



DESIGNING WITH INTUITION¹ & LOGIC²

November 16, 2023

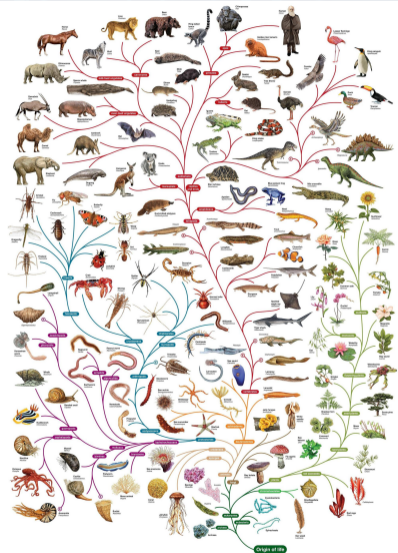
Thomas Schiex

Co-chairs: S. Barbe, S. de Givry, G. Katsirelos, D. Simoncini

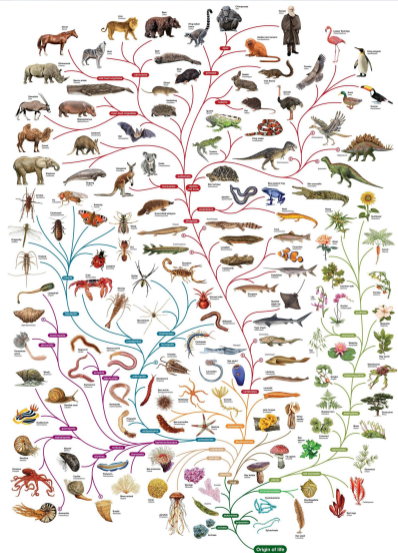
PhDs: M. Defresne, V. Durante, P. Montalbano



- ▶ Most active molecules of life (virus to humans)
- ▶ Useful in health to green chemistry



- ▶ Most active molecules of life (virus to humans)
- ▶ Useful in health to green chemistry



DVVGKVVVDGKDD···GVKVGDKVKVKKV

Organizes different types of atoms in 3D

Sequence \rightsquigarrow Structure \rightsquigarrow Function

DVVGKVVVDGKDD···GVKVGDKVKVKKV

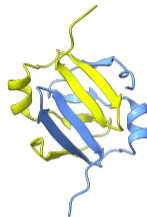


Organizes different types of atoms in 3D

Sequence \rightsquigarrow Structure \rightsquigarrow Function

Protein folding

DVVGKVVVDGKDD···GVKVGDKVKVKKV



Organizes different types of atoms in 3D

Sequence \rightsquigarrow Structure \rightsquigarrow Function

X

Amino acid sequence
(20 letters alphabet)

 Φ

Continuous SE(3)-invariant
3D structure

Organizes different types of atoms in 3D

Sequence \rightsquigarrow Structure \rightsquigarrow Function

X

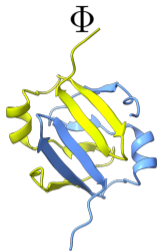
Amino acid sequence
(20 letters alphabet)

 Φ

Continuous SE(3)-invariant
3D structure

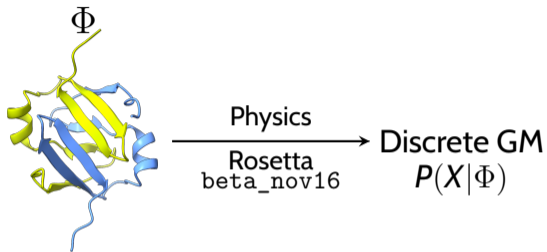
Organizes different types of atoms in 3D

Sequence \rightsquigarrow Structure \rightsquigarrow Function



A quite successful all physics+logic generative process

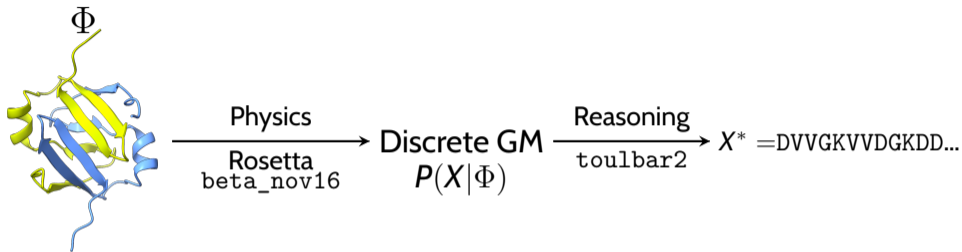
The Toulbar package for WCSPs significantly improved the state-of-the-art efficiency for protein design
Com. ACM-20, B. Donald et al.



