Neural Learning of Graphical Models for Protein Design

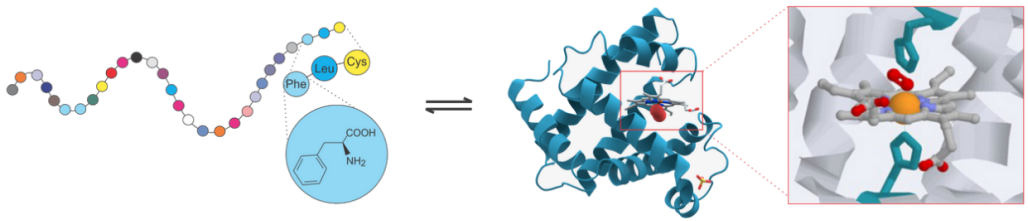
ANITI ARDM Workshop

Marianne Defresne
(INSA, INRAE, ANITI)
-
PhD thesis under the supervision of
Thomas Schiex (MIAT) and Sophie Barbe (TBI)

June, the 22th 2022

Computational Protein Design (CPD)

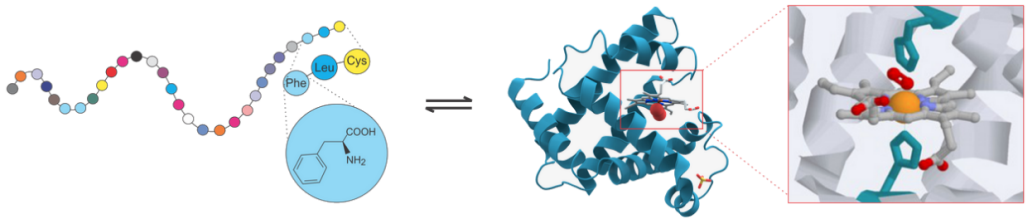
- Protein = sequence of amino acids (among 20 natural types)



Taken from Zhang's lab website

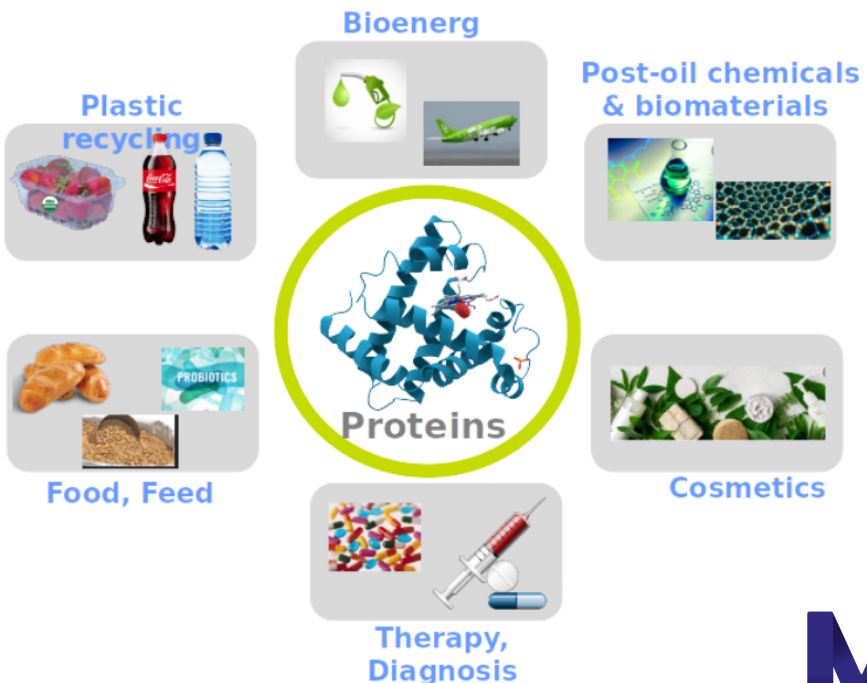
Computational Protein Design (CPD)

- Protein = sequence of amino acids (among 20 natural types)



