Representing, predicting, and generating simple and complex peptides

Chemoinformatics Strasbourg Summer School 2022

Massina Abderrahmane, Brice Hoffmann, Nicolas Devaux, Maud Jusot

IKTOS

Agenda

1. Peptides in drug design
2. A matter of representation…
3. To predict…
4. And generate new peptides
5. Conclusion

Peptide therapeutics

Around 80 peptide drugs on the global market

More than 150 peptides in clinical development

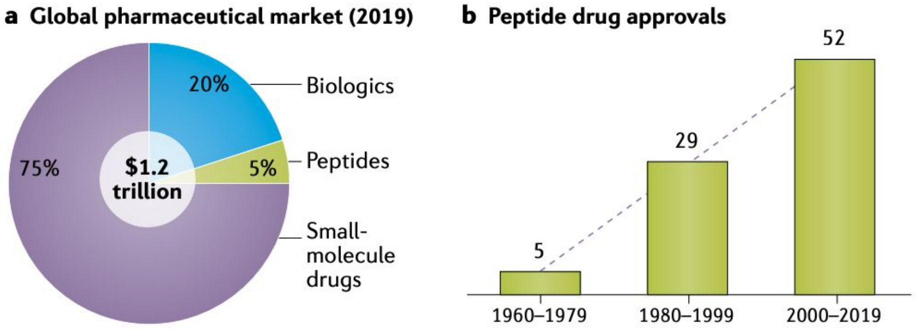
400–600 peptides undergoing preclinical studies

Some limitations:

90% of all peptide drugs are delivered by injection

lack of oral bioavailability remains the major limiting barrier in peptide drug development

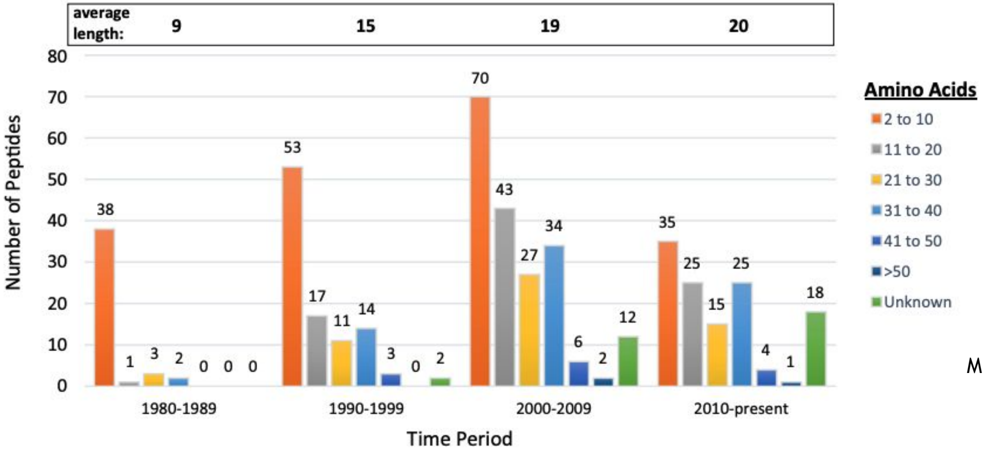
Most peptide drugs modulate peripheral extracellular targets



Muttenthaler, M. et al. Trends in peptide drug discovery. Nat Rev Drug Discov 20, 309–325 (2021)

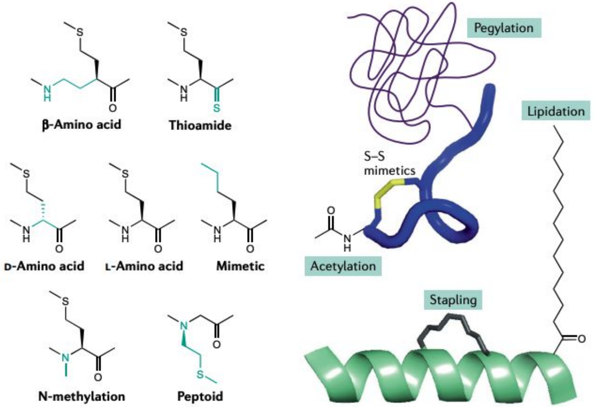
Peptides are heterogeneous by their size and type

Length of peptides entering clinical development, by decade.

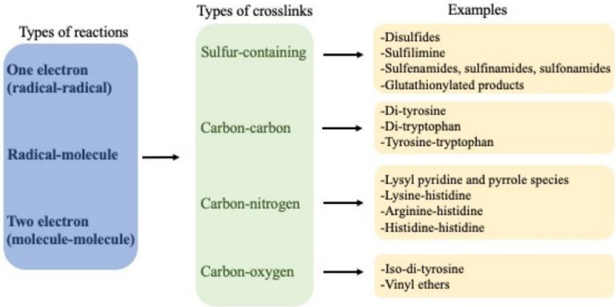


Lau, J. L., & Dunn, M. K. (2018). Bioorganic & medicinal chemistry, 26(10), 2700-2707.

Overview of well-validated chemical modifications used in peptide drug development to increase metabolic stability and bioavailability



Muttenthaler, M. et al. Trends in peptide drug discovery. Nat Rev Drug Discov 20, 309–325 (2021)



Types of reactions

One electron (radical-radical)

Radical-molecule

Two electron (molecule-molecule)

Types of crosslinks

Sulfur-containing

Carbon-carbon

Carbon-nitrogen

Carbon-oxygen

Examples

-Disulfides
-Sulfilimine
-Sulfenamides, sulfinamides, sulfonamides
-Glutathionylated products

-Di-tyrosine
-Di-tryptophan
-Tyrosine-tryptophan

-Lysyl pyridine and pyrrole species
-Lysine-histidine
-Arginine-histidine
-Histidine-histidine

-Iso-di-tyrosine
-Vinyl ethers

Fuentes-Lemus, E., et. al. (2021). Molecules, 27(1), 15.

Objective

Build and evaluate **predictive** and **generative** models for peptides

Take into account **complex peptides**, including modified amino-acids, crosslink, linkers, terminal modifications, etc…

Agenda

1. Peptides in drug design
2. **A matter of representation…**
3. To predict…
4. And generate new peptides
5. Conclusion

Classical representations for small molecules

Many **machine learning models** for small molecules rely on vectorial representations. Two categories have been heavily used:

- physical-chemical descriptors (logP, TPSA, HBA, HBD, MW, etc…)
- molecular fingerprints

ECFP / Morgan fingerprints are a way to represent molecules as **mathematical objects.** They are computed from the atomic representation of molecules.

Starting From **the atomic graph of molecules**, the algorithms takes place in two main steps :

- Initial integer identifier to each non-hydrogen atom (invariant) of the input molecule
- A number of iterations are performed to combine the initial atom identifiers with identifiers of neighboring atoms until a specified diameter is reached

Typical atom invariants :

- atomic number
- number of "heavy" (non-hydrogen) neighbor atoms
- number of attached hydrogens (both implicit and explicit)
- formal charge
- additional property that indicates whether the atom is part of at least one ring

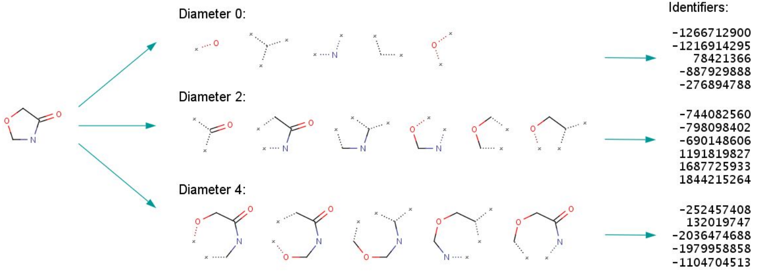


Figure from:
https://docs.chemaxon.com/display/docs/extended-connectivity-fingerprint-ecfp.md#src-1806333-extendedconnectivityfingerprintecfp-introduction

Graph representation of peptides

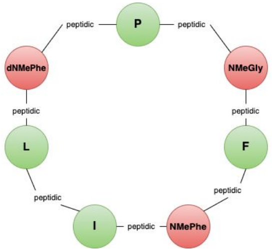
We introduced new graph representation of peptides at the different levels

1- Simple Peptide Graph

It is the basic graph representation for peptides and the most intuitive.

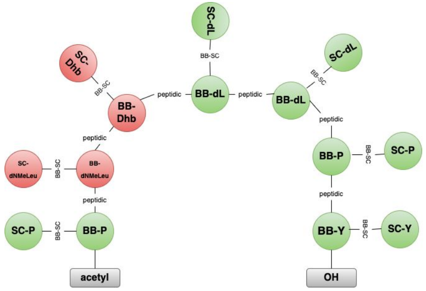
Each node of the graph corresponds to an amino-acid.

Can deal with **natural** and **modified** amino acid, cyclic, **crosslinks**, **linkers**, **terminal modifications**, etc…



2- BBSC Peptide graph (backbone and side chain)

In this representation, each amino acid node is splitted into a backbone node and a side-chain node.



Detection of amino-acids is made using Proteax (PLN format)

These graphs are then converted in a vectorial representation using the Morgan algorithm

A matter of representation

Definition of invariants for peptides

We tested two different invariants to represent each node in the graph:

- **Amino acid names (tokens)**

P	Y	NMeAla	Dhb
0	0	1	0

- **Amino acid descriptors**

Given a list of descriptors with their thresholds, descriptor values are computed on each node then binned into intervals. (Descriptors and Number of intervals depends on user given input)

mw	rb	tpsa	logp	charge
1	1	2	3	1

Peptide fingerprints

Each type of graph combined with invariant, and using morgan fingerprints algorithm, we could build **4 different Peptide fingerprints** representations computed on peptide graph.