A quite successful all physics+logic generative process

The Toulbar package for WCSPs significantly improved the state-of-the-art efficiency for protein design
Com. ACM-20, B. Donald et al.

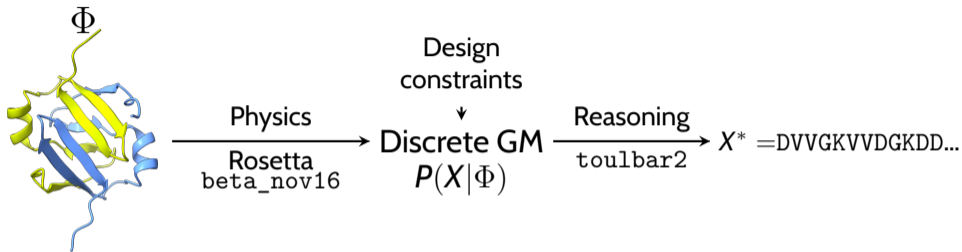
Designing Proteins with logic²



A quite successful all physics+logic generative process

The Toulbar package for WCSPs significantly improved the state-of-the-art efficiency for protein design
Com. ACM-20, B. Donald et al.

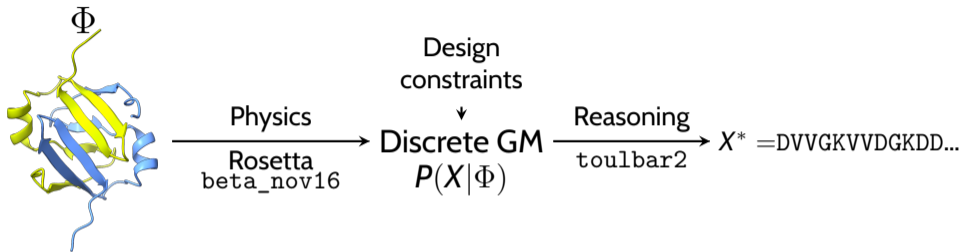
Designing Proteins with logic²



A quite successful all physics+logic generative process

The Toulbar package for WCSPs significantly improved the state-of-the-art efficiency for protein design
Com. ACM-20, B. Donald et al.

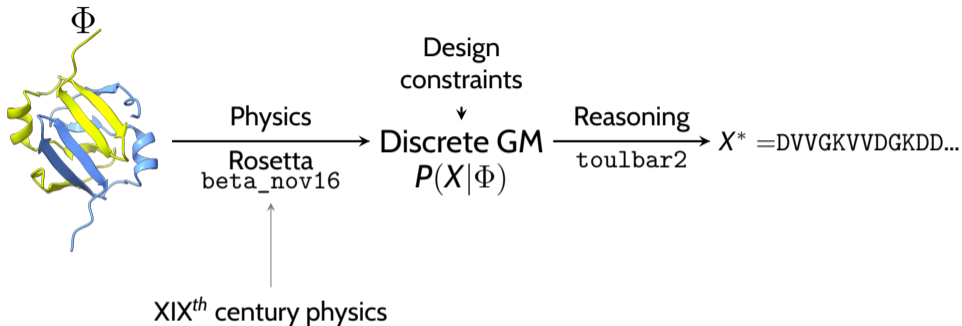
Designing Proteins with logic²



A quite successful all physics+logic generative process

The Toulbar package for WCSPs significantly improved the state-of-the-art efficiency for protein design
Com. ACM-20, B. Donald et al.