Taken from Zhang's lab website

- Diversity of applications

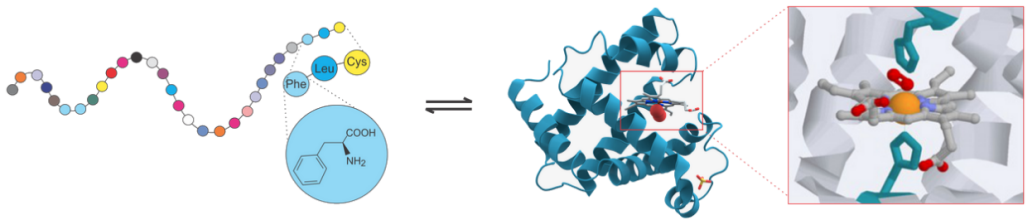


MIA TOULOUSE

tbi Toulouse Biotechnology Institute Bio & Chemical Engineering

Computational Protein Design (CPD)

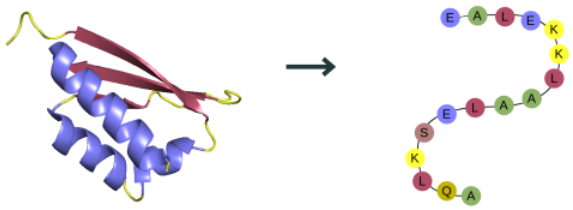
- Protein = sequence of amino acids (among 20 natural types)



Taken from Zhang's lab website

- Diversity of applications
- CPD = modifying/creating proteins for a **function of interest**

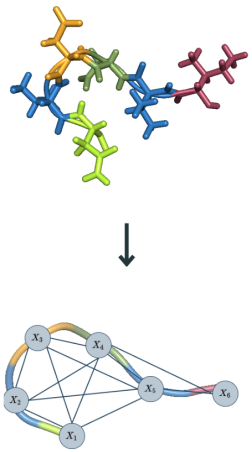
Protein design as a constraint optimization problem



- **Inverse folding problem**: backbone $\mapsto$ sequence
- Criteria: maximum stability = minimum energy
  - $r^* = argmin_r E(r)$
  - Score function = Energy + Design objectives

Protein design as a constraint optimization problem

- **Inverse folding problem**: backbone $\mapsto$ sequence
- Criteria: maximum stability = minimum energy
  - $r^* = argmin_r E(r)$



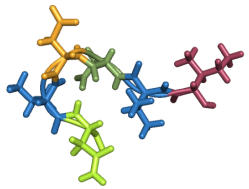
- Energy function as a **graphical model** (GM)
  - One variable X_i per amino acid
  - Domain D_i = amino acid types
  - Cost functions = energy terms

$$E(r) = E_\varnothing + \sum_{i=1}^{n} E_i(r_i) + \sum_{i<j} E_{i,j}(r_i, r_j)$$

