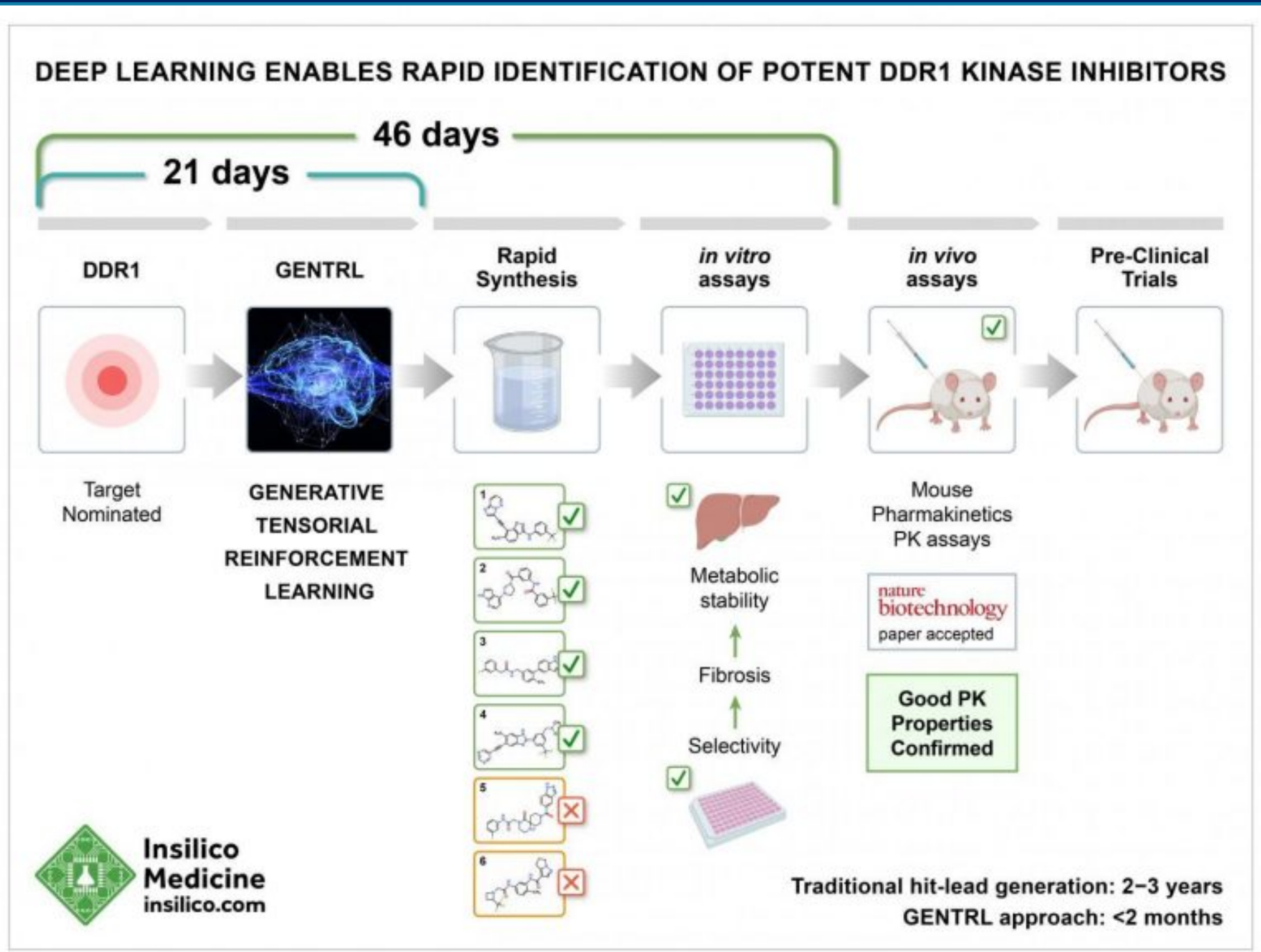
Modern AI for Drug Discovery

Truyen Tran
Deakin University

truyen.tran@deakin.edu.au
truyentran.github.io
@truyenoz
letdataspeak.blogspot.com
goo.gl/3jJ1O0