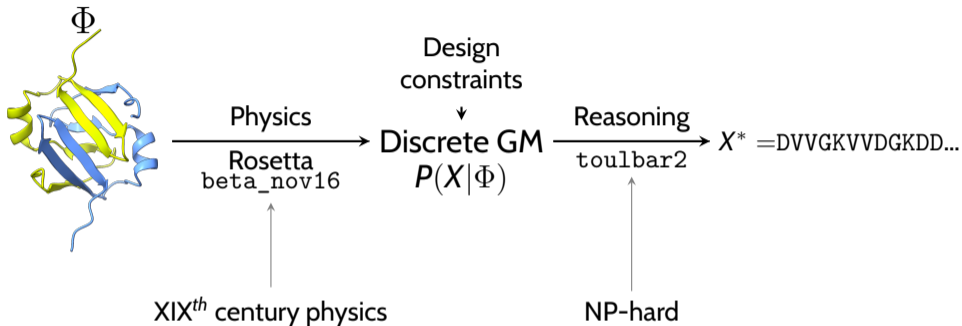
Designing Proteins with logic²



A quite successful all physics+logic generative process

The Toulbar package for WCSPs significantly improved the state-of-the-art efficiency for protein design
Com. ACM-20, B. Donald et al.

Designing Proteins with logic²



A quite successful all physics+logic generative process

The Toulbar package for WCSPs significantly improved the state-of-the-art efficiency for protein design
Com. ACM-20, B. Donald et al.

Discrete (and continuous) optimization

(generic)

- ▶ LR-BCD: SDP relaxation + low rank solver for discrete GMs
- ▶ Discrete solver: Low rank SDP + bundle + branch and cut
- ▶ Parallel Best/Depth-first search algorithm
- ▶ Multiple choice Knapsack constraints
- ▶ Second level of the k -consistency hierarchy bounds
- ▶ ...

ICML'2022

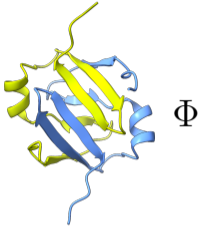
To be submitted

CP'2022

CPAIOR'2022

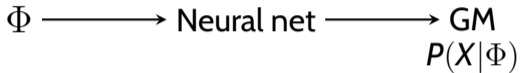
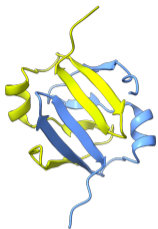
CPAIOR'2023

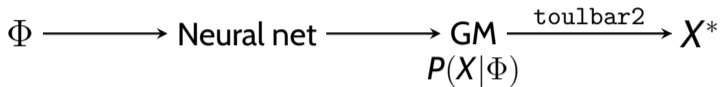
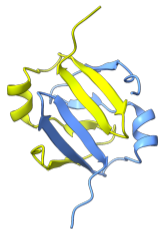
V. Durante and P. Montalbano PhDs (defenses on December 15th, all day, INRAE-MIAT)