- Goal: estimating a conditional energy from data

General context: learning a graphical model

- Learning constraints from real examples
  - > Proteins: laws of physics



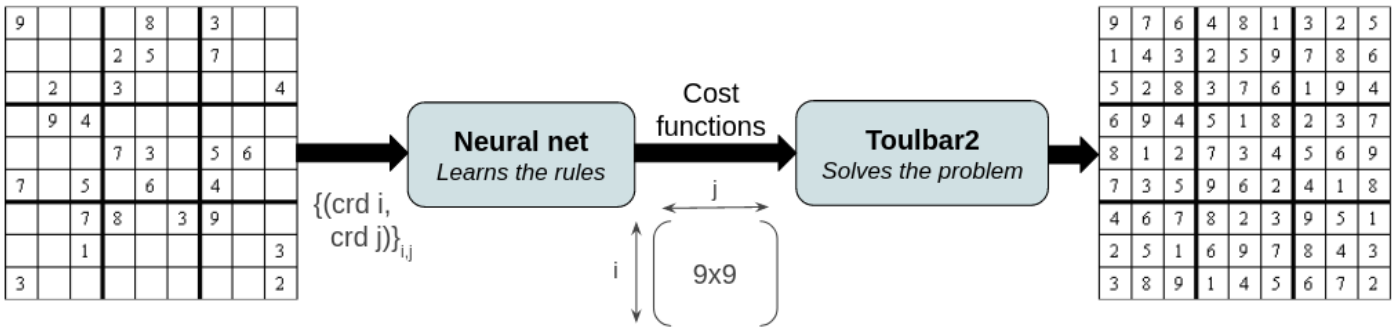
  - > Toy problem: sudoku rules

9	7	6	4	8	1	3	2	5
1	4	3	2	5	9	7	8	6
5	2	8	3	7	6	1	9	4
6	9	4	5	1	8	2	3	7
8	1	2	7	3	4	5	6	9
7	3	5	9	6	2	4	1	8
4	6	7	8	2	3	9	5	1
2	5	1	6	9	7	8	4	3
3	8	9	1	4	5	6	7	2

- Why? To be able to use learned constraints on new examples
- How? By interfacing 2 branches of AI:
  - > **Deep Learning** criteria from examples
  - > **Automated reasoning** to identify the optimal solution

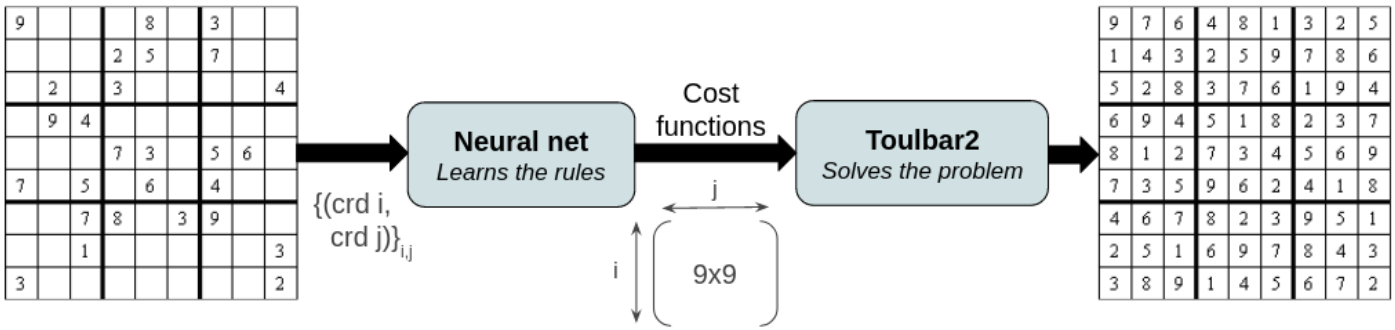
Toy example: the sudoku

- Learning the rules from sudoku examples



Toy example: the sudoku

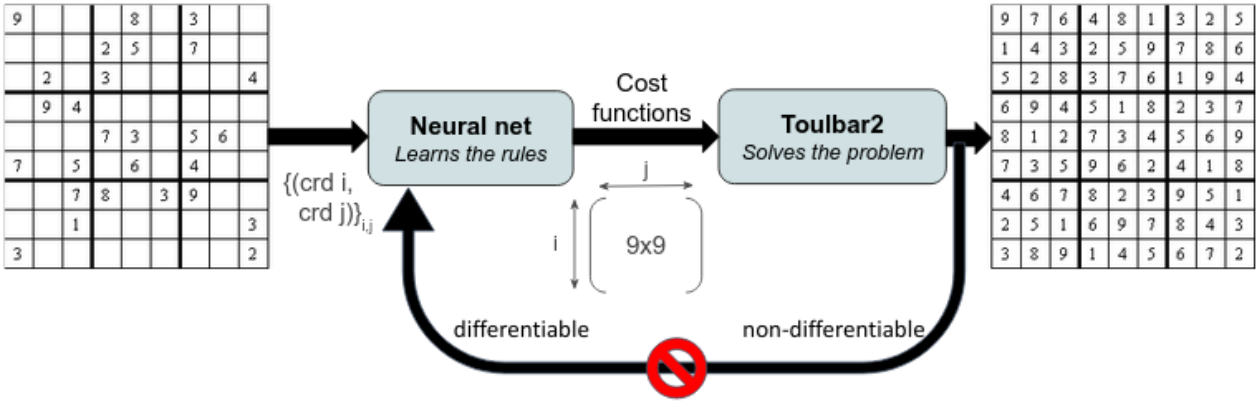
- Learning the rules from sudoku examples