Representation name	**Type of peptide graph**	**Type of node attributes**
AA_tokens	SIMPLE	Tokens
AA_descriptors	SIMPLE	Descriptors (Different List of descriptors and thresholds)
BB-SC_tokens	BB-SC	Tokens
BB-SC_descriptors	BB-SC	Descriptors (Different lists of descriptors and thresholds)

Next step: compare these representations with morgan on atomic level and with molecular descriptors for classification tasks

Agenda

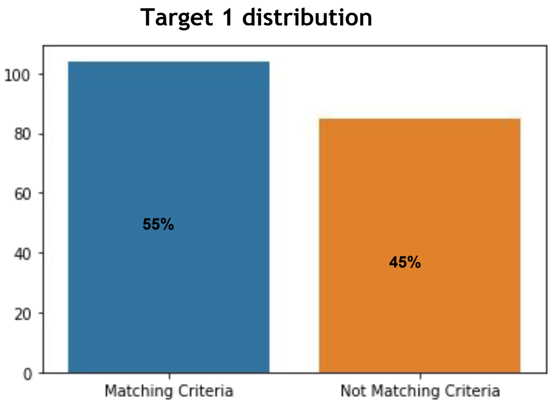
1. Peptides in drug design
2. A matter of representation…
3. To predict…
4. And generate new peptides
5. Conclusion

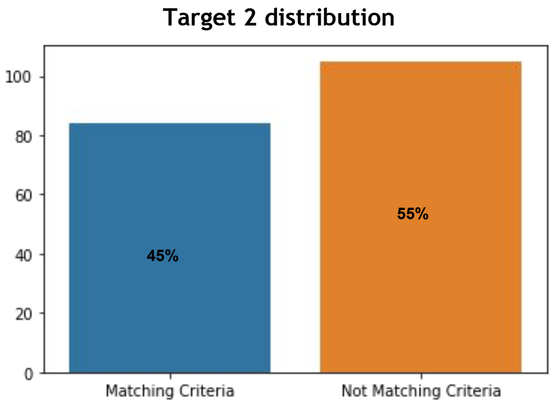
Dataset 1

We worked in collaboration with a pharmaceutical company on predicting activity of a series of peptides on two targets.

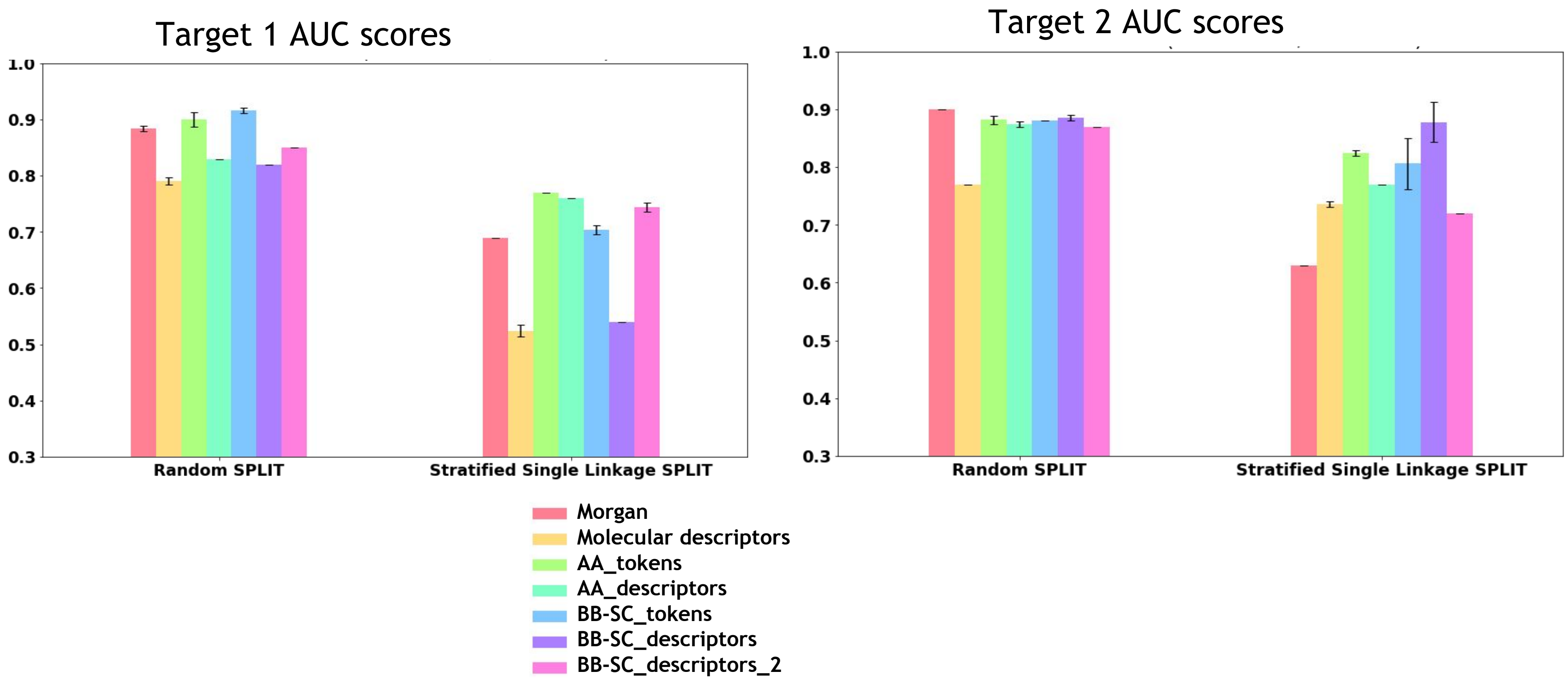
Objective was to be able to generate peptides achieving **activity on target1** and **selectivity on target 2**.

Given dataset was composed of **189 small linear peptides** with their measured target1 and target2 PIC50. Peptides of the dataset include **modified amino acids** and **other specific** components used to enhance peptide stability and permeability.





Prediction results (Random Forest)

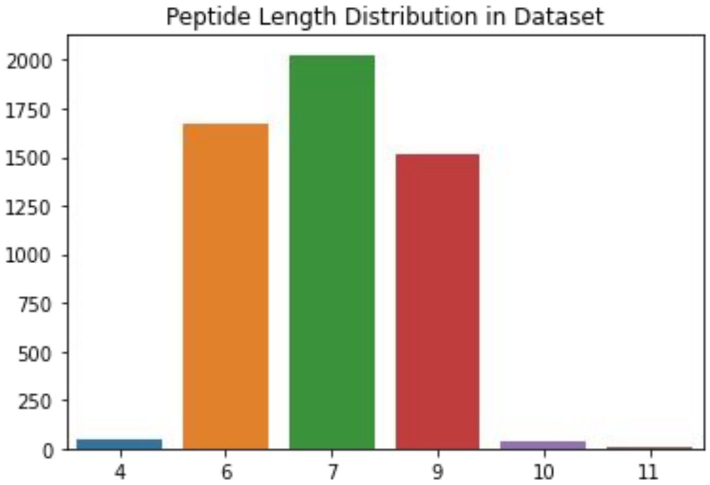


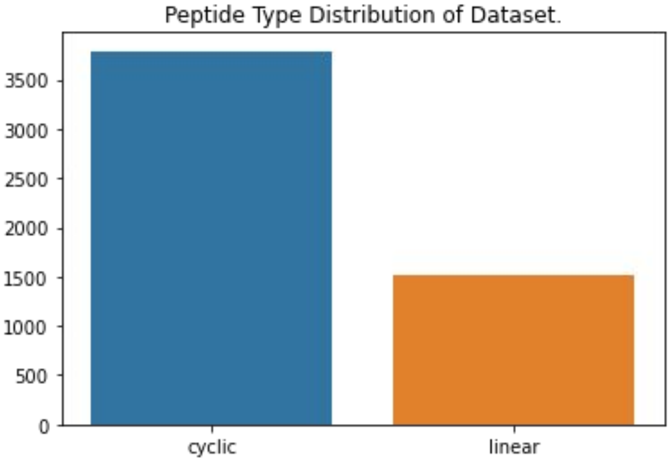
Dataset 2

We worked in collaboration with a second pharma company on **predicting peptides permeability.**

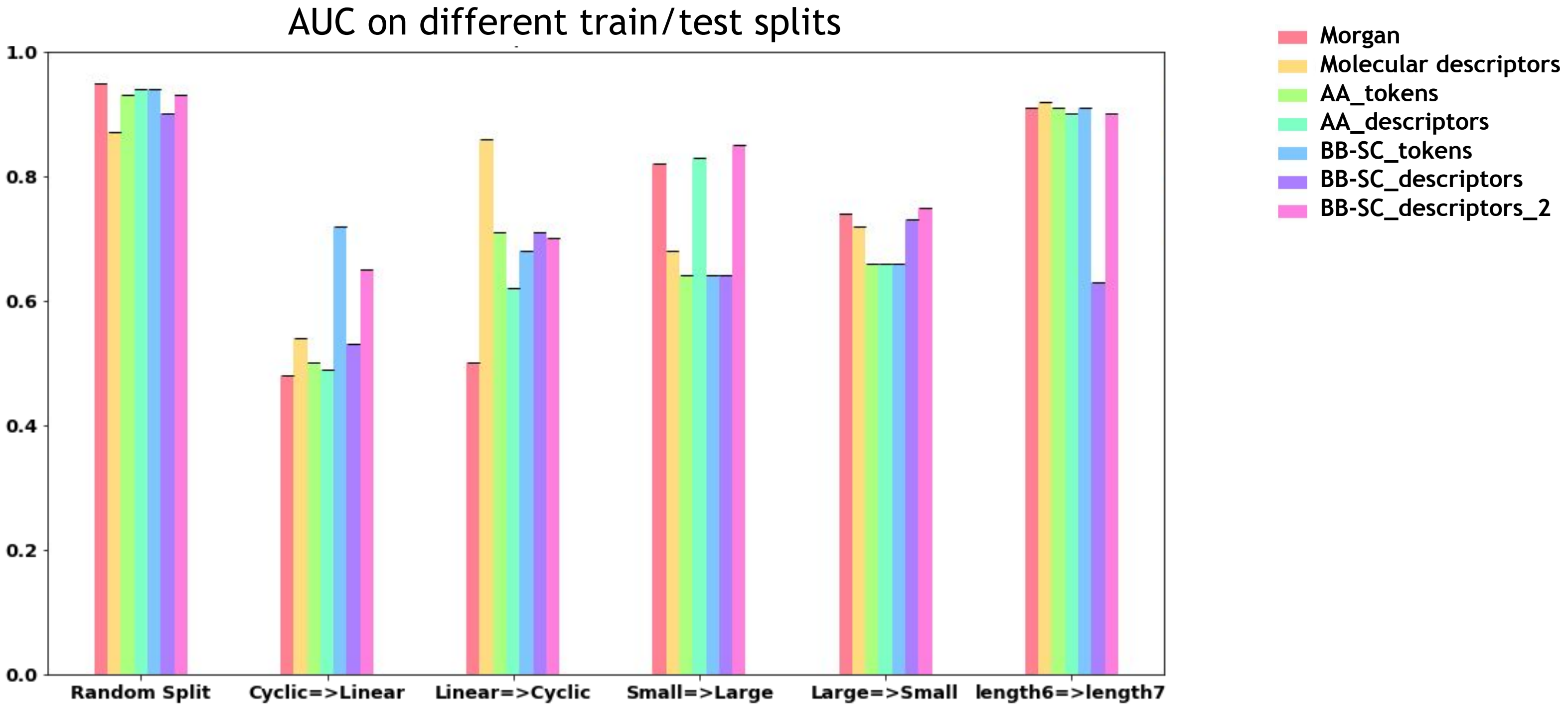
Given dataset is composed of **5339 peptides** (Linear and cyclic peptides) with their measured permeability value in PIC50.

Peptides of the dataset include **modified amino acids**.





Prediction results (Random Forest)



Agenda

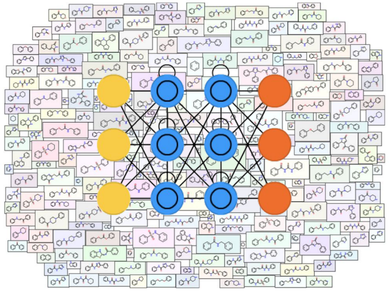
1. Peptides in drug design
2. A matter of representation…
3. To predict…
4. And generate new peptides
5. Conclusion

LSTM generation optimized with reinforcement

Generative AI

Generative model

small molecule: CN1C=NC2=C1C(=O)N(C(=O)N2C)C
peptide: Nter [STyr] W P H W [NMePhe] Cter

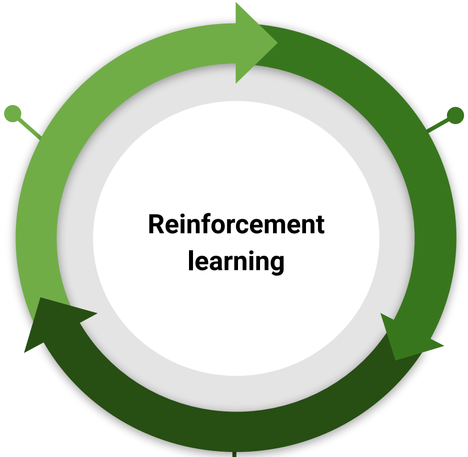


Peptide database

Sampled database of peptides

Reinforcement learning (AI)

Policy gradient



1. Peptides are generated

2. Peptides are scored by the multi-objective fitness function

3. The weights of the model are adjusted to maximize the probability of generating peptides similar to those maximizing the global score using a policy gradient algorithm

Predictors

Machine learning

- Local models
- Global (generic) models

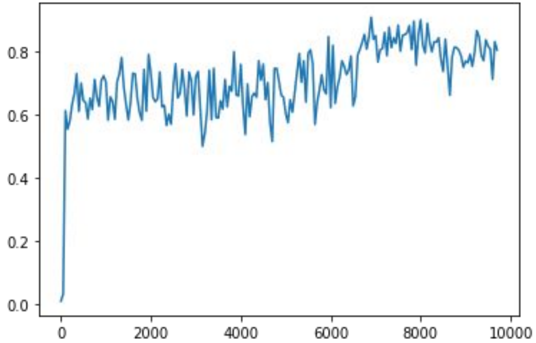
Internal scores

- Metrics: similarity score, Quality Score, Confidence Score.

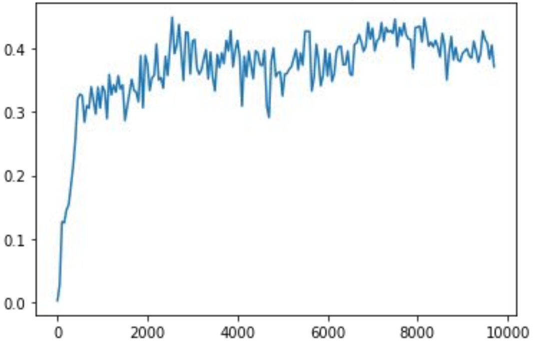
Peptides Generation using predictors trained on project 1

Evolution of scores of generated peptides shows that step by step we are able to optimise different scores of generated peptides.

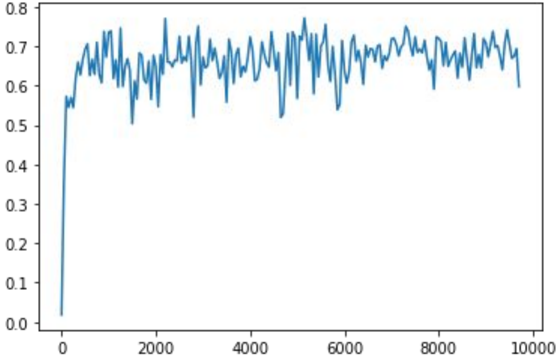
Target 1 predicted SCORE evolution by step



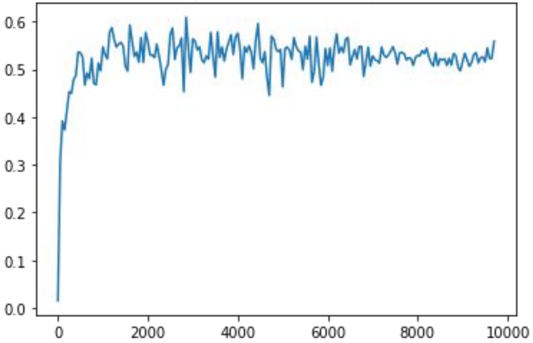
Final SCORE evolution by step.



Target 2 predicted SCORE evolution by step

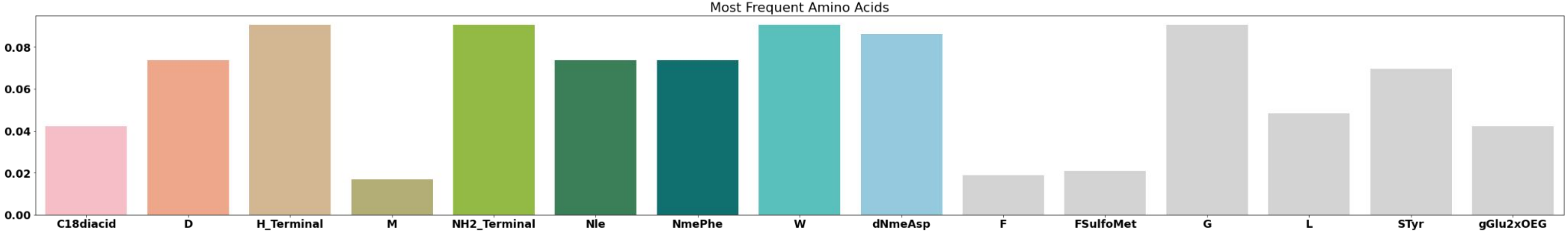


Similarity SCORE evolution by step.

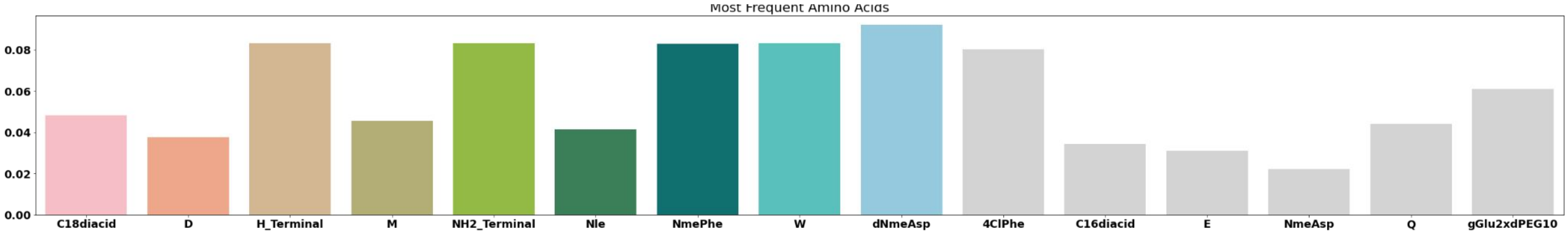


Peptides Generation

Initial Dataset (actives)



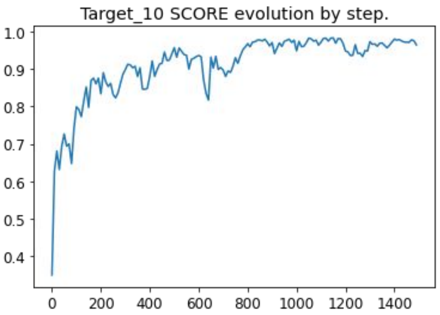
Generated peptides

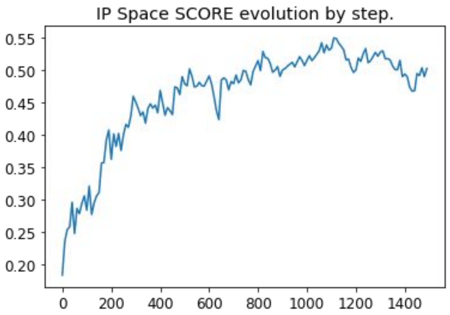


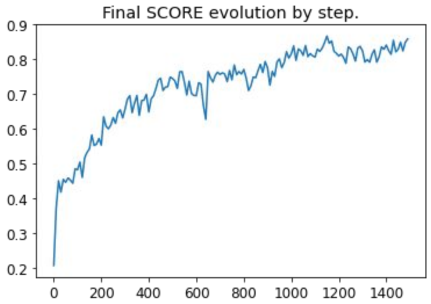
Peptides Generation using predictors trained on project 2

Generation of 62000 peptides satisfying constraints. (Target activity prediction > 0.8, Quality scores)

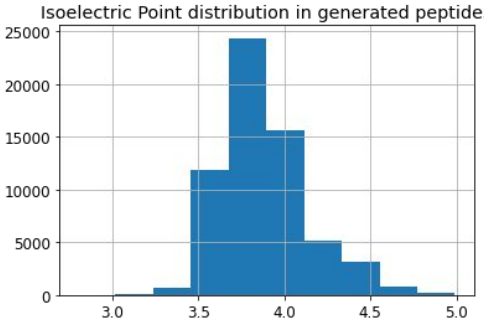
Different scores evolution by steps of optimization.

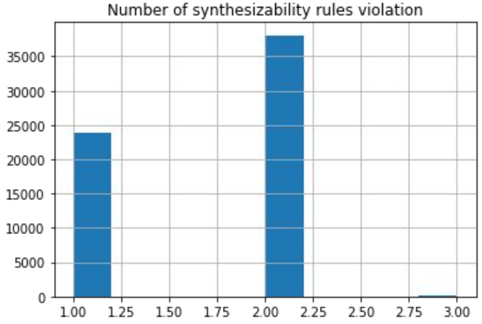

Target_10 SCORE evolution by step.

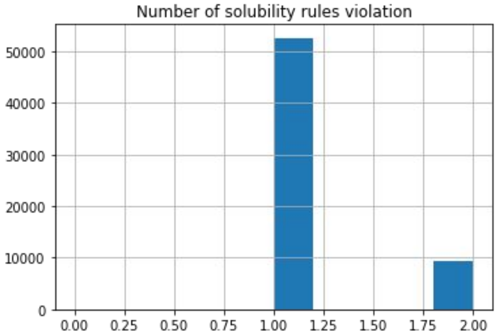

IP Space SCORE evolution by step.


Final SCORE evolution by step.

Checkers distribution in generated peptides.


Isoelectric Point distribution in generated peptides

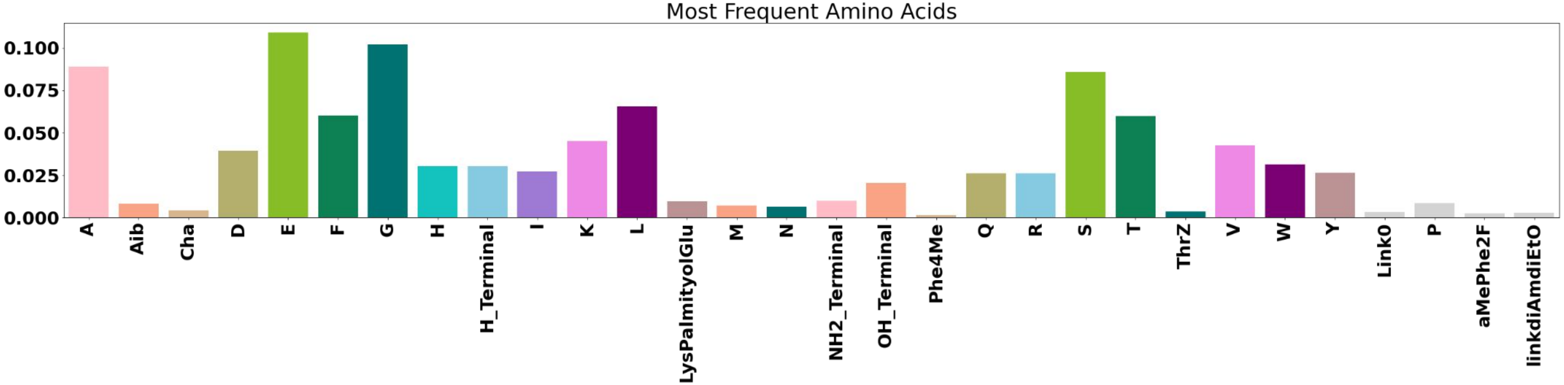

Number of synthesizability rules violation


Number of solubility rules violation

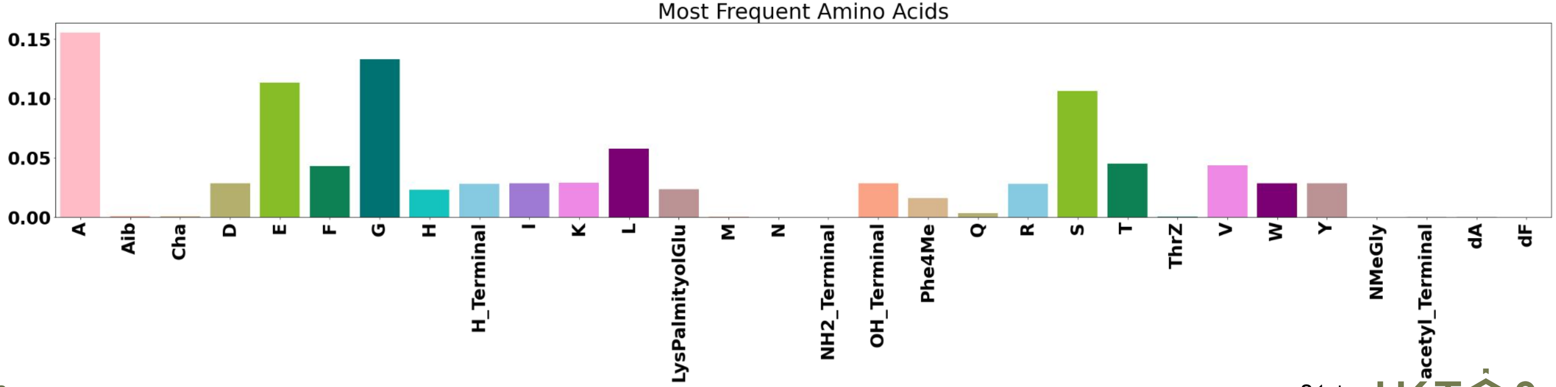
Peptides Generation

Amino acid distribution:

Initial Dataset (actives)



Generated peptides



Agenda

1. Peptides in drug design
2. A matter of representation…
3. To predict…
4. And generate new peptides
5. Conclusion

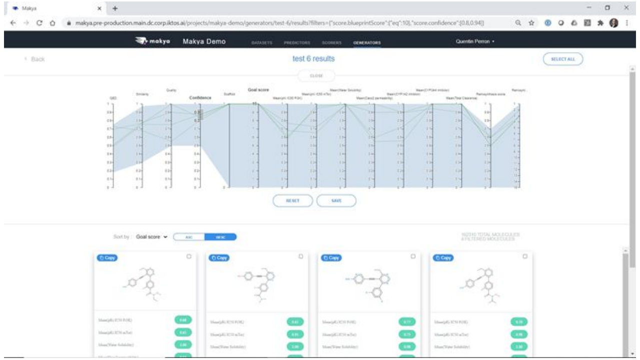
Conclusion

- We introduced new graph representation of peptides at different levels
- These graphs are then converted in a vectorial representation using the Morgan algorithm
- We compared these new representations with classical morgan fingerprint on atomic graph and with molecular descriptors on classification tasks
- We obtained promising results on two different datasets, depending on the splitting scheme
- We used these predictors to generate new peptide with optimized predicted properties

Iktos Our technologies

makya

Deep Generative Chemistry for *de novo* drug design



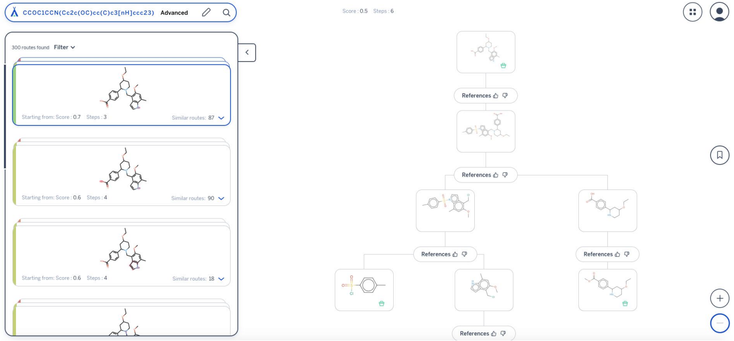
What molecules should I make next?

- Multi-parameter optimization
- Multiple different goal-oriented modes
- Compatible with external tools
- Incorporate IP awareness
- Take advantage of structural knowledge:

makya
API
GENERATOR
DOCKING

spaya.ai

Data-driven Retrosynthesis Analysis


CCOC1CCN(Cc2c(OC)cc(C)c3[nH]ccc23)
Advanced
300 routes found
Filter
Starting from: Score : 0.7 Steps : 3
Similar routes: 87
Starting from: Score : 0.6 Steps : 4
Similar routes: 90
Starting from: Score : 0.6 Steps : 4
Similar routes: 18
Score : 0.5 Steps : 6
References

How can I make these new molecules?

- Find novel synthesis routes for diverse applications
- Explore, share, and collaborate within a team
- Find reference information for all proposed reactions
- Incorporate internal knowledge into synthesis planning AI
- Ensure you are always working with realistic, synthesizable compounds

Try it, it’s free !

www.spaya.ai