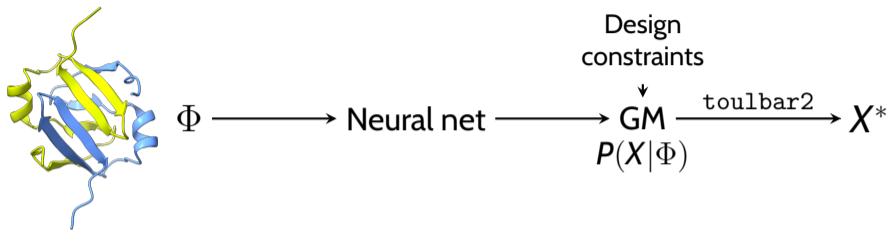


Φ \longrightarrow Neural net

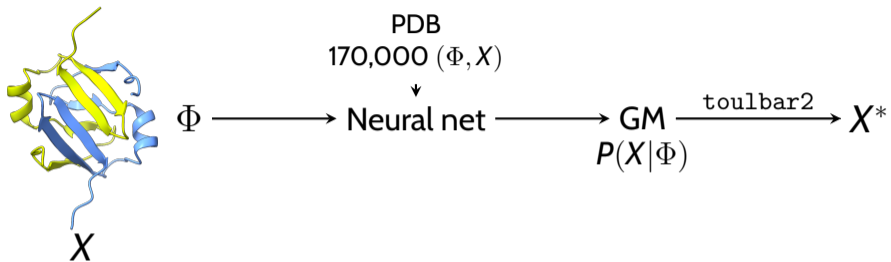




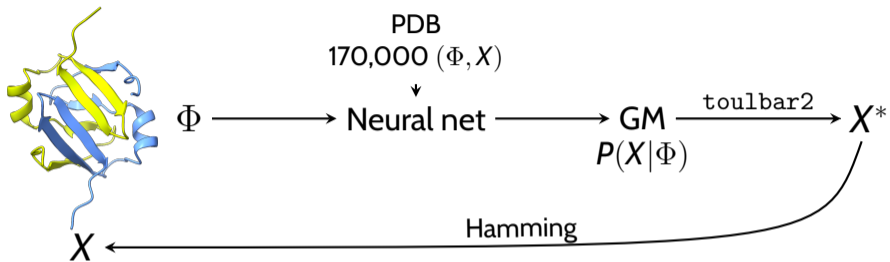
Injecting intuition¹



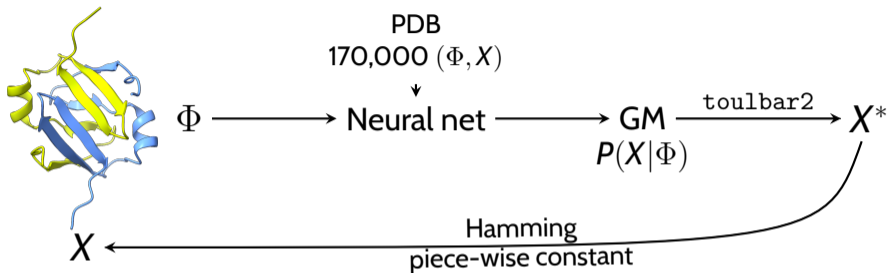
Injecting intuition¹



Injecting intuition¹



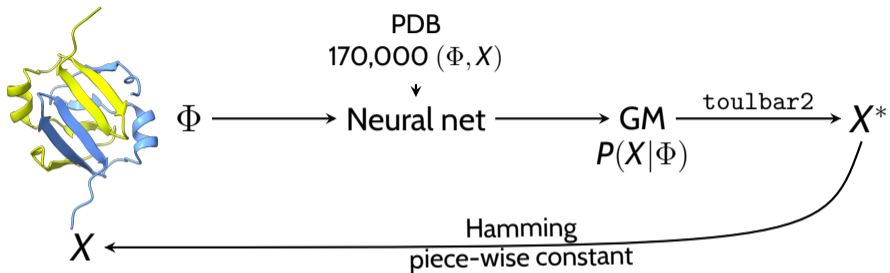
Injecting intuition¹



Issues

- ▶ Gradients either zero or undefined
- ▶ Requires to repeatedly solve random NP-hard instances

Injecting intuition¹



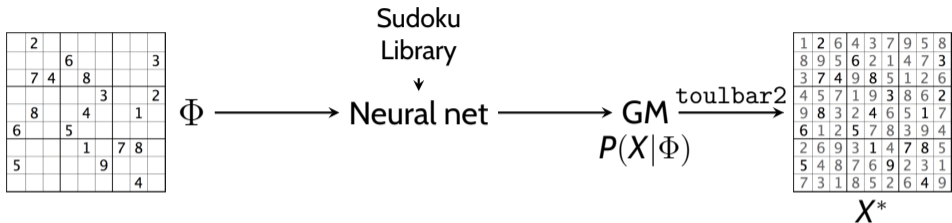
Our solution

- ▶ Introduced a dedicated loss: the E-Pseudo Log Likelihood
- ▶ Kicked the solver out of the training loop

IJCAI'2023

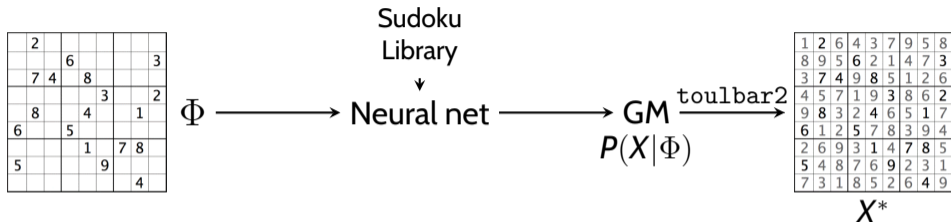
(Defresne et al. 2023)