HCM City, Nov 2019



re.com/articles/s41587-019-0224-x

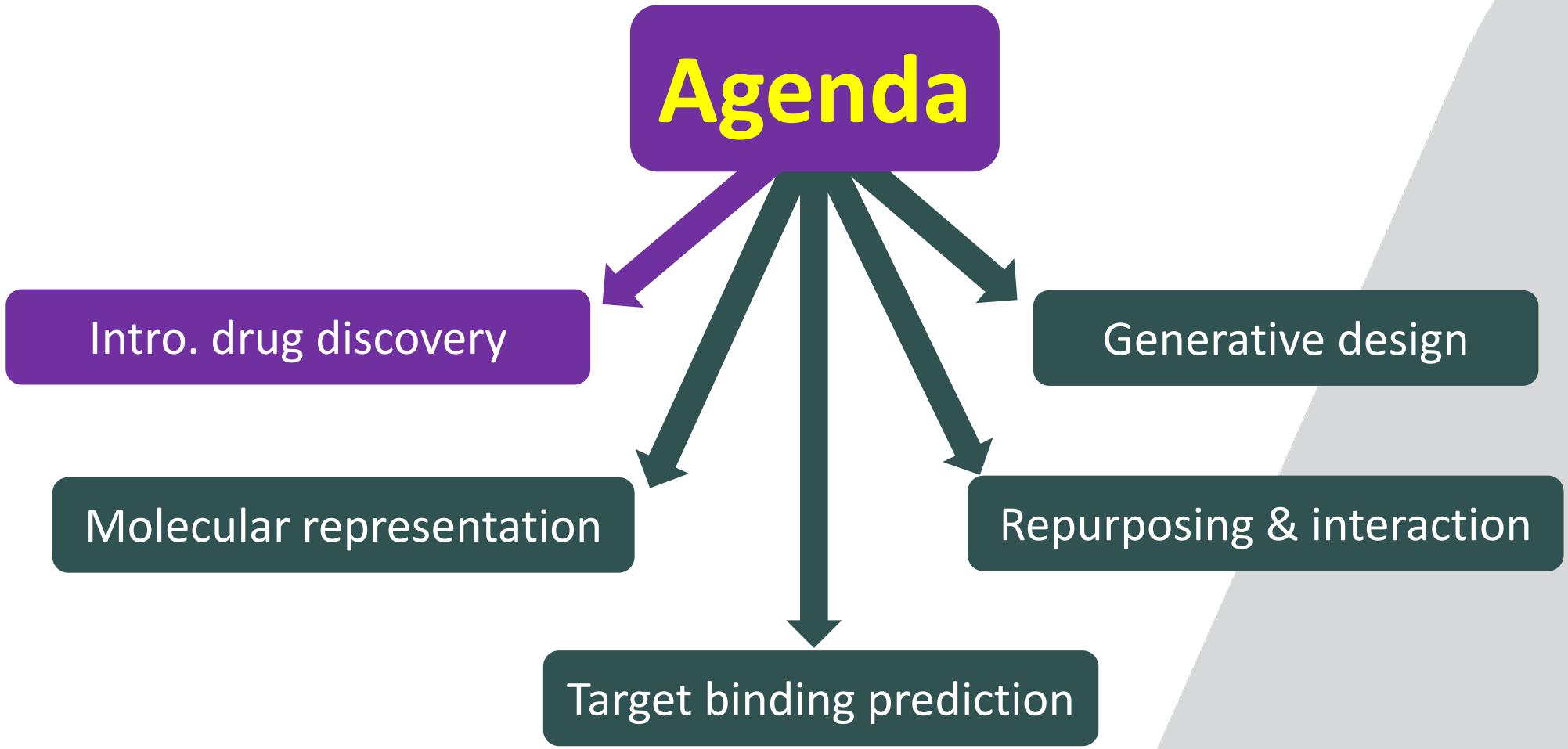
ning enables rapid
tion of potent DDR1
hibitors

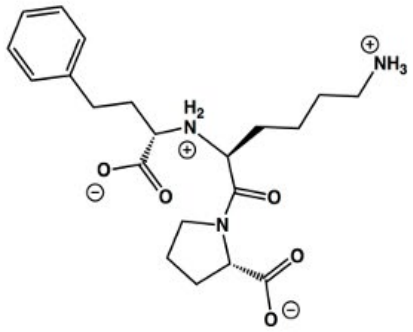
n A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A.
Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy,
ip Asadulaev, Yury Volkov, Artem Zholus, Rim R.
er Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy,
Soll, David Madge, Li Xing, Tao Guo & Alán Aspuru-Guzik

7, 1038–1040 (2019) | Download Citation
tions | 1701 Altmetric | Metrics »

d a deep generative model, generative
ement learning (GENTRL), for de novo
esign. GENTRL optimizes synthetic
, and biological activity. We used GENTRL
t inhibitors of discoidin domain receptor 1
rget implicated in fibrosis and other
s. Four compounds were active in
s, and two were validated in cell-based
andidate was tested and demonstrated
okinetics in mice.

Agenda


Intro. drug discovery
Generative design
Molecular representation
Repurposing & interaction
Target binding prediction

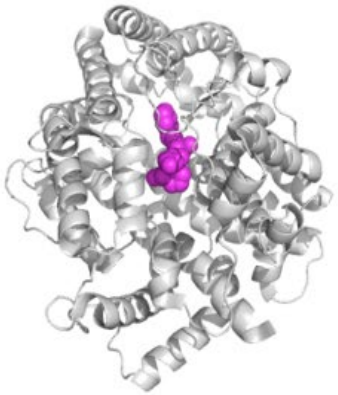


free drug ligand

+

free target protein

⇌



bound drug-target co-complex

Source: c4xdiscovery.com

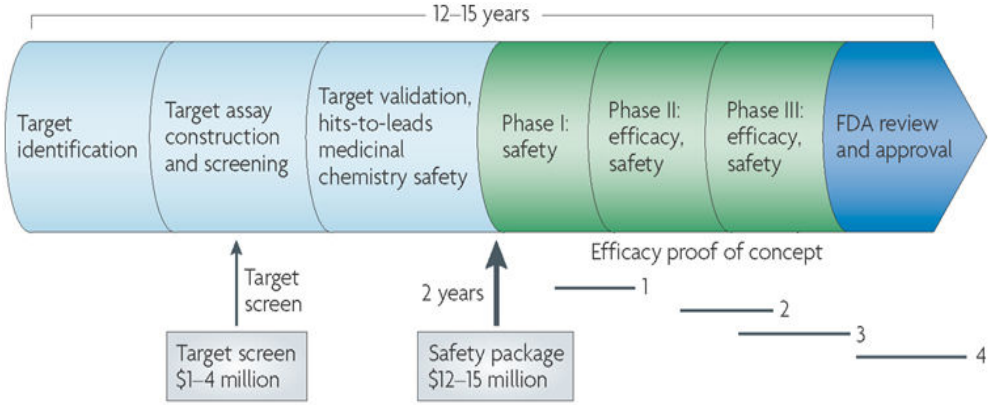
Drug-likeness:

- Solubility in water and fat, e.g., measured by LogP. Most drugs are admitted orally → pass through membrance.
- Potency at the bio target → target-specific binding.
- Ligand efficiency (low energy binding) and lipophilic efficiency.
- Small molecular weight → affect diffusion
- Rule of Five

- Drug is a small molecule that binds to a bio target (e.g., protein) and modifies its functions to produce useful physiological or mental effects.
  - Proteins are large biomolecules consisting of chains of amino acid residues.
    - **Drug discovery** is the process through which potential new medicines are identified. It involves a wide range of scientific disciplines, including **biology**, **chemistry** and **pharmacology** (*Nature, 2019*).

Knowledge-driven

$500M - $2B



Nature Reviews | Drug Discovery

→ thousands of small molecules → a few lead-like molecules → one in ten of these molecules pass clinical trials in human patients.

#REF: Roses, Allen D. "Pharmacogenetics in drug discovery and development: a translational perspective." *Nature reviews Drug discovery* 7.10 (2008): 807-817.

AI-driven

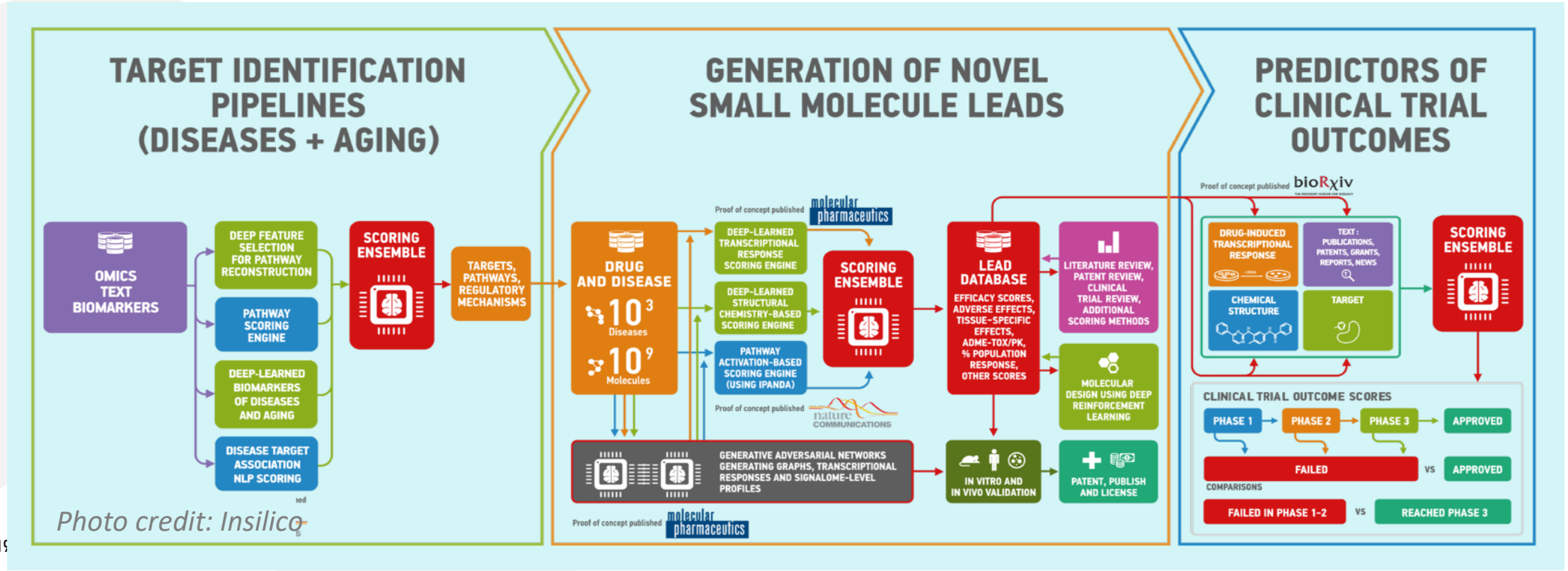


Photo credit: Insilico

The three basic questions

Given a molecule, is this drug? Aka properties/targets/effects prediction.

- Drug-likeness
- Targets it can modulate and how much
- Its dynamics/kinetics/effects/metabolism if administered orally or via injection

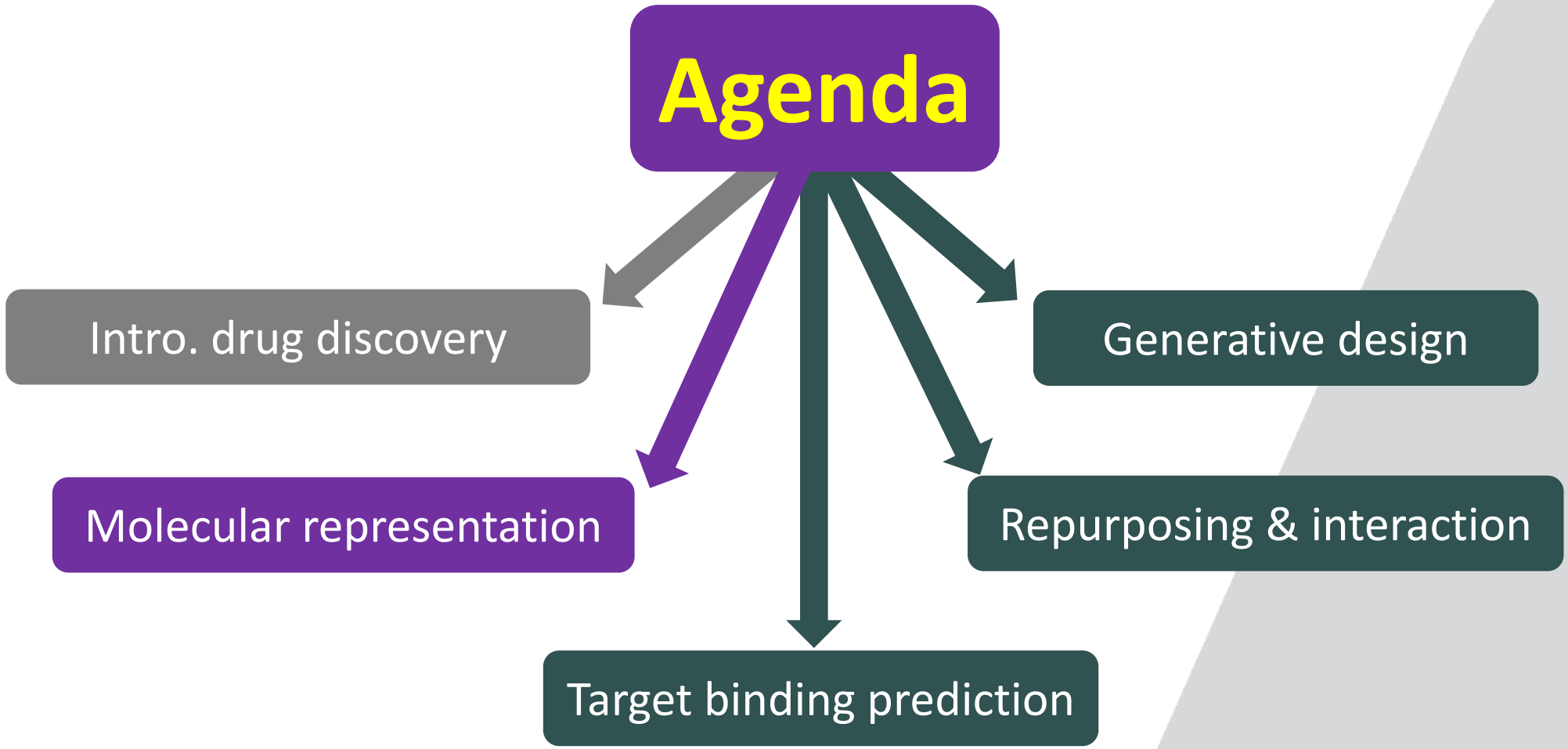
Given a target, what are molecules?

- If the list of molecules is given, pick the good one. If evaluation is expensive, need to search, e.g., using BO.
- If no molecule is found, need to generate from scratch → generative models + BO, or RL.
- How does the drug-like space look like?

Given a molecular graph, what are the steps to make the molecule?

- Synthetic tractability
- Reaction planning, or retrosynthesis

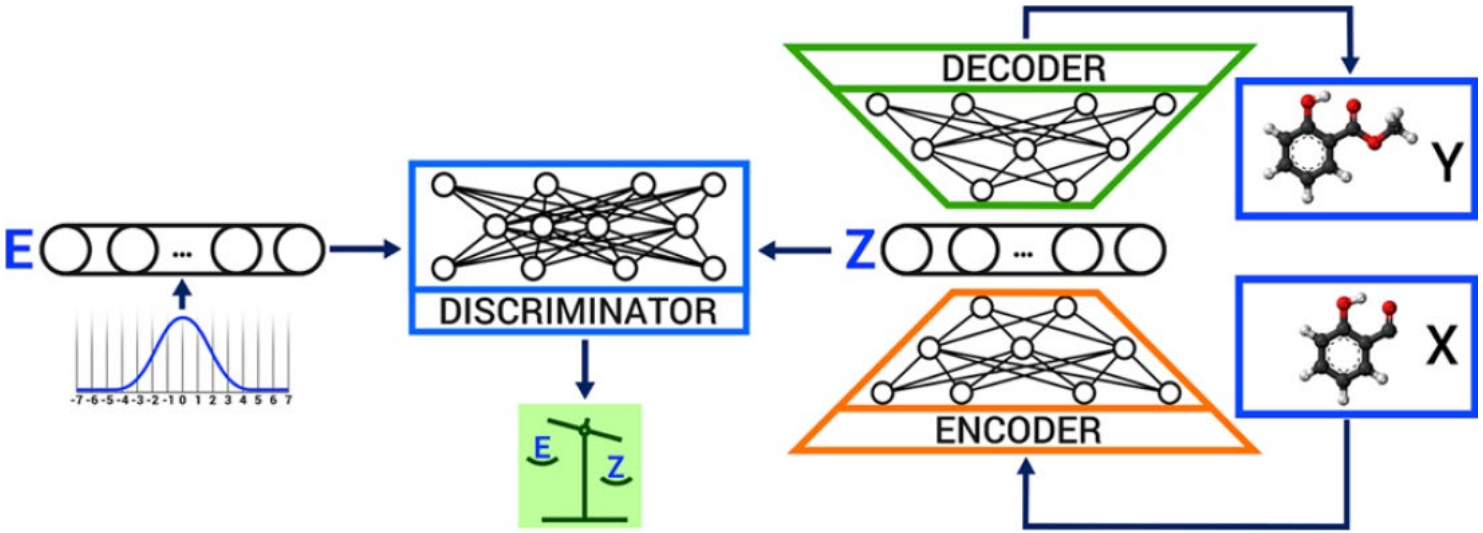
Agenda


Intro. drug discovery
Molecular representation
Target binding prediction
Repurposing & interaction
Generative design

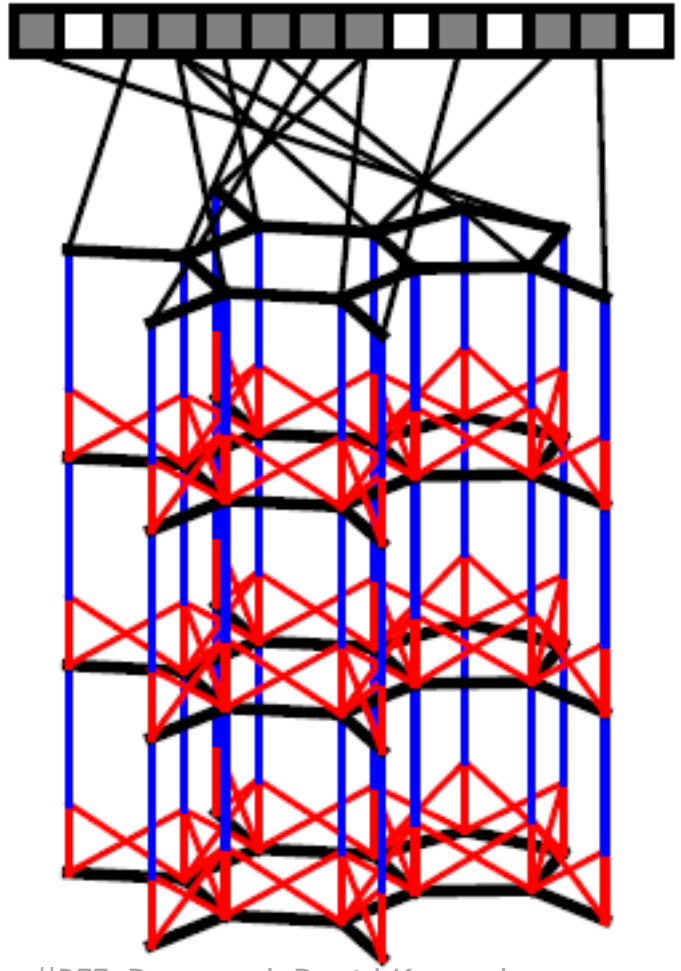
Molecule → fingerprints

Graph → vector. Mostly discrete. Substructures coded.

Vectors are easy to manipulate. Not easy to reconstruct the graphs from fingerprints.



Kadurin, Artur, et al. "The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology." *Oncotarget* 8.7 (2017): 10883.



#REF: Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." *Advances in neural information processing systems*. 2015.

Molecule → string

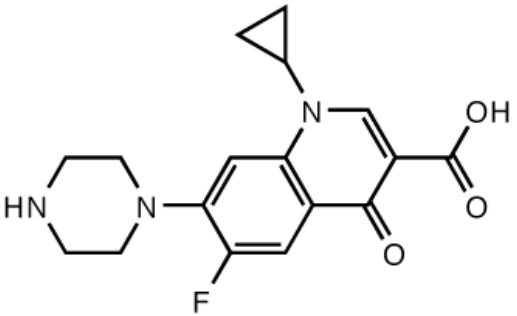
SMILES = Simplified Molecular-Input Line-Entry System

Ready for encoding/decoding with sequential models (seq2seq, MANN, RL).

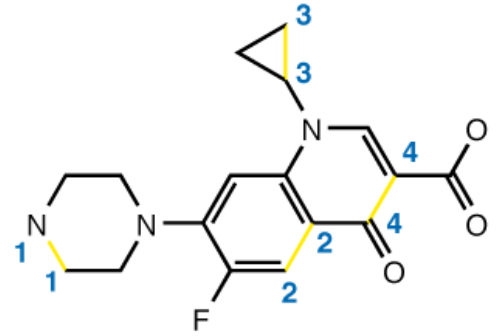
BUT …

- String → graphs is not unique!
- Lots of string are invalid
- Precise 3D information is lost
- Short range in graph may become long range in string

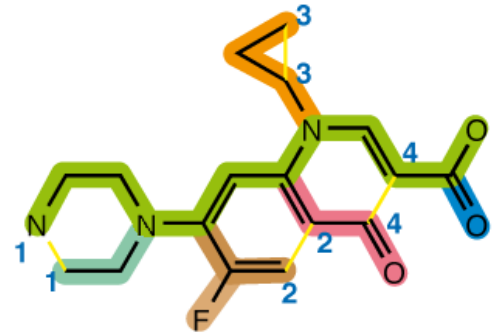
A



B



C



D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Source: wikipedia.org

#REF: Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *arXiv preprint arXiv:1610.02415* (2016).

Molecule → graphs

No regular, fixed-size structures

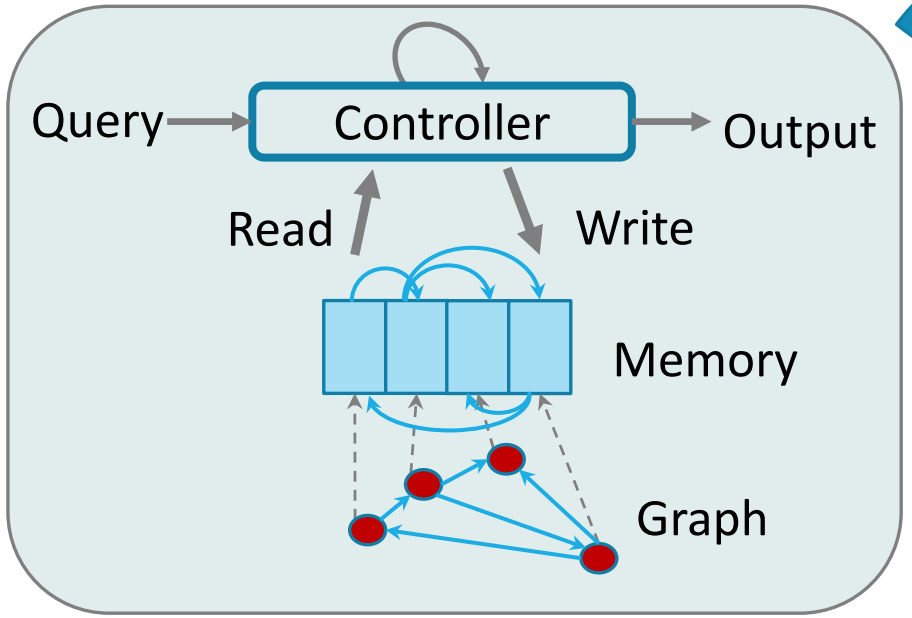
Graphs are ***permutation invariant***:

