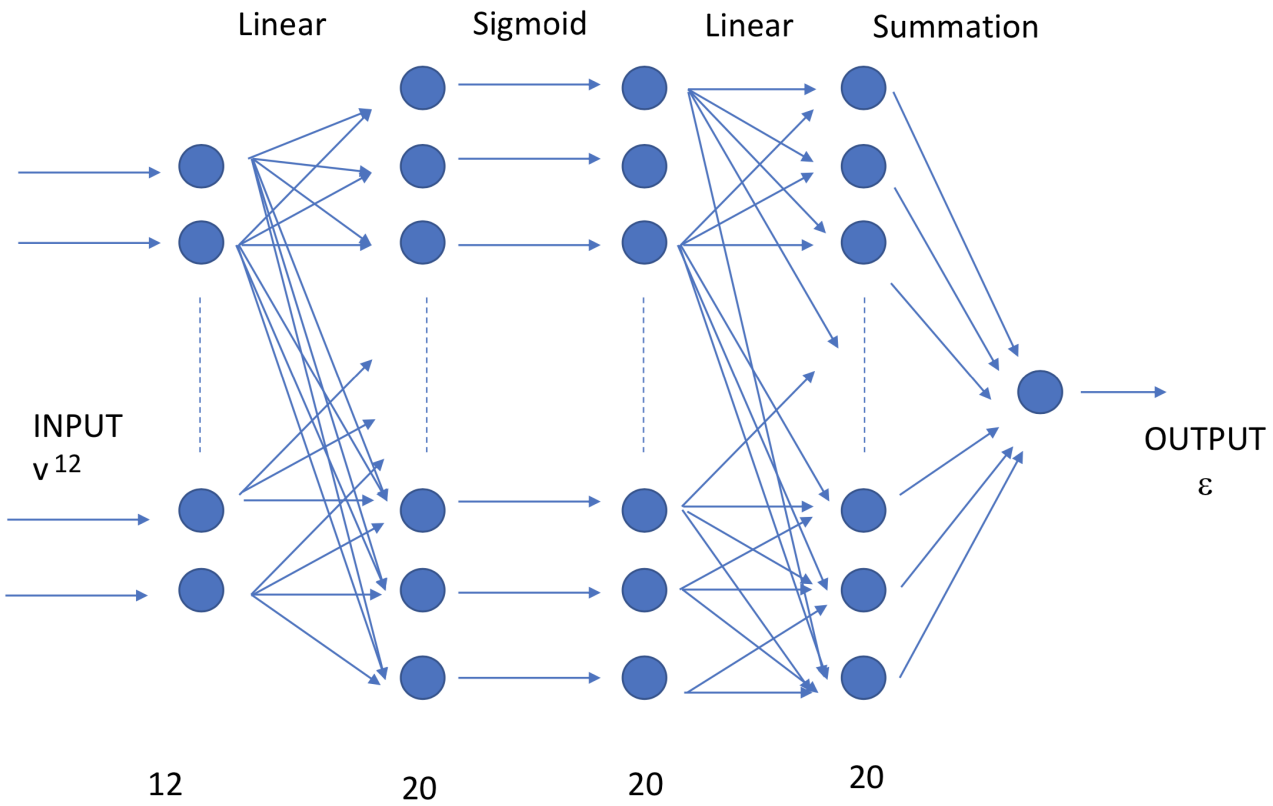
1.
   - Virus structure and assembly
   - Toy model and evolutionary fitness landscape

2.
   - Neural networks
   - Predictions

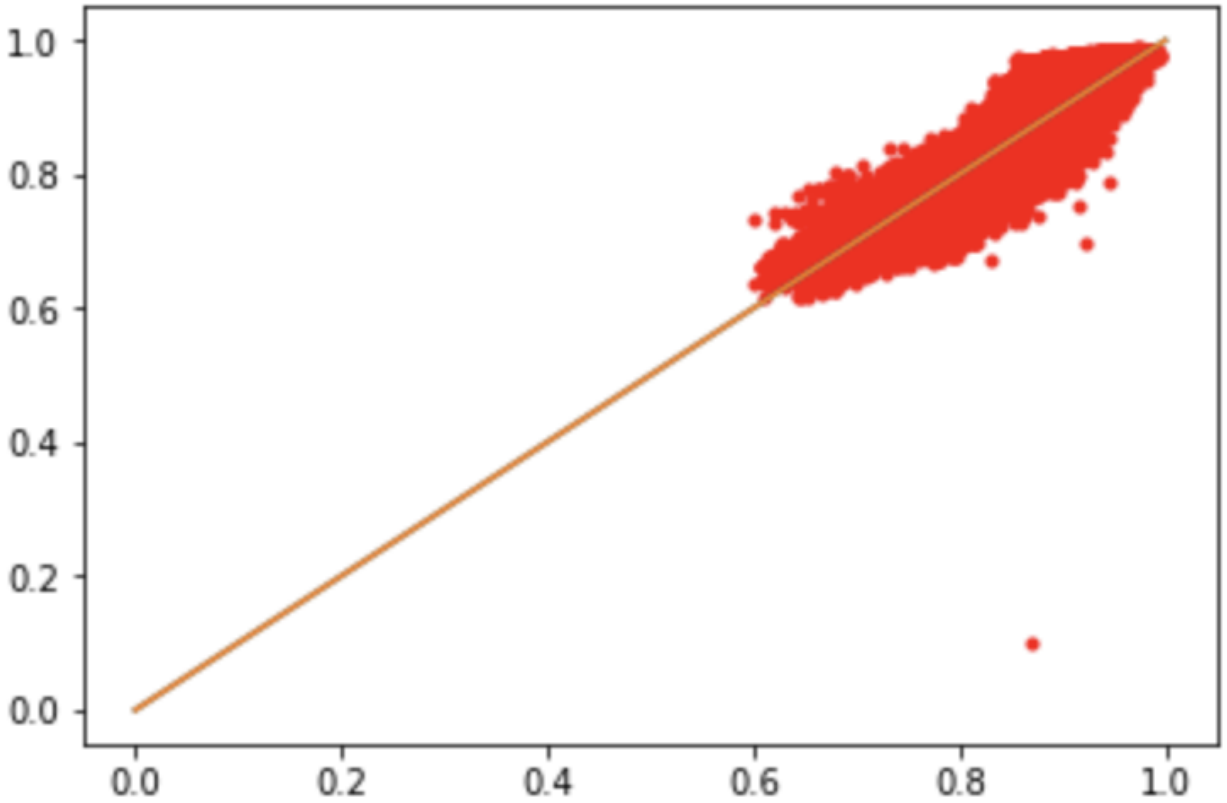
Rationale

- Input vector: Genotype/Phenotype of length 12 (packaging signal strengths in 3 bands)
- Output vector: Assembly efficiency (out of 2000 possible capsids)
- Black box: Machine learning via a neural network

Machine Learning with a Neural Network

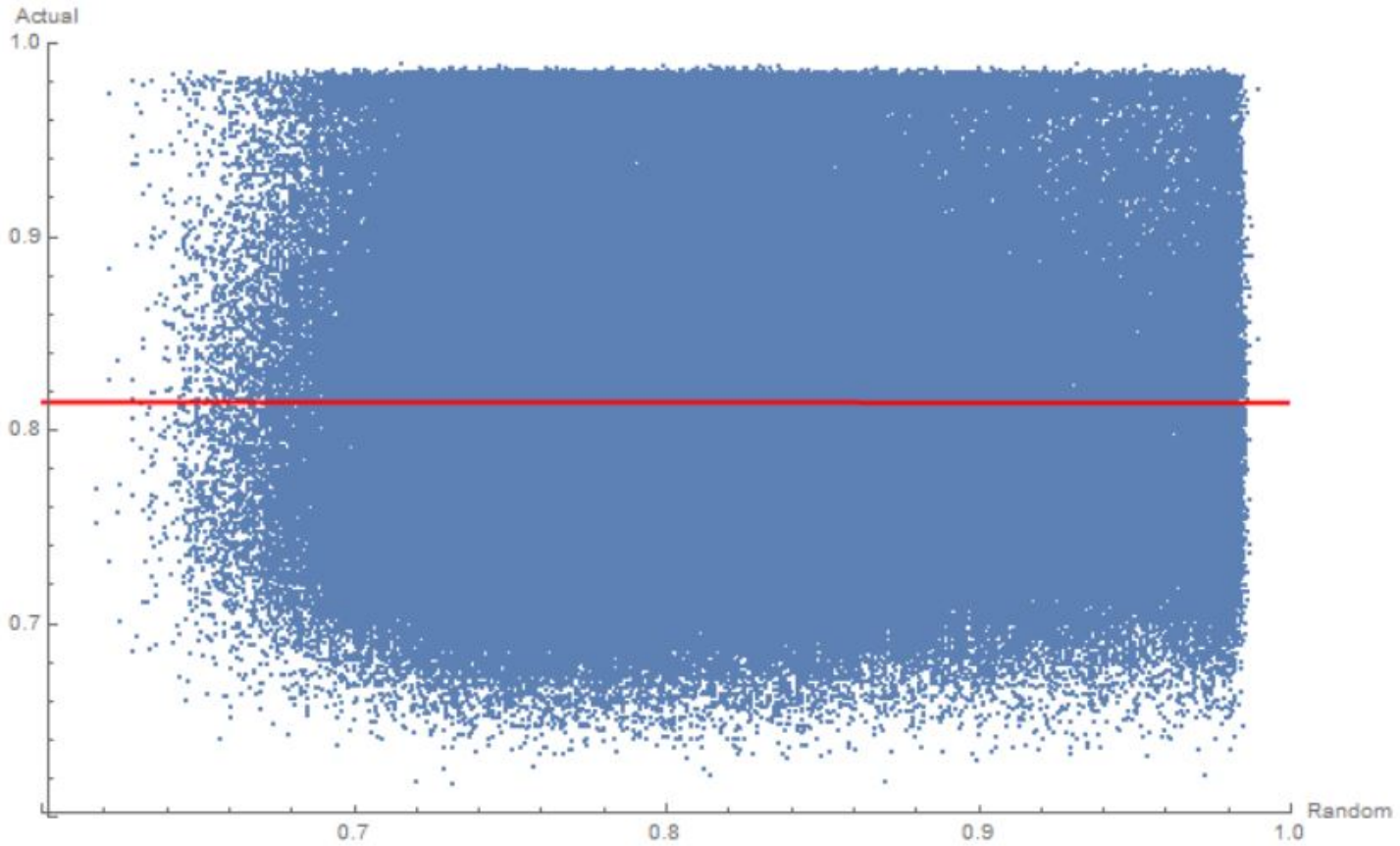


Predictions



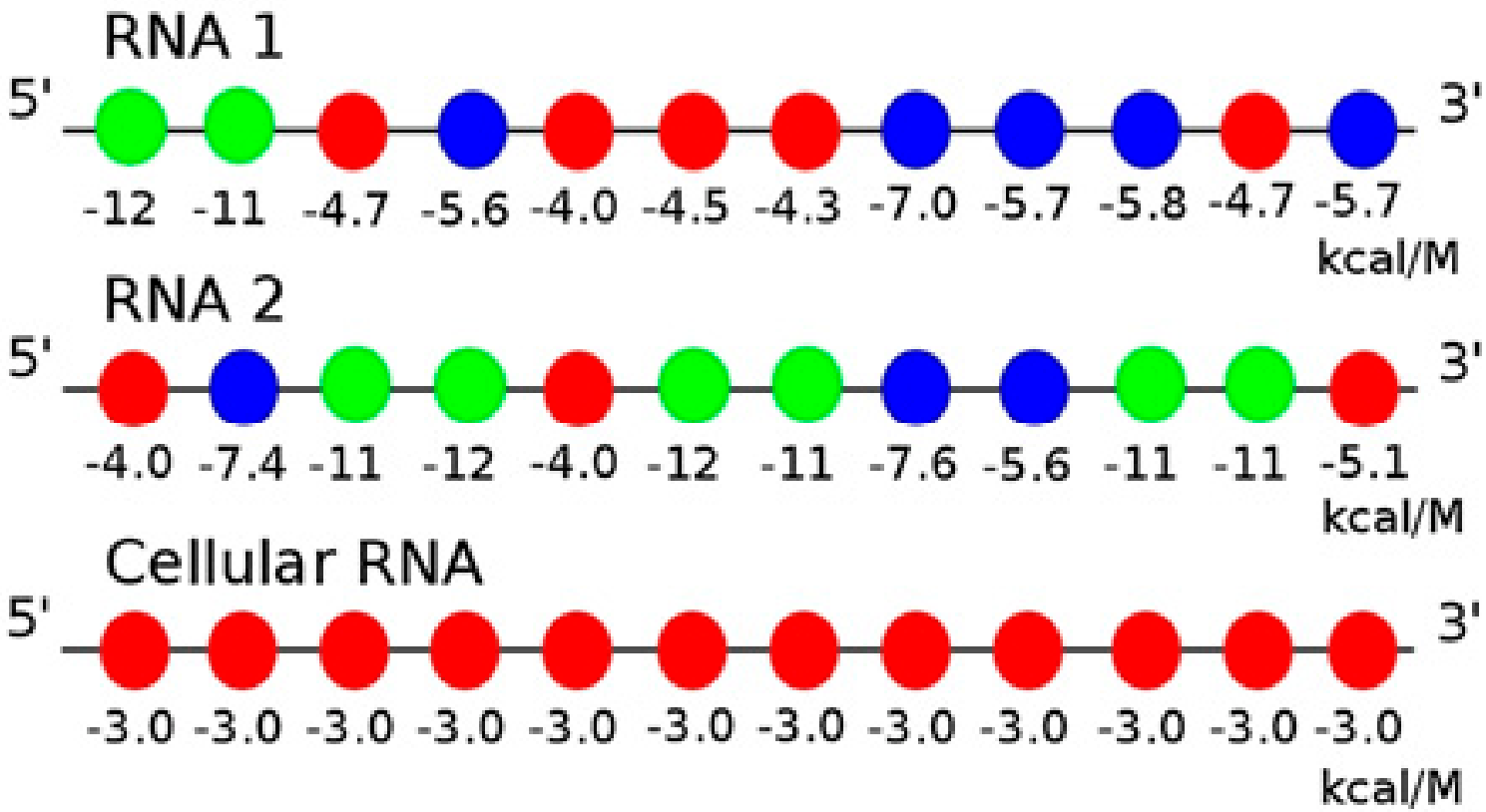
predicted vs actual value of assembly efficiency

Predictions



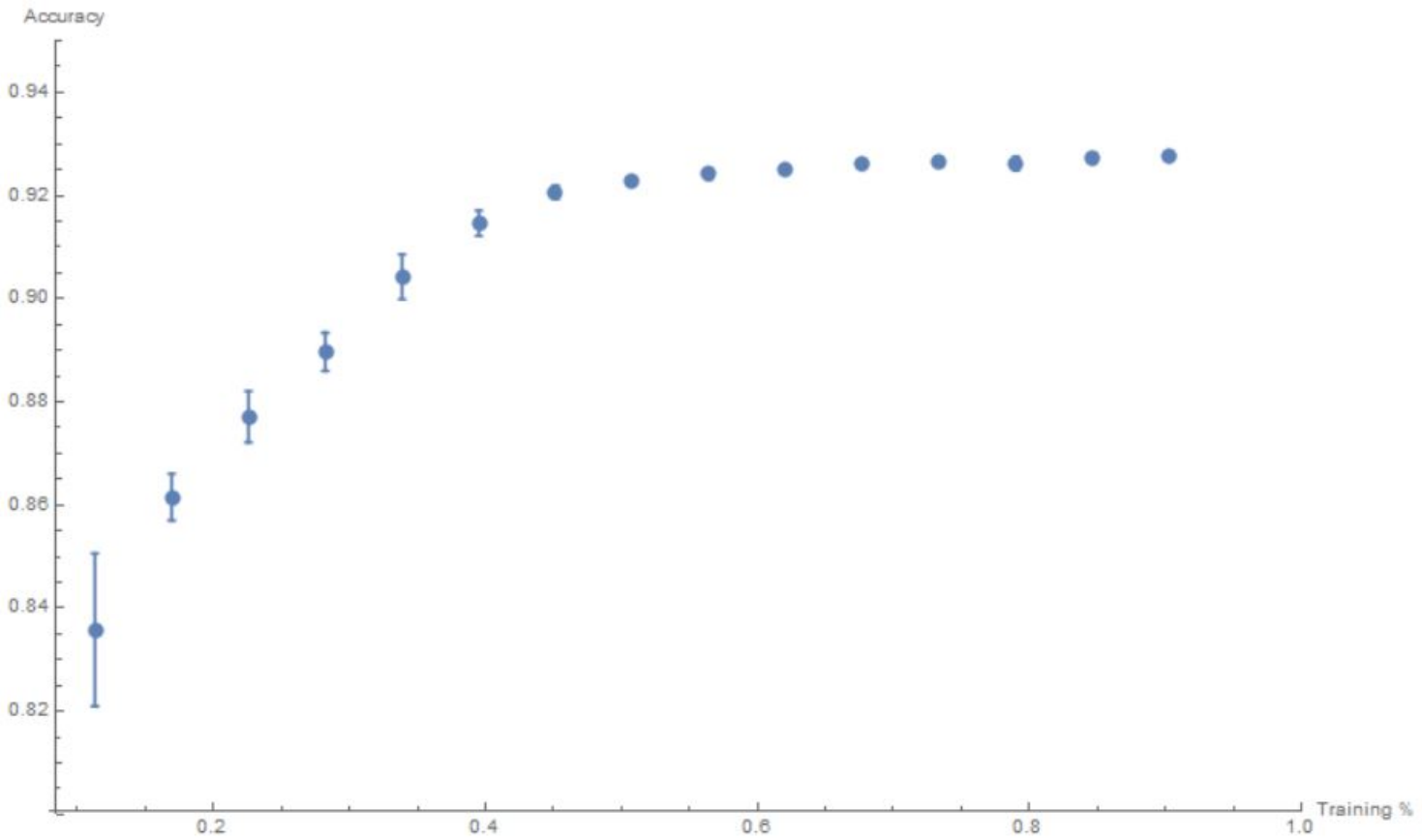
vs random assignments of assembly efficiency

Not just random, intrinsic features?

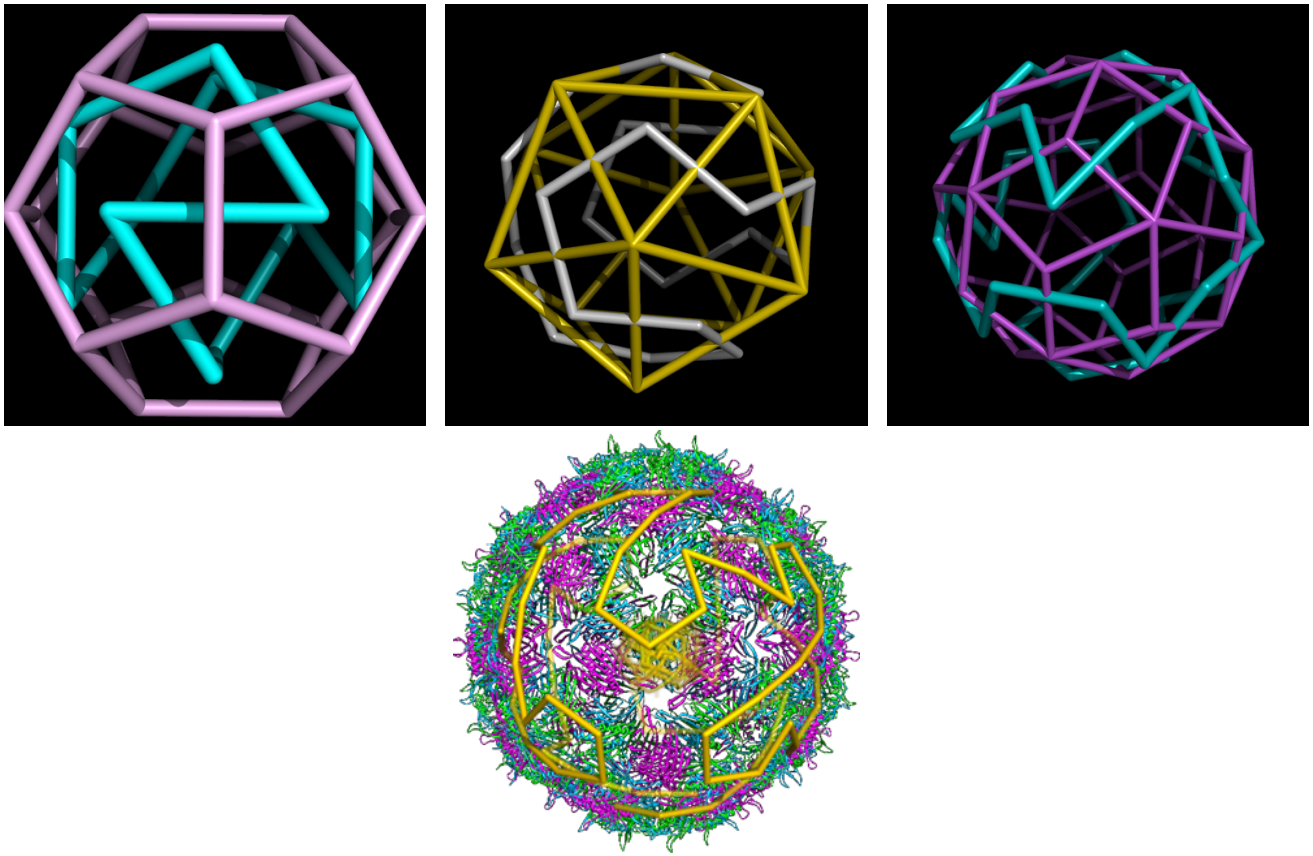


Definite starting point with strong binding, then weaker binding in an error-correcting bit, driven to completion by thermodynamics

Learning Curve



Conclusions



Do more realistic models in future – geometry, binding gradation.
Partially explore the landscape and predict the rest (procedurally)?

Thank you!

Machine-learning a virus assembly fitness landscape
P-P Dechant, Y-H He, arXiv preprint arXiv:1901.05051, 2019