Learning to play Sudoku



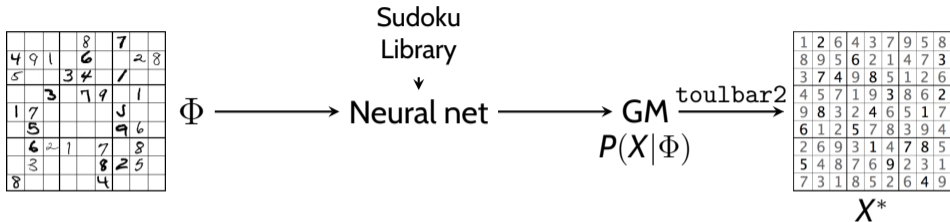
Approach	Architecture	Acc.	Grids	Training set
RRN NeurIPS18	GNN	96.6%	Hard	180,000
SATNet ICML19	Relaxation	99.8%	Easy	9,000
Hybrid IJCAI23	E-PLL	100%	Hard	200

Learning to play Sudoku



Approach	Architecture	Acc.	Grids	Training set
RRN <small>NeurIPS18</small>	GNN	96.6%	Hard	180,000
SATNet <small>ICML19</small>	Relaxation	99.8%	Easy	9,000
Hybrid <small>IJCAI23</small>	E-PLL	100%	Hard	200

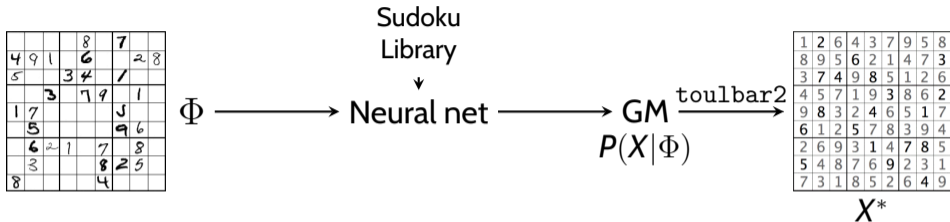
Learning to play Visual Sudoku



Simultaneously learns to recognize digits and to play the Sudoku

SATNet	Theoretical (no corrections)	Hybrid
63.2 %	74.2%	94.1 ± 0.8%

Learning to play Visual Sudoku

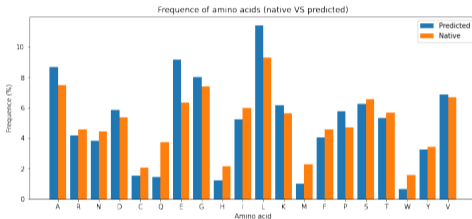
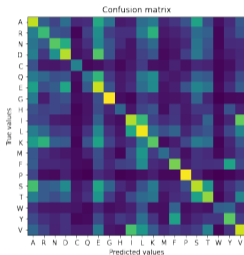


Simultaneously learns to recognize digits and to play the Sudoku

SATNet	Theoretical (no corrections)	Hybrid
63.2 %	74.2%	94.1 ± 0.8%

Recovering amino acid properties

- ▶ Correctly predicts 51% of amino acids from their environment



Zero-shot prediction of the effect of single mutations

- ▶ 79% accuracy on ATOM3D benchmark
- ▶ 0.4 correlation stability score/predicted energy (Rocklin et al. 2017)

Optimizing a complete protein sequence

Full redesign of large proteins in the test set

- ▶ Guaranteed `trRosetta` solution expensive
- ▶ Using LR-BCD instead (Durante et al. 2022)

Outperforms all-atoms XIXth-century physics

- ▶ Metric: Native Sequence Recovery rate (NSR)

