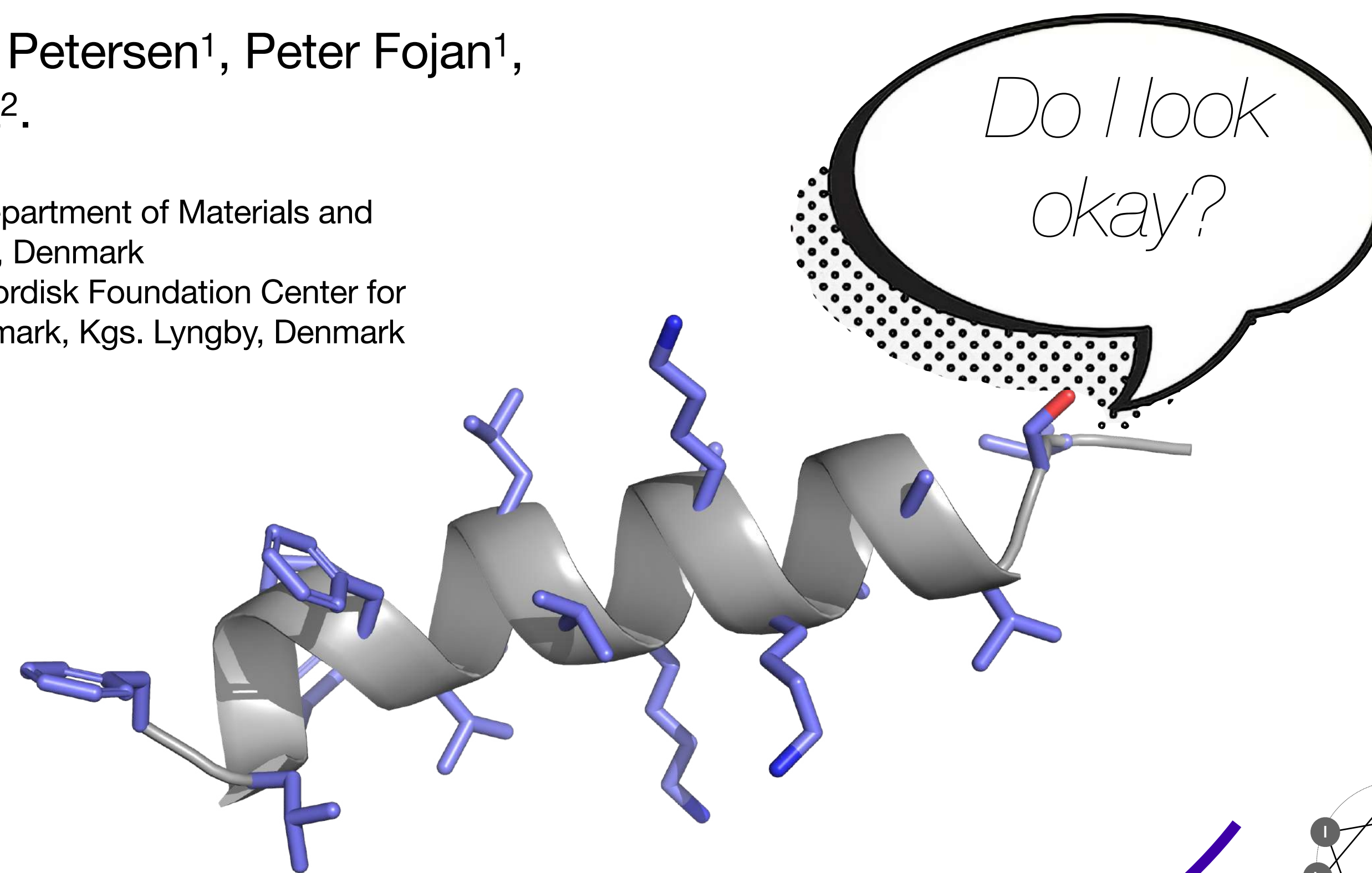
Physicochemical symmetries restrict AI/DL success in predicting antimicrobial peptide activity: Breaking permutation invariance with geometric deep learning.

Niklas G. Madsen[1,2], Evamaria Petersen[1], Peter Fojan[1], and Carlos G. Acevedo-Rocha[2].

Correspondence: nikma@dtu.dk
[1]Material Science and Engineering Group, Department of Materials and Production, Aalborg University, 9000 Aalborg, Denmark
[2]Computational Protein Engineering, Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark

Abstract (short):
Antimicrobial peptides (AMPs) remain a staple in last-resort treatment against antibiotic resistant organisms, yet state-of-the-art computational methods result in low success rates in vivo. We computationally investigate which numerical representation of amino acid sequences correlate with antimicrobial activity. It is shown that state-of-the-art methods can not discriminate a sequence from its shuffled permutation. Naturally, a shuffled amino acid sequence leads to differential activity in vivo. This failure mode is necessarily the case, as most physicochemical descriptors are permutation invariant, making the task of classifying shuffled sequences impossible. We stress the importance of careful embeddings and their associated symmetries when using AI/DL for biological tasks. We develop a geometric deep learning method to overcome permutation invariance and predict activity from sequence.



GVSVAGAKKVKVLFVFPFLF
MIC >256 ug/mL†

a) *Introduction*: *Antimicrobial Peptides (AMPs) as a necessary tool to battle Antimicrobial Resistance.*

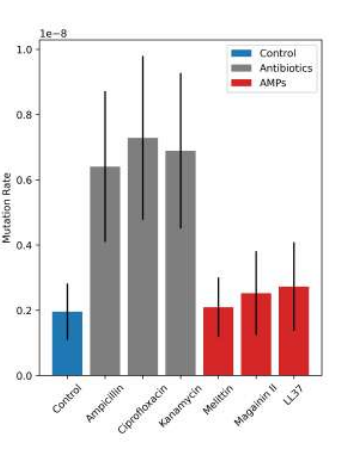


Figure 1.1: Mutation rate for *E. coli* cultures treated at MIC50 for four hours as determined using the Luria-Delbrück fluctuation test. Conventional antibiotics (ampicillin, ciprofloxacin, and kanamycin) significantly increase mutation rates compared to the control. Adapted from [23]. The article from which the figure is reproduced is an open-access article and as such, is distributed under a CC BY 4.0 license.

i) AMPs do not significantly alter the mutation rate. (emergence of resistance)
ii) AMPs target the membrane (primarily), which is less prone to resistance.
iii) AMPs are cationic and structurally amphiphilic.
iv) Canonical AMPs are natural and everywhere! Over 40 are expressed in your mouth [1]. They are the platform which organisms use for host-defense.
v) AMPs are currently last-resort drugs against antibiotic resistance microbes.
vi) Discovering new AMPs is crucial

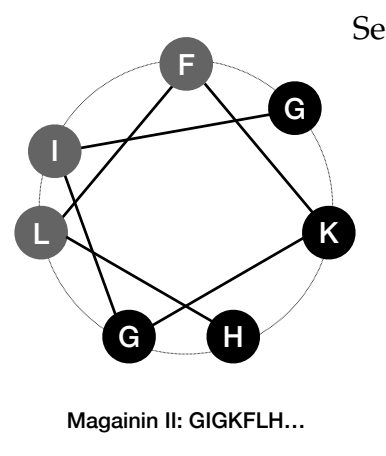


Figure 1.4: Helical wheel representation of the first seven residues of Magainin II. Facial amphiphilicity is well illustrated by the segregation of polar and apolar residues.

Table 2.2: AMPs frequently have intracellular targets, a few examples of which are shown in the table. The table is adapted from [2]. Sequences around twenty in length were selected to illustrate the diverse modes of action accessible with twenty residues.

AMP Name	Sequence	Mode(s) of Action
Buforin II	TRSSRAGLOFPVGRVHRLLRK	Inhibits DNA, inhibits RNA
Microcin J25	VGIGTPIFSYGGGAGHVPEYF	Inhibits RNA polymerase
Pyrrhocoricin	VDKGSYLPRPTPPRPIYNRN	Inhibits DnaK and GroEL, binds LPS
Mersacidin	CTETLPGGGGVCTLTSECIC	Inhibits lipid II in peptidoglycan biosynthesis
Magainin I	GIGKFLHSAGKFGKAFVGEIMKS	Inhibits energy metabolism proteins
Melittin	GIGAVLKVLTTGLPALISWIKRKRQQ	Pore-formation and membrane permeabilisation

a) *Observation: using State-of-the-art methods results in a zero-success rate* in vitro….

Table 5.1: Overview of investigated peptides and summary of results obtained.

Acronym	Sequence	Discovery Method	Results Summary
CFT_cons	FLGKVLKKASKVVKAVFKKV	Consensus sequence as a baseline	non-haemolytic, AMP (28.8 μM).
C4K	TLFKRIKGQRVCVWVHTKSV	Random walk, cross-filtering against haemolysis	non-haemolytic, non-AMP.
KAKCP	KAKFFFACPGCAFFFKAK	Rationally designed cationic self-assembling peptide from an old project	strongly haemolytic, N/A.

….
…
…

Pulling everything out of the peptide freezer and characterising the minimum inhibitory concentration (MIC, lower is better) comparing with predictions from published methods. I wanted to do an experimental MSc. but realised this made little sense as the success rate was zero for me.

b) *Attempt: Trying to understand the impact of descriptor on prediction accuracy?*

Table 6.1: Comparison of state-of-the-art methods for the recognition of AMPs. Acronyms: support vector machine (SVM), artificial neural network (ANN), discriminate analysis (DA), random forest (RF), fuzzy K-nearest neighbor (FKNN), convolutional neural network (CNN), long short-term memory (LSTM). Values obtained from [76] Table 2 which were based on the Veltri et al. benchmark [74]. The largest value of each column is marked in bold.

State-of-the-art	Descriptor	Sn(%)	Sp(%)	ACC(%)	MCC	AUC(%)
AntiBP2 (SVM)	Amino acid composition	87.91	90.80	89.37	0.7876	89.36
CAMPr3-ANN	Unclear: "sixty-four best peptide descriptors"	83.00	85.11	84.05	0.6813	84.05
CAMPr3-DA	Unclear: "sixty-four best peptide descriptors"	87.07	80.75	83.91	0.6797	89.97
CAMPr3-RF	Unclear: "sixty-four best peptide descriptors"	92.69	82.44	87.57	0.7553	93.63
CAMPr3-SVM	Unclear: "sixty-four best peptide descriptors"	88.62	80.47	84.55	0.6933	90.62
iAMP-2L (FKNN)	Pseudo amino acid composition & physiocochemical	83.99	85.86	84.90	0.6983	84.90
iAMPpred (SVM)	Pseudo amino acid composition & physiocochemical structural propensity	89.33	87.22	88.27	0.7656	94.44
gkmSVM	Gapped k-mer amino acid composition	88.34	90.59	89.46	0.7895	94.98
AMPScanner (CNN + LSTM)	Amino acid encoding	89.88	92.69	91.29	0.8261	96.30
ACEP (Three-track CNN + LSTM + Attention)	Amino acid composition, amino acid one-hot encoding, position-specific scoring matrices (PSSM)	92.41	**93.67**	**93.04**	**0.8610**	**97.78**
Tests (internal)						
SVM (linear kernel)	Amino acid composition	87.57	88.14	87.85	0.7571	94.95
	1-gap dipeptide composition	82.77	90.96	86.86	0.7398	94.32
	3-gap dipeptide composition	83.90	93.22	88.56	0.7746	95.06
	4-gap dipeptide composition	82.77	93.50	88.14	0.7671	95.06
	Tripeptide composition	17.80	100	58.90	0.3125	91.95
	Physiocochemical	87.29	88.14	87.71	0.7543	58.90
MLP (ReLU, 4 hidden layers, Adam)	Amino acid composition	**92.66**	85.88	89.27	0.7871	95.80
	1-gap dipeptide composition	84.46	90.68	87.57	0.7529	93.69
	3-gap dipeptide composition	88.98	83.62	86.30	0.7270	94.32
	4-gap dipeptide composition	91.24	86.16	88.70	0.7750	93.95
	Tripeptide composition	83.05	83.33	83.19	0.6638	88.03
	Physiocochemical	87.29	88.42	87.85	0.7571	95.18
Huggingface (RF)	Concatenated compositional features (AAC, 4-gap DPC, PCP)	88.98	90.40	89.69	0.7939	89.69

i) SOTA methods have flatlined and it is often unclear how sequence is encoded. Badly reported methods.
ii) Why does composition alone fare so well?
iii) Why did none of the predicted peptides work if test metrics are so high?

a) *Observation: All published methods seem to fail at shuffled peptides.*

Method	Descriptor	ACC(%)
MLP (ReLU, 4 hidden layers, Adam)	Amino acid composition	53.12
Huggingface (RF)	Concatenated compositional features (AAC, 4-gap DPC, PCP)	50.
AMPScanner V2.0 (CNN + LSTM)	Amino acid encoding	46.87

Finding purposely shuffled sequences in literature [2]