- Why is it similar to protein design?
  - Grid = backbone; cell = residue
  - Pairwise interactions
  - Cost function depends only on relative coordinates

Toy example: the sudoku

- Learning the rules from sudoku examples

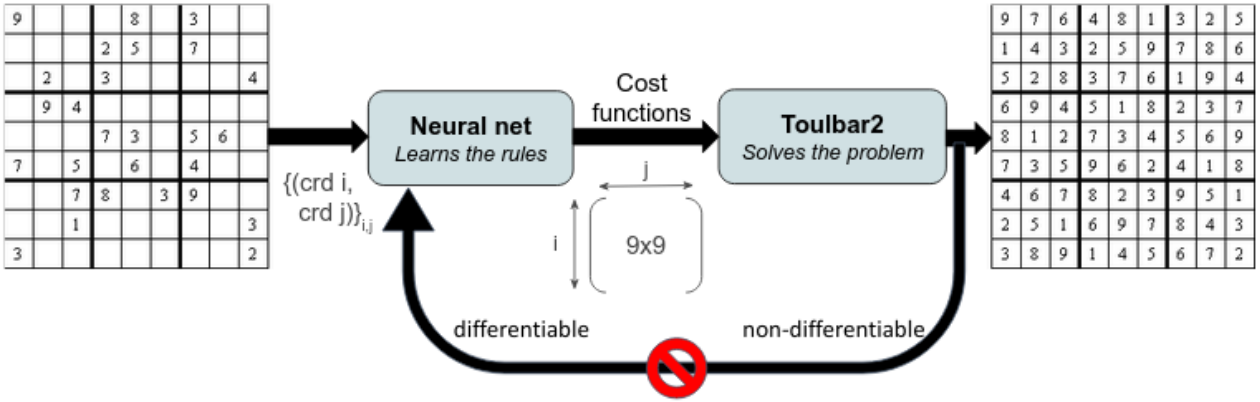


- Objective:

$$L = Hamming(y, \hat{y}) = \frac{1}{81} \sum_{i=1}^{81} \mathbb{1}[y_i \neq \hat{y}_i]$$

Toy example: the sudoku

- Learning the rules from sudoku examples



- Objective:

$$L = Hamming(y, \hat{y}) = \frac{1}{81} \sum_{i=1}^{81} \mathbb{1}[y_i \neq \hat{y}_i]$$

- Difficulty: **Discrete** objective vs **gradient** descent
  - Differentiable relaxation: SATNet *(Wang et al. 2019)*
  - Continuous interpolation: Blackbox solver *(Pogančić et al. 2019)*
  - Differentiable & informative upper bound: Hinge loss *(Tsochantaridis et al. 2005)*

Results on the toy example

- Performance + data-efficiency

Approach	Accuracy	Train size	Reference
Pure DL	96.6%	180,000	*(Palm, Paquet, and Winther. NeurIPS2018)*
SATNet *	99.8%	9,000	*(Wang et al. ICML2019)*
ML + toulbar2	100%	9,000	*(Brouard, Givry, and Schiex. CP2020)*
DL + toulbar2	**100%**	**1,000**	-

** Much easier dataset*

Results on the toy example

- Performance + data-efficiency

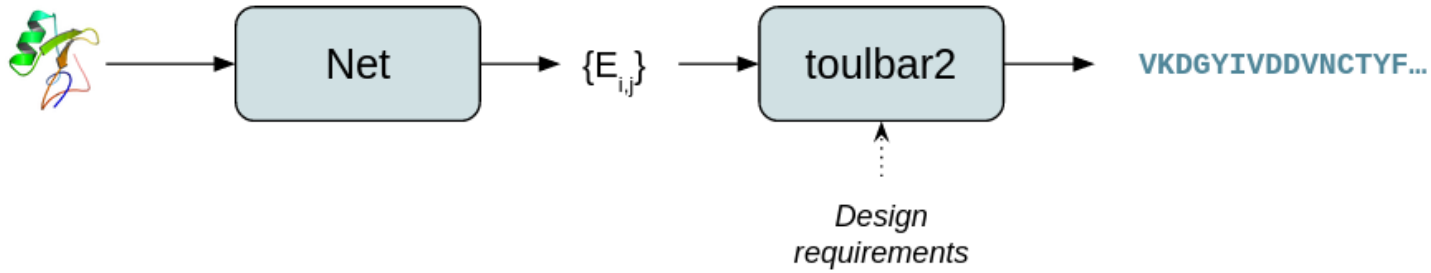
Approach	Accuracy	Train size	Reference
Pure DL	96.6%	180,000	*(Palm, Paquet, and Winther. NeurIPS2018)*
SATNet *	99.8%	9,000	*(Wang et al. ICML2019)*
ML + toulbar2	100%	9,000	*(Brouard, Givry, and Schiex. CP2020)*
DL + toulbar2	**100%**	**1,000**	-

** Much easier dataset*

- Understanding what is learned:
  - **Checking the rules** learned (if known)
  - **Interpreting** the rules (reasonable size)
    - Non-redundant rules of sudoku

Back to proteins

- **Objective**: reconstruct the natural sequence



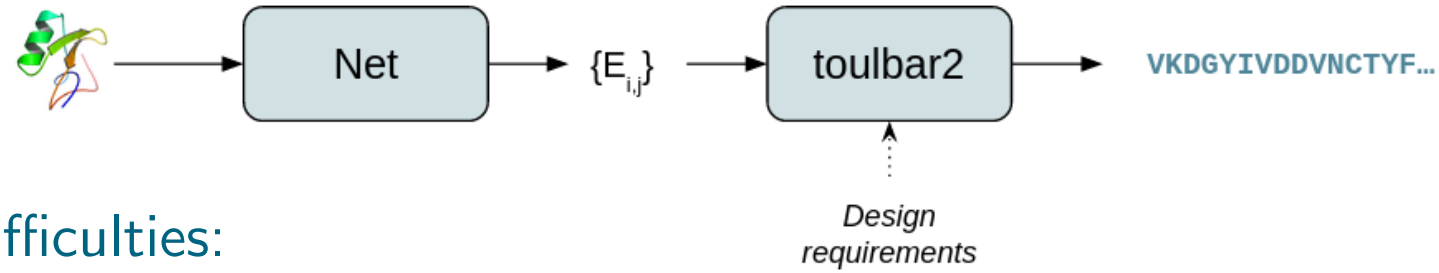
- **Dataset**: RCSB PDB PROTEIN DATA BANK
  - Split like *(Ingraham et al. 2019)*[1]
  - 17,000 training proteins, 1,200 for test and 600 for validation

[1] J. Ingraham et al. (2019). "Generative models for graph-based protein design". In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

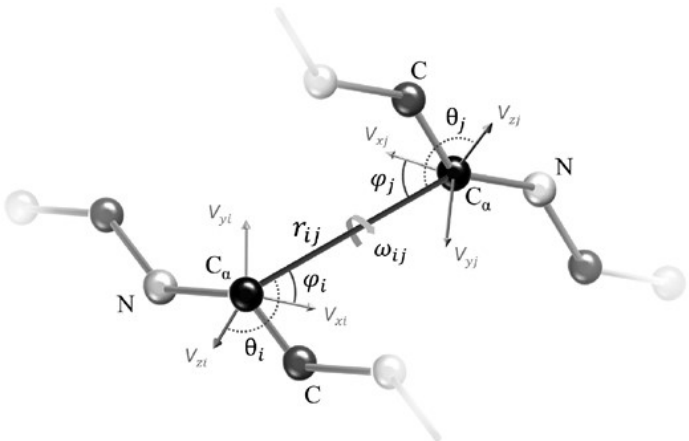
MIA TOULOUSE
tbi Toulouse Biotechnology Institute Bio & Chemical Engineering

Back to proteins

- **Objective**: reconstruct the natural sequence



- Difficulties:
  - **Protein representation**: no consensus
    - Chose pairwise features (relative distance & orientation)



(López-Blanco and Chacón 2019)

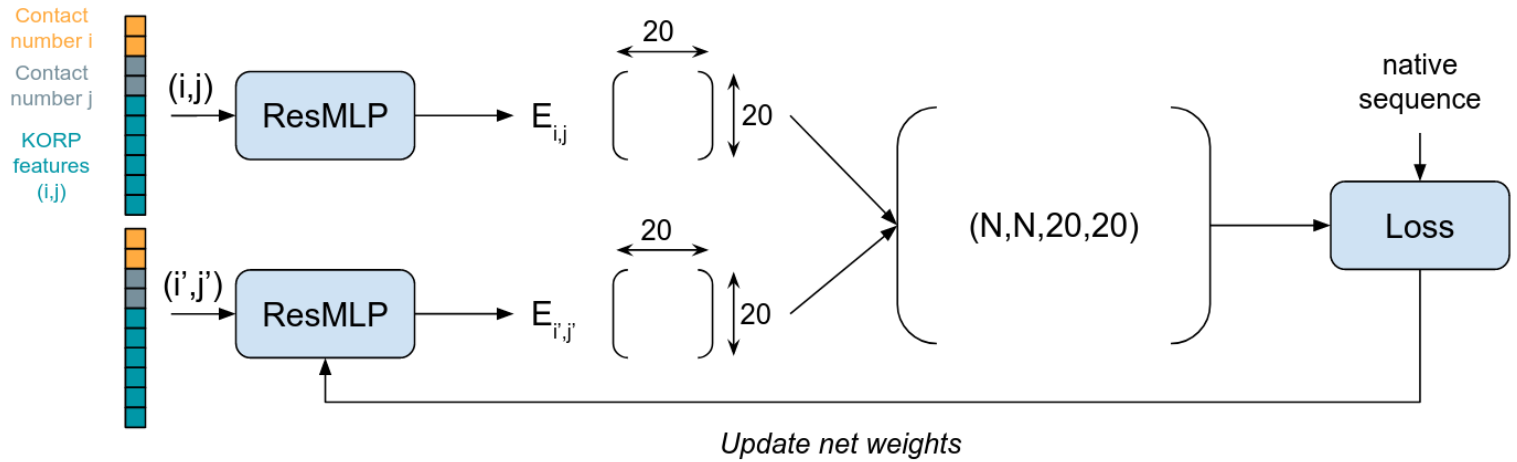
  - Training alongside toulbar2 too slow

MIA TOULOUSE

tbi Toulouse Biotechnology Institute Bio & Chemical Engineering

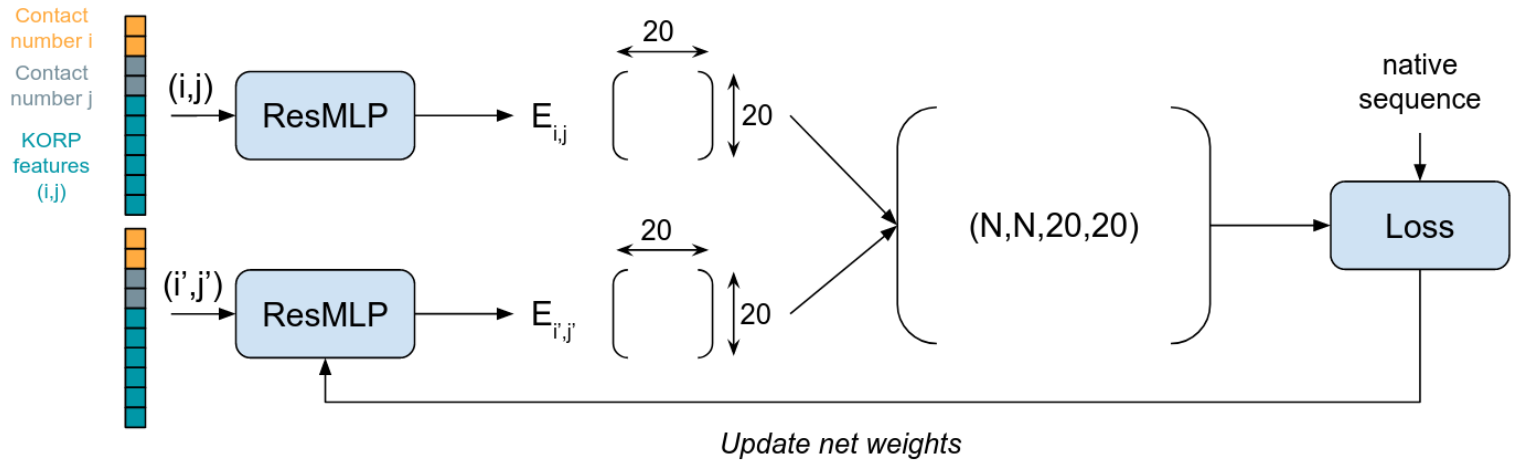
Direct adaptation of the sudoku pipeline to proteins

- **Pure neural training** with likelihood-based loss
- Toulbar2 used for inference (full protein design)



Direct adaptation of the sudoku pipeline to proteins

- **Pure neural training** with likelihood-based loss
- Toulbar2 used for inference (full protein design)



- Initial model; in progress
  - Global rotation/translation invariance
  - Taking into account the environment of each residue

MIA TOULOUSE

tbi Toulouse Biotechnology Institute Bio & Chemical Engineering

Results on proteins

No direct metric to assess the learned GM $\rightarrow$ **auxiliary tasks**

- Predicting one masked residue
  - Accuracy: 42%

[2] Valentin Durante, George Katsirelos, and Thomas Schiex (July 2022). "Efficient low rank convex bounds for pairwise discrete Graphical Models". In: *Thirty-ninth International Conference on Machine Learning*.

Results on proteins

No direct metric to assess the learned GM $\rightarrow$ **auxiliary tasks**

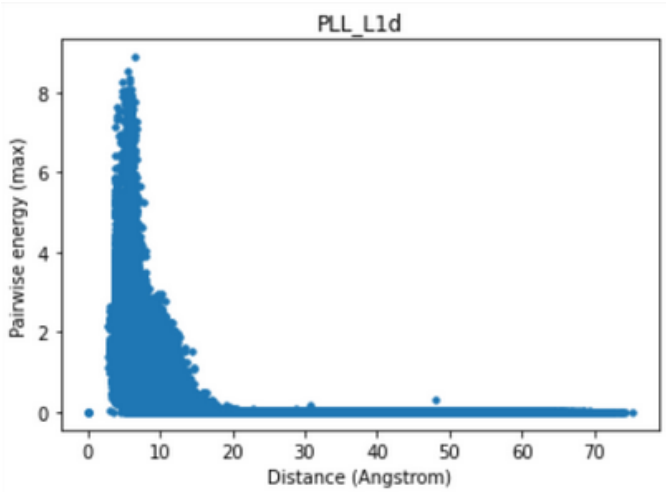
- Predicting one masked residue
  - Accuracy: 42%
- Predicting full sequence
  - Inference with a convex relaxation of toulbar2 *(Durante, Katsirelos, and Schiex 2022)*[2]
  - Recovery: 35.6%

[2] Valentin Durante, George Katsirelos, and Thomas Schiex (July 2022). "Efficient low rank convex bounds for pairwise discrete Graphical Models". In: *Thirty-ninth International Conference on Machine Learning*.

Results on proteins

No direct metric to assess the learned GM $\rightarrow$ **auxiliary tasks**

- Predicting one masked residue
  - Accuracy: 42%
- Predicting full sequence
  - Inference with a convex relaxation of toulbar2 *(Durante, Katsirelos, and Schiex 2022)*[2]
  - Recovery: 35.6%
- Trend of energy vs distance



[2] Valentin Durante, George Katsirelos, and Thomas Schiex (July 2022). "Efficient low rank convex bounds for pairwise discrete Graphical Models". In: *Thirty-ninth International Conference on Machine Learning*.

MIA TOULOUSE

tbi Toulouse Biotechnology Institute Bio & Chemical Engineering

Results on proteins: decoy task

- KORP dataset *(López-Blanco and Chacón 2019)*[3]
  - Task: Identifying the natural protein among decoys (~ 100)
  - Comparison with state-of-the art statistical potential **KORP**
  - Correct: 200/200 (vs 193/200 pour KORP)

[3] José Ramón López-Blanco and Pablo Chacón (Jan. 2019). "KORP: knowledge-based 6D potential for fast protein and loop modeling". In: *Bioinformatics* 35.17, pp. 3013–3019. ISSN: 1367-4803.

[4] Hahnbeom Park et al. (2016). "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules". In: *Journal of Chemical Theory and Computation* 12.12, pp. 6201–6212.

Results on proteins: decoy task

- KORP dataset *(López-Blanco and Chacón 2019)*[3]
  - Task: Identifying the natural protein among decoys (~ 100)
  - Comparison with state-of-the art statistical potential **KORP**
  - Correct: 200/200 (vs 193/200 pour KORP)
- Rosetta dataset *(Park et al. 2016)*[4]
  - Task: Ranking decoys quality (measured by TM score)
  - Comparison with **Rosetta** energy function (all-atom)

Approach	Spearman	TM best
Rosetta	0.798	92.2
Effie	**0.813**	**93.0**

[3] José Ramón López-Blanco and Pablo Chacón (Jan. 2019). "KORP: knowledge-based 6D potential for fast protein and loop modeling". In: *Bioinformatics* 35.17, pp. 3013–3019. ISSN: 1367-4803.

[4] Hahnbeom Park et al. (2016). "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules". In: *Journal of Chemical Theory and Computation* 12.12, pp. 6201–6212.

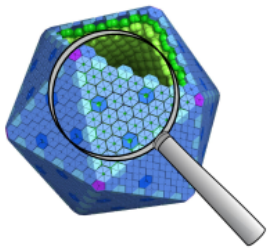
Conclusion & Perspectives

- Advantage of the method
  - **Quality** of the decision
  - **Handling output**: adding constraints, criteria
  - **Interpretable output**: understanding what is learned

Conclusion & Perspectives

- Advantage of the method
  - **Quality** of the decision
  - **Handling output**: adding constraints, criteria
  - **Interpretable output**: understanding what is learned

- **Perspectives**: production & experimental characterization of designed proteins
  - Hexamer from microbial compartment
  - Nanotech application *(spatial organization of enzymes)*



MIA TOULOUSE

tbi Toulouse Biotechnology Institute Bio & Chemical Engineering

Acknowledgment

- **Sophie Barbe** and **Thomas Schiex** for supervision
- **CALMIP** for computational resources
- The organizers of this ARDM workshop

Thanks for your attention!

This work has been supported by the Agence Nationale de la Recherche (ANR) [grant ANR-18-EURE-0021]

MIA TOULOUSE

tbi Toulouse Biotechnology Institute Bio & Chemical Engineering

References I

- Durante, Valentin, George Katsirelos, and Thomas Schiex (July 2022). "Efficient low rank convex bounds for pairwise discrete Graphical Models". In: *Thirty-ninth International Conference on Machine Learning*.
- Ingraham, J. et al. (2019). "Generative models for graph-based protein design". In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.
- López-Blanco, José Ramón and Pablo Chacón (Jan. 2019). "KORP: knowledge-based 6D potential for fast protein and loop modeling". In: *Bioinformatics* 35.17, pp. 3013–3019. ISSN: 1367-4803.

References II

- Park, Hahnbeom et al. (2016). "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules". In: *Journal of Chemical Theory and Computation* 12.12, pp. 6201–6212.
- Pogančić, Marin Vlastelica et al. (2019). "Differentiation of blackbox combinatorial solvers". In: *International Conference on Learning Representations*.
- Tsochantaridis, Ioannis et al. (2005). "Large margin methods for structured and interdependent output variables.". In: *Journal of machine learning research* 6.9.
- Wang, Po-Wei et al. (Sept. 2019). "SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. PMLR.