Approach	Rosetta	Effie
NSR	17.9%	32.8%

M. Defresne PhD defense (November 30th, 2:30 PM, INRAE-MIAT).

Optimizing a complete protein sequence

Full redesign of large proteins in the test set

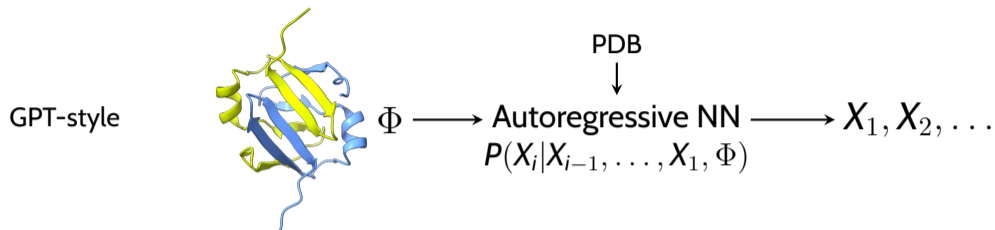
- ▶ Guaranteed `trRosetta` solution expensive
- ▶ Using LR-BCD instead (Durante et al. 2022)

Outperforms all-atoms XIXth-century physics

- ▶ Metric: **Native Sequence Recovery** rate (NSR)

Approach	Rosetta	Effie
NSR	17.9%	32.8%

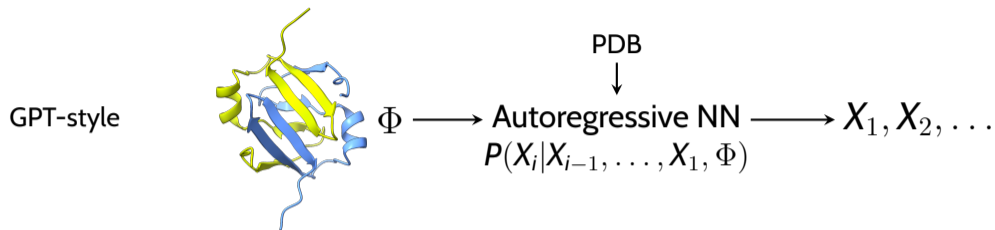
M. Defresne PhD defense (November 30th, 2:30 PM, INRAE-MIAT).



Pros and cons

- ▶ Efficient sampling instead of NP-hard solving
- ▶ Capacity to capture higher-order interactions
- ▶ Limited control for design constraints

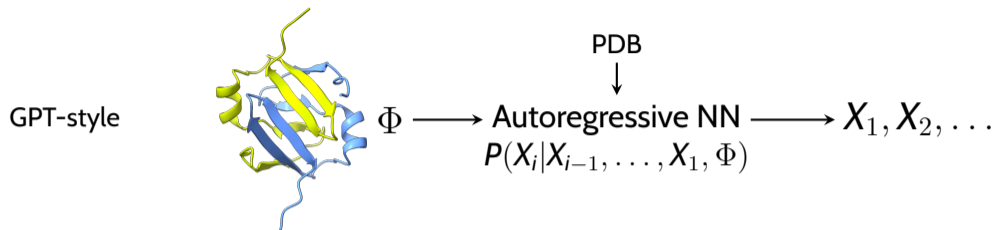
	ProteinMPNN	Effie
NSR	45.9%	48.4%



Pros and cons

- ▶ Efficient sampling instead of NP-hard solving
- ▶ Capacity to capture higher-order interactions
- ▶ Limited control for design constraints

	ProteinMPNN	Effie
NSR	45.9%	48.4%



Pros and cons

- ▶ Efficient sampling instead of NP-hard solving
- ▶ Capacity to capture higher-order interactions
- ▶ Limited control for design constraints

	ProteinMPNN	Effie
NSR	45.9%	48.4%

Predicting SARS-CoV2 variants



Enumerate CoViD variants with a bounded number of mutations

- ▶ Uses only the initial March 2020 RBD-ACE2 structure + Effie/toulbar2
- ▶ Relies on (Montalbano et al. 2022) constraint to bound mutations
- ▶ Predicts all the first SARS-CoV2 VoCs (α , β , γ , δ , κ , ι , λ and μ)
- ▶ In a few seconds, on one CPU-thread.

Not achievable by pure autoregressive models (ProteinMPNN)



DVVGKVVVDGKDD · · · GVKVGDKVKVKKV






Designing the RNA polymerase double Ψ - β -barrel with simple chemistry

- ▶ Done with 7 AA types (physics+logic (Yagi et al. 2021))
- ▶ Sequences with 6, 5, and 4 AA types correctly folded by AlphaFold
- ▶ Relies on (Montalbano et al. 2022) constraint to bound the # of AA types
- ▶ To be synthesized and crystallized (RIKEN collab.)

Not achievable by pure autoregressive models (ProteinMPNN)

Design of an enzyme organizing platform




Design of an heteromeric hexamer

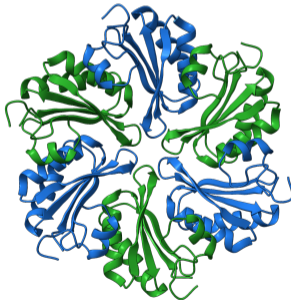
- ▶ Design ▲ and ▲ that self-assemble as  but not as  or 
- ▶ Physics+logic: requires bi-level optimization (NP^{NP} -complete) (Vucinic et al. 2020)
- ▶ Can be solved by Effie+tb2 (NP-complete) or ProteinMPNN, using bi-criteria optimization



Design of an enzyme organizing platform






Design of an heteromeric hexamer

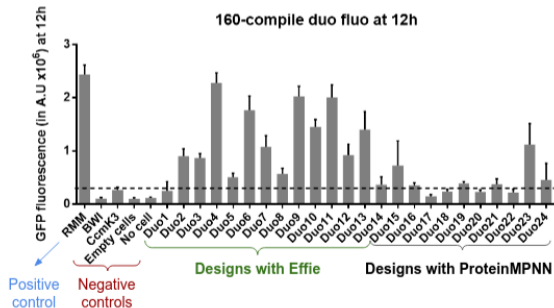
- ▶ Design ▲ and ▲ that self-assemble as  but not as  or 
- ▶ Physics+logic: requires bi-level optimization (NP^{NP} -complete) (Vucinic et al. 2020)
- ▶ Can be solved by Effie+tb2 (NP-complete) or ProteinMPNN, using bi-criteria optimization



Design of an enzyme organizing platform

Design of a heteromeric hexamer

- ▶ Design  and  that self-assemble as  but not as  or 
- ▶ Physics+logic: requires bi-level optimization (NP^{NP} -complete) (Vucinic et al. 2020)
- ▶ Can be solved by Effie+tb2 (NP-complete) or ProteinMPNN, using bi-criteria optimization



A Neural Net, a Graphical Model and a discrete solver in a hybrid autoencoder

- ▶ A hybrid generic Generative AI that benefits from each component
- ▶ Neural Network: ideal to extract a representation of $P(X|\Phi)$ from raw inputs
- ▶ Represented as a GM in a fully explorable and controllable latent layer
- ▶ Using decoding by discrete reasoning (toulbar2 or LR-BCD)
- ▶ All this with scalable training

Future: SDP solver, hidden variables, higher-order dependencies,

A Neural Net, a Graphical Model and a discrete solver in a hybrid autoencoder

- ▶ A hybrid generic Generative AI that benefits from each component
- ▶ Neural Network: ideal to extract a representation of $P(X|\Phi)$ from raw inputs
- ▶ Represented as a GM in a fully explorable and controllable latent layer
- ▶ Using decoding by discrete reasoning (toulbar2 or LR-BCD)
- ▶ All this with scalable training

Future: SDP solver, hidden variables, higher-order dependencies,

A Neural Net, a Graphical Model and a discrete solver in a hybrid autoencoder

- ▶ A hybrid generic Generative AI that benefits from each component
- ▶ Neural Network: ideal to extract a representation of $P(X|\Phi)$ from raw inputs
- ▶ Represented as a GM in a fully explorable and controllable latent layer
 - ▶ Using decoding by discrete reasoning (toulbar2 or LR-BCD)
 - ▶ All this with scalable training

Future: SDP solver, hidden variables, higher-order dependencies,

A Neural Net, a Graphical Model and a discrete solver in a hybrid autoencoder

- ▶ A hybrid generic Generative AI that benefits from each component
- ▶ Neural Network: ideal to extract a representation of $P(X|\Phi)$ from raw inputs
- ▶ Represented as a GM in a fully explorable and controllable latent layer
- ▶ Using decoding by discrete reasoning (toulbar2 or LR-BCD)
- ▶ All this with scalable training