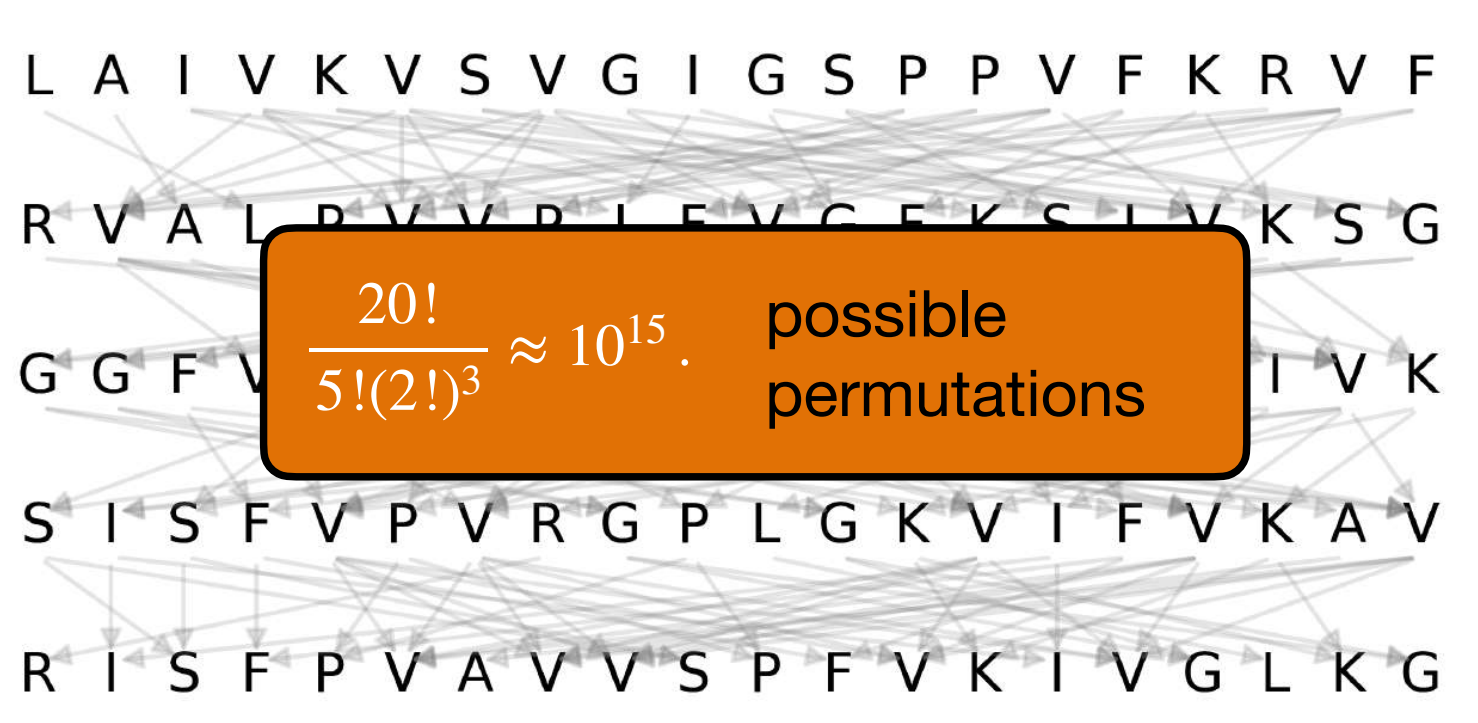
Inactive, shuffled
GVSVAGAKKVKVLFVFPFLF

Active
FLGVVFKLASKVFPAVFGKV

b) *Symmetries of physicochemical representations and their implications.*

i) Most physicochemical features are global sequence averages and thus many sequences encode the same feature.
ii) Proposition: a ML/DL framework based on these cannot possibly learn to discriminate permutations.

All 10^{15} unique permutations in sequence result in the same physiocochemical representation.

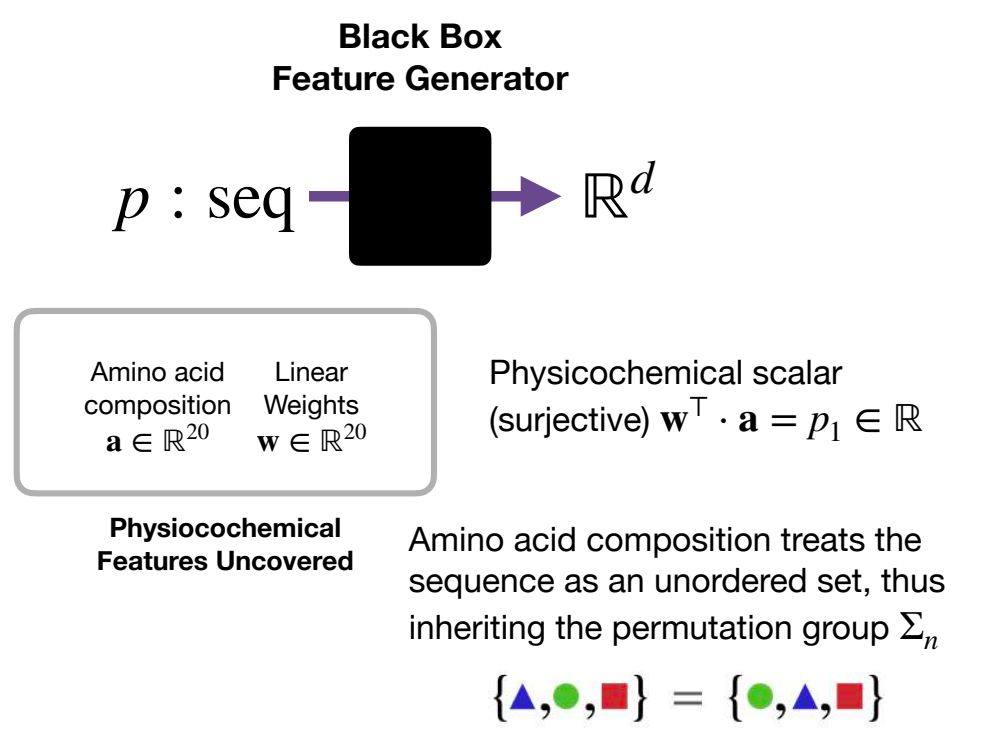

L A I V K V S V G I G S P P V F K R V F
R V A L … K S G
G G F V … I V K
S I S F V P V R G P L G K V I F V K A V
R I S F P V A V V S P F V K I V G L K G

$\frac{20!}{5!(2!)^3} \approx 10^{15}$. possible permutations

pI	Mw
11.17	2113.62
11.17	2113.62
…	…
11.17	2113.62

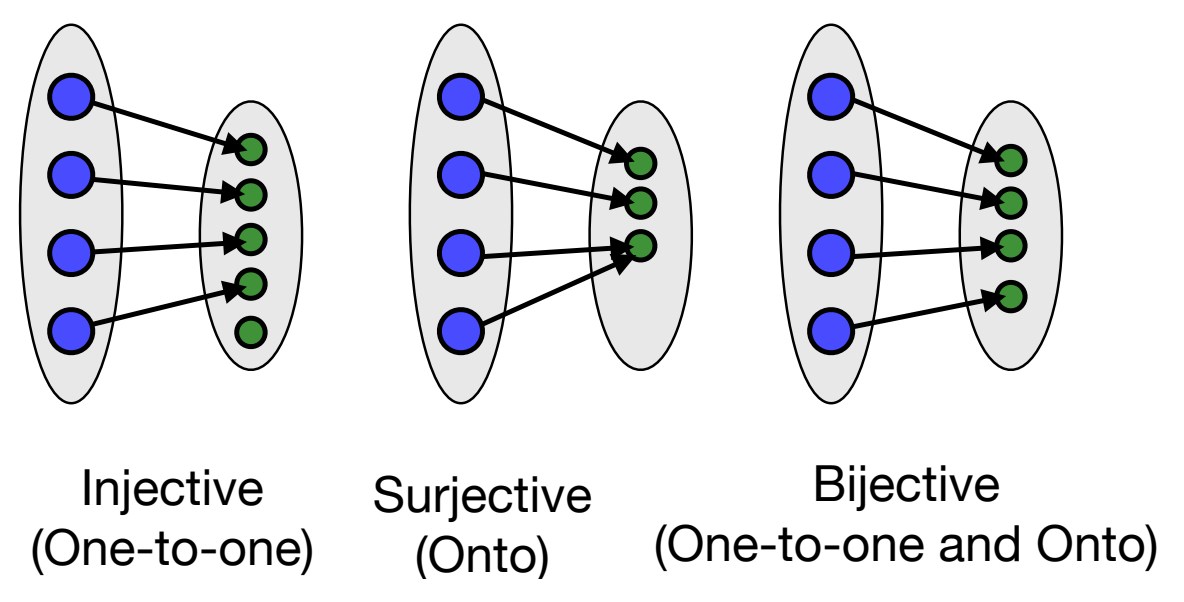
c) *Why do methods lack robustness? Group theoretic answers.*

i) Most features are surjective, thus degeneracy arises.
ii) Amino acid composition is an orthogonal basis set, from which most physiocochemical descriptors can be derived.
iii) An MLP can provably learn all these descriptors, one-per-neuron.


Black Box
Feature Generator
p : seq
$\mathbb{R}^d$
Amino acid composition $\mathbf{a} \in \mathbb{R}^{20}$
Linear Weights $\mathbf{w} \in \mathbb{R}^{20}$
Physiocochemical Features Uncovered
Physicochemical scalar (surjective) $\mathbf{w}^\top \cdot \mathbf{a} = p_1 \in \mathbb{R}$

Amino acid composition treats the sequence as an unordered set, thus inheriting the permutation group Σ_n

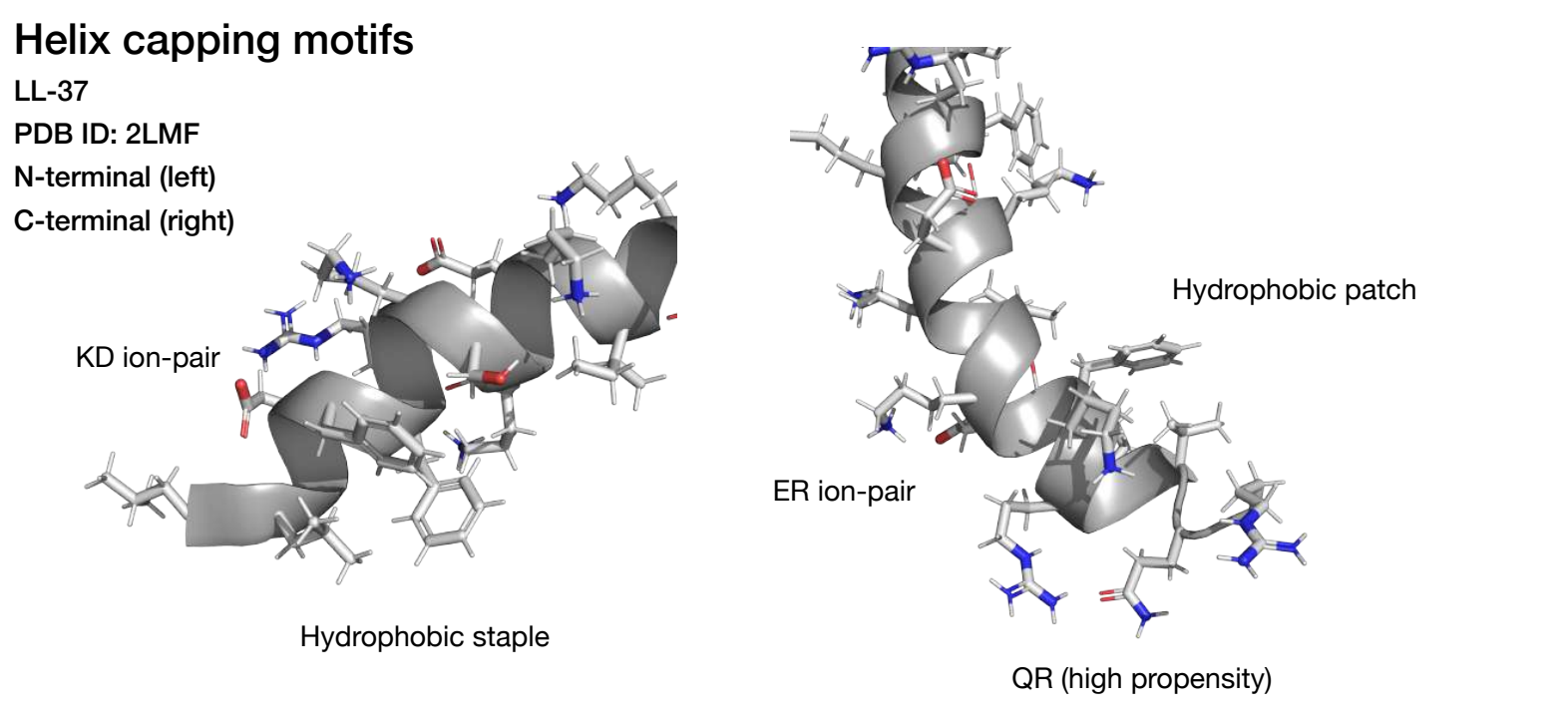
{▲,●,■} = {●,▲,■}



a) *What explains why AMPs are not permutation invariant, unlike physicochemical descriptors commonly used.*

Table 5.4: Helix capping motifs in common α-helical AMPs. These have an NMR-resolved structure, which allows for an analysis of intra-chain motifs.

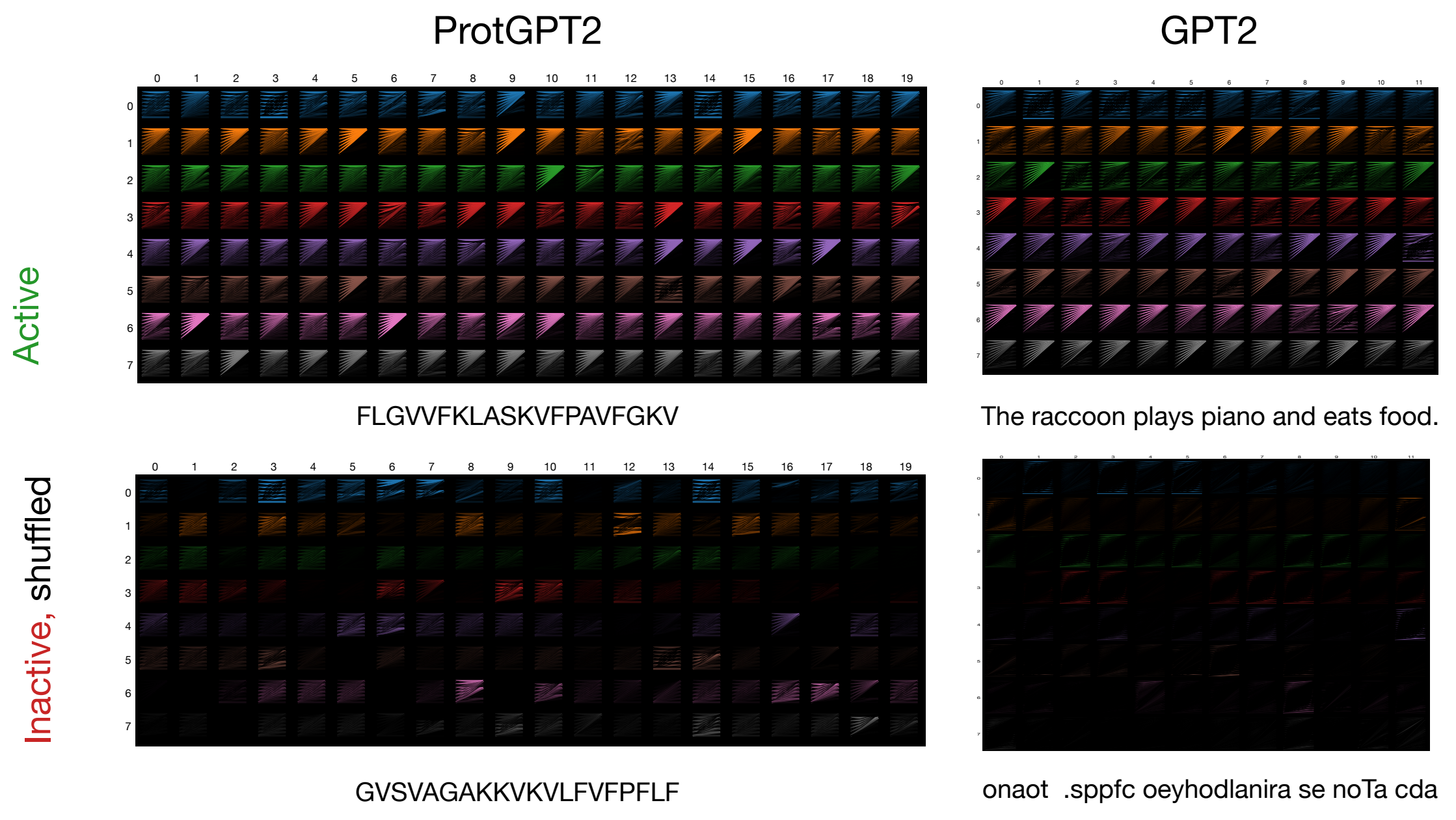
	Melittin	LL-37	Brevinin-1BYa
Sequence	GIGAVLKVLTTGLPALISWIKRKRQQ	LLGDFFRKSKEKIGKEFKRIVQR	FLPILASLAAKFGPKLFCLVTKKC
N-terminal Motif	G (high propensity)	Hydrophobic staple	P (high propensity)
C-terminal Motif	KRQ (high propensity)	QR (high propensity)	KK (high propensity)
Stabilising pair-wise	I2L6, V5L9, W19R22	E11K15,K12E16,E16R19	F1L5,I4L8, F12L16
Notes	P14 causes kink	L2F5F6 form hydrophobic patch	S-S bond at C-terminus
PDB ID	6DST	2LMF	6G4I



i) Simply: biophysics. A permutation in order changes structure. (the obvious)
ii) More subtly, permutations disrupt helix capping motifs that thermodynamically favour adsorption to the membrane.
iii) Permutations in order change the facial amphiphilicity

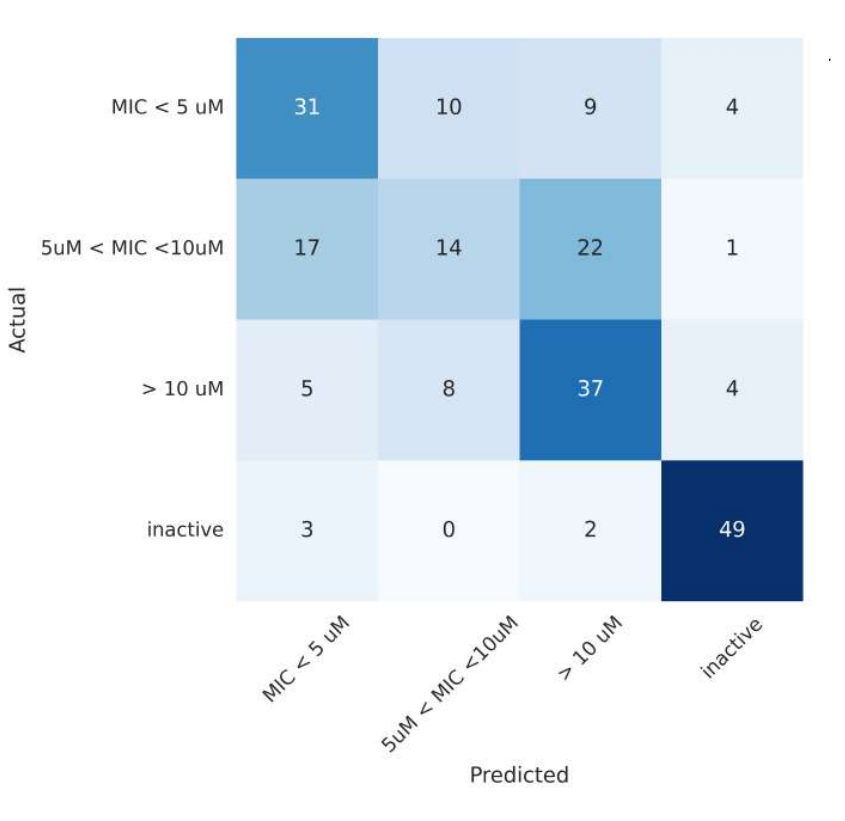
a) *Protein language models break permutation invariance.*

i) Positional encoding!
ii) Attention looks dramatically different for a shuffled-inactive vs. active AMPs.
iii) Perplexity indicates the difference also: 97.9 < 13615
iv) As is done routinely for all other protein related tasks now, language model embeddings are powerful



a) *The beginnings of QSAR, but lacking validation due to failed chemical synthesis of peptide.*

1) Method can predict activity reasonably.
2) Experimental validation of new global-surface method hindered by chemical synthesis.
3) But, MD simulations indicate strong membrane affinity.



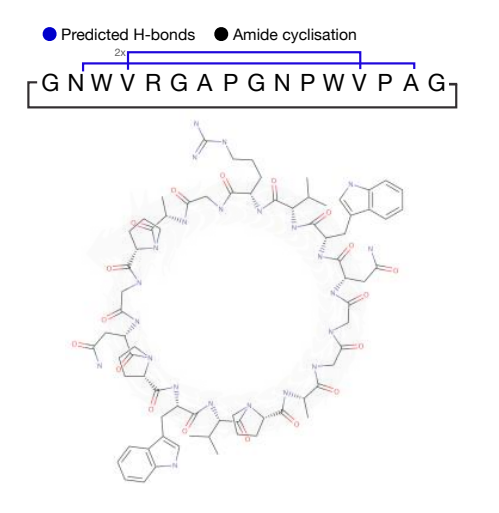
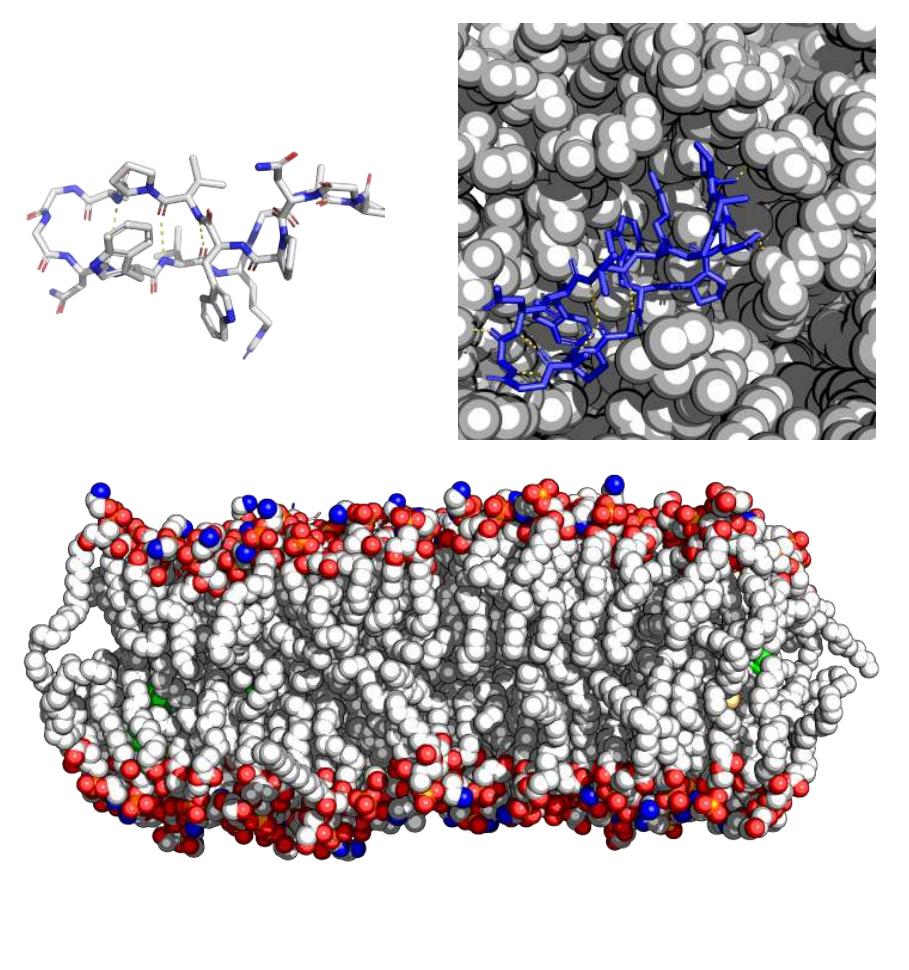
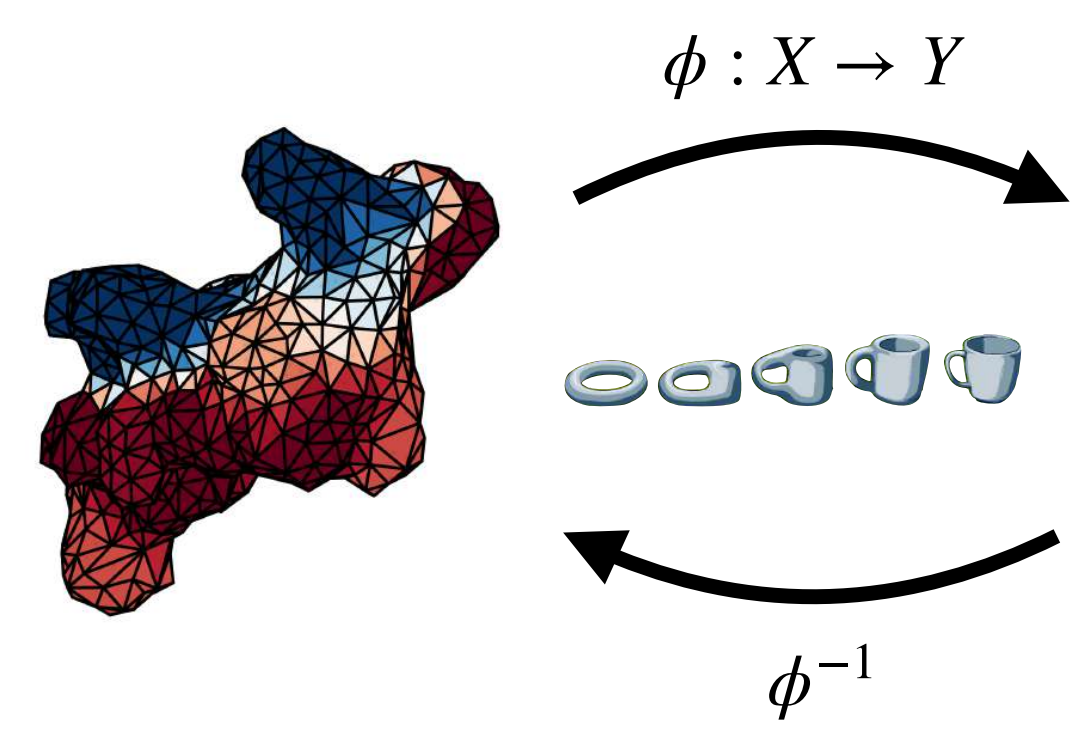
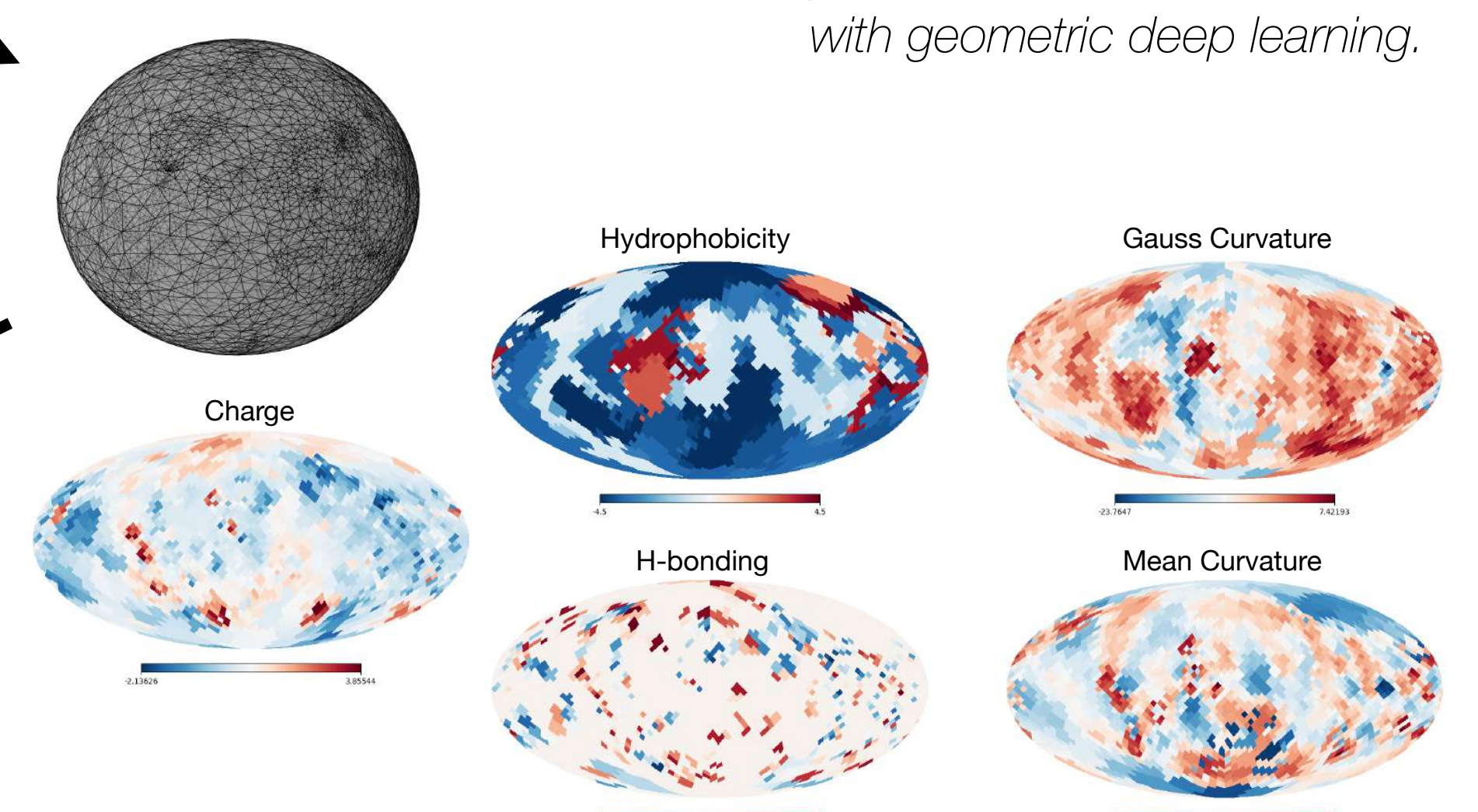


Figure 5.26: Linear representation of fixbb21 showing intramolecular hydrogen bonds and amide cyclisation. The Kekulé form of the molecular structure is visualised below.



a) *Homeomorphisms to enable global surface representations with geometric deep learning.*





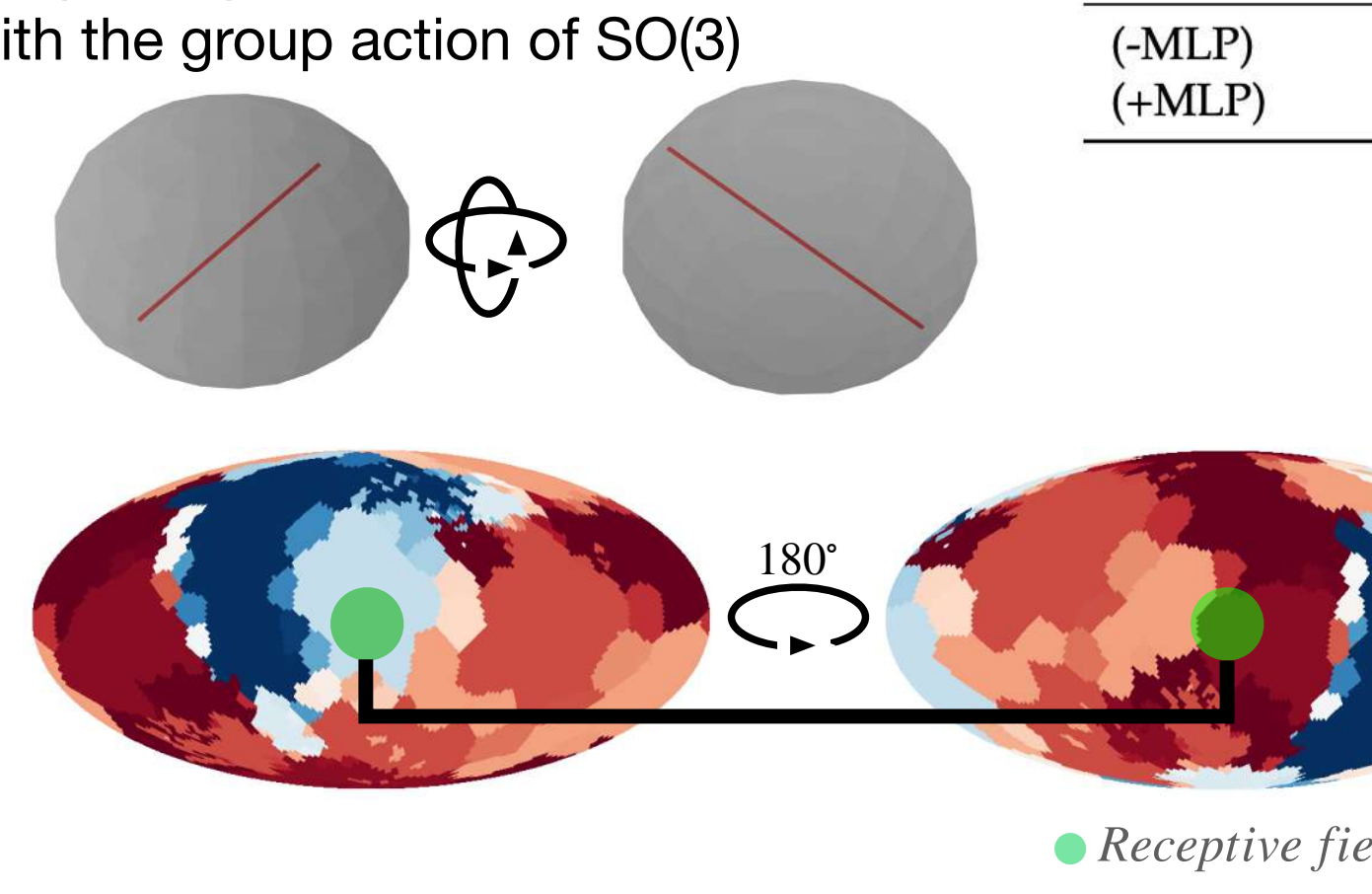
Building on available surface method MaSIF [3] for deep learning on surface representations, but extend upon it to avoid random sampling of surface.

b) $\mathcal{S}^2$ *representation has a brilliant symmetry for antimicrobial peptides.*

Method developed can accurately learn global surface motifs to classify antimicrobial peptides, thus integrating structure. (Using a spherical CNN [4])

	ACC	F1	MCC	AUC-ROC
(-MLP)	0.8305	0.8545	0.7097	0.9232
(+MLP)	0.8573	0.8573	0.7153	0.9419

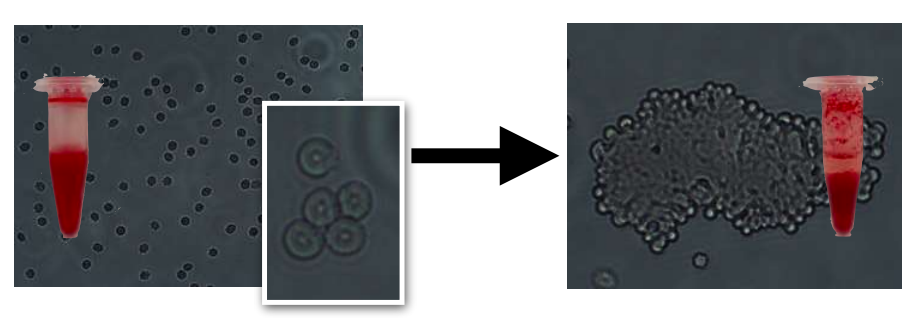
Antipodal points commute with the group action of SO(3)

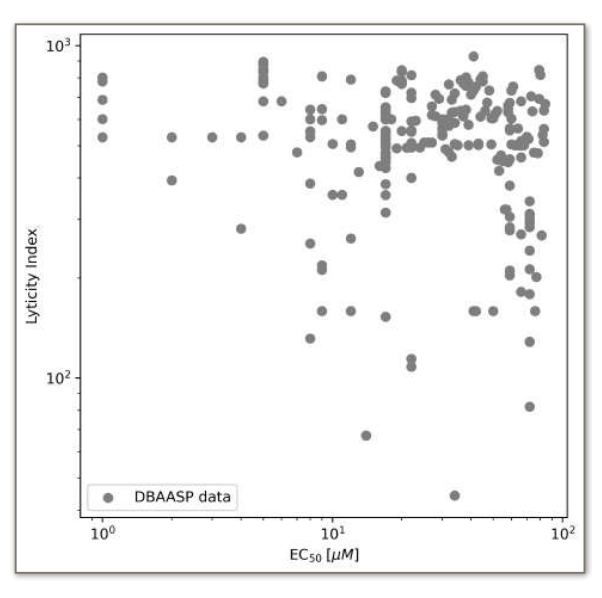


● *Receptive field*

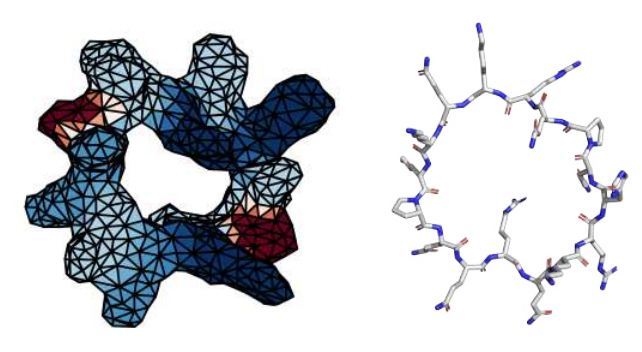
Feature	SO(3)/SO(3)
All features	0.835 ± 0.002
Chemical	0.853 ± 0.007
Geometric	0.656 ± 0.013
Charge	0.832 ± 0.009
Hydrophobicity	0.736 ± 0.014
H-bonding	0.761 ± 0.011
Gauss Curvature	0.650 ± 0.010
Mean Curvature	0.651 ± 0.019

***Extra**: Haemolysis is a critical problem of AMPs, we found that the lyticity index sets a lower bound for the EC50*



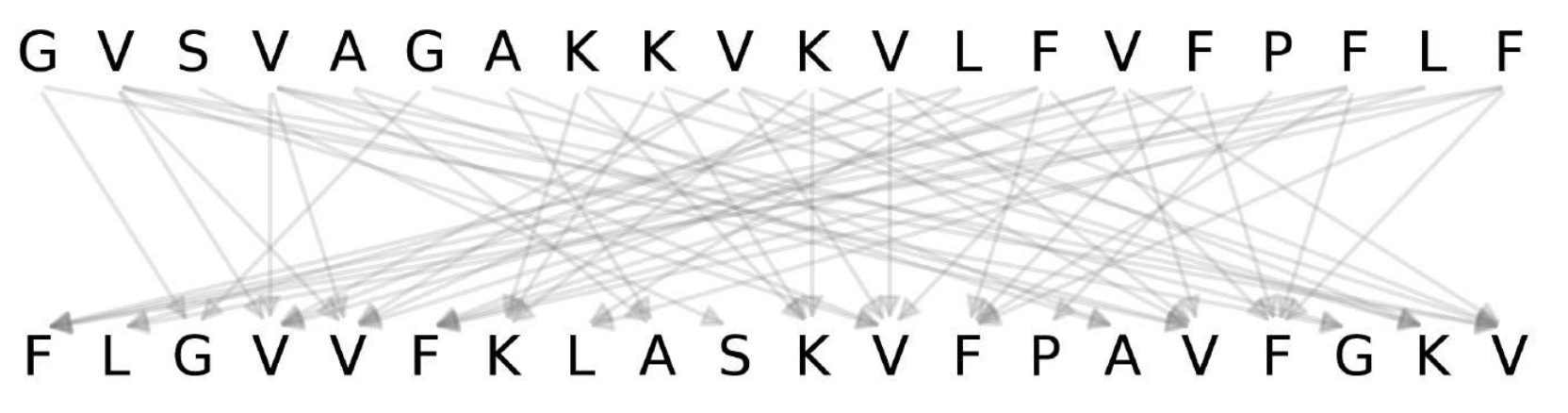


***Extra**: Counter example which the method cannot reliably predict due to genus of surface.*



Cyclic: RHQPQRKQKKPQQRQK
Genus: 1

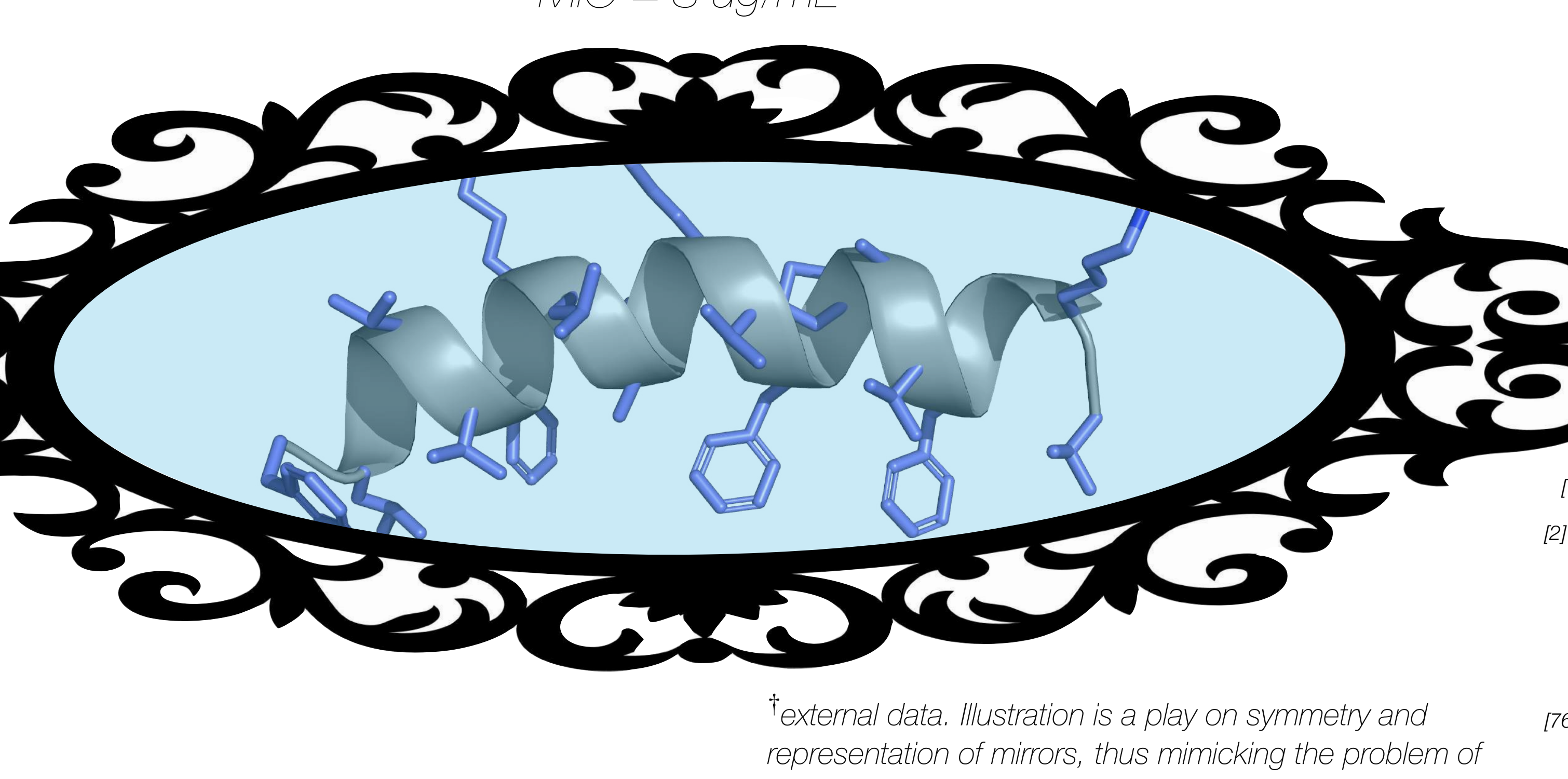
***Extra**: Symmetric group of the set is LARGE and thus only one embedding for all is problematic.*


G V S V A G A K K V K V L F V F P F L F
F L G V V F K L A S K V F P A V F G K V

> 256 ug/mL

8 ug/mL

FLGVVFKLASKVFPAVFGKV
MIC = 8 ug/mL†



†*external data. Illustration is a play on symmetry and representation of mirrors, thus mimicking the problem of determining which sequence permutation 'looks okay'.*

a) Conclusion.

*The need for introspection for **Robust** AI in Biology:*
i) Which symmetries are your methods blind to?
ii) How could you test this? (adversarial)
iii) Is the blind spot, due to symmetry, acceptable for the use case?

Results here:
1) Sequence (linguistic) and structure (geometric) form well-complimented representations.
2) Syntactically 'correct' AMP sequences grow sub-exponentially in sequence space rather than as 20^n.
3) Global surface motif recognition with geometric deep learning and antipodal symmetries.

[1] Neeloffer Mookherjee et al. 'Antimicrobial host defence peptides: functions and clinical potential'. In: Nat Rev Drug Discov 19.5 (2020),
[2] Christopher Loose et al. 'A linguistic model for the rational design of antimicrobial peptides'. en. In: Nature 443.7113 (2006).
[3] Gainza, P., Sverrisson, F., Monti, F. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods **17**, 184–192 (2020).
[4] Nathanaël Perraudin et al. 'DeepSphere: Efficient spherical Convolutional Neural Network with HEALPix sampling for cosmological applications'.
[74] Daniel Veltri, Uday Kamath, and Amarda Shehu. 'Deep learning improves antimicrobial peptide recognition'. In: Bioinformatics 34.16 (Aug. 2018), pp. 2740–2747. doi: 10.1093/bioinformatics/ bty179.
[76] Haoyi Fu et al. 'ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding'. In: BMC Genomics 21.1 (Aug. 2020)