Future: SDP solver, hidden variables, higher-order dependencies,

A Neural Net, a Graphical Model and a discrete solver in a hybrid autoencoder

- ▶ A hybrid generic Generative AI that benefits from each component
- ▶ Neural Network: ideal to extract a representation of $P(X|\Phi)$ from raw inputs
- ▶ Represented as a GM in a fully explorable and controllable latent layer
- ▶ Using decoding by discrete reasoning (toulbar2 or LR-BCD)
- ▶ All this with scalable training

Future: SDP solver, hidden variables, higher-order dependencies,

A Neural Net, a Graphical Model and a discrete solver in a hybrid autoencoder

- ▶ A hybrid generic Generative AI that benefits from each component
- ▶ Neural Network: ideal to extract a representation of $P(X|\Phi)$ from raw inputs
- ▶ Represented as a GM in a fully explorable and controllable latent layer
- ▶ Using decoding by discrete reasoning (toulbar2 or LR-BCD)
- ▶ All this with scalable training

Future: SDP solver, hidden variables, higher-order dependencies,



Acknowledgments



AI/toulbar2

S. de Givry (INRA)
G. Katsirelos (INRA)
M. Zytnicki (PhD, INRA)
D. Allouche (INRA)
M. Ruffini (INRA)
V. Durante (ANITI, PhD)
H. Nguyen (PhD, INRA)
C. Brouard (ML, INRA)
S. Buchet (INRAE/ANITI)
P. Montalbano (ANITI, PhD)
M. Cooper (IRIT, Toulouse)
J. Larrosa (UPC, Spain)
F. Heras (UPC, Spain)
M. Sanchez (Spain)
E. Rollon (UPC, Spain)
P. Meseguer (CSIC, Spain)
G. Verfaillie (ONERA, ret.)
JH. Lee (CU. Hong Kong)
C. Bessiere (LIMM, Montpellier)
JP. Métivier (GREYC, Caen)
S. Loudni (GREYC, Caen)
M. Fontaine (GREYC, Caen),...










DL/Protein Design

A. Voet (KU Leuven)
A. Olichon (INSERM)
D. Simoncini (UFT, Toulouse)
S. Barbe (INSA, Toulouse)
M. Defresne (INRAE, PhD)
Y. Bouchiba (INSA, PhD)
C. Dumont (INSA, Toulouse)
J. Vucinic (INRA/INSA)
S. Traoré (PhD, CEA)
C. Viricel (PhD)
K. Zhang (Riken, CBDR)
S. Yagi (Riken, CBDR)
S. Tagami (Riken, CBDR)
RosettaCommons (U. Washington)
W. Sheffler (U. Washington)
V. Mulligan (Flatiron Institute, NY)
C. Bahl (IPI, Boston)
PyRosetta (U. John Hopkins)
B. Donald (U. North Carolina)
K. Roberts (U. North Carolina)
T. Simonson (Polytechnique)
J. Cortes (LAAS/CNRS),...



My apologies to those missing in these lists. Even imperfect lists seem better than no list

-  Dauparas, J. et al. (2022). “Robust deep learning-based protein sequence design using ProteinMPNN”. In: [Science](#) 378.6615, pp. 49–56. DOI: [10.1126/science.add2187](#).
-  Defresne, Marianne et al. (2023). “Scalable Coupling of Deep Learning with Logical Reasoning”. In: [32nd International Joint Conference on Artificial Intelligence, IJCAI 2023](#). Macao, SAR, China: [ijcai.org](#), pp. 3615–3623. DOI: [10.24963/IJCAI.2023/402](#).
-  Durante, Valentin et al. (July 2022). “Efficient low rank convex bounds for pairwise discrete Graphical Models”. In: [Thirty-ninth International Conference on Machine Learning](#).
-  Montalbano, Pierre et al. (2022). “Multiple-Choice Knapsack Constraint in Graphical Models”. In: [Proc. of CPAIOR’22](#).
-  Rocklin, Gabriel J et al. (2017). “Global analysis of protein folding using massively parallel design, synthesis, and testing”. In: [Science](#) 357.6347, pp. 168–175.
-  Vucinic, Jelena et al. (2020). “Positive multistate protein design”. In: [Bioinformatics](#) 36.1, pp. 122–130.
-  Yagi, Sota et al. (2021). “Seven amino acid types suffice to create the core fold of RNA polymerase”. In: [Journal of the American Chemical Society](#) 143.39, pp. 15998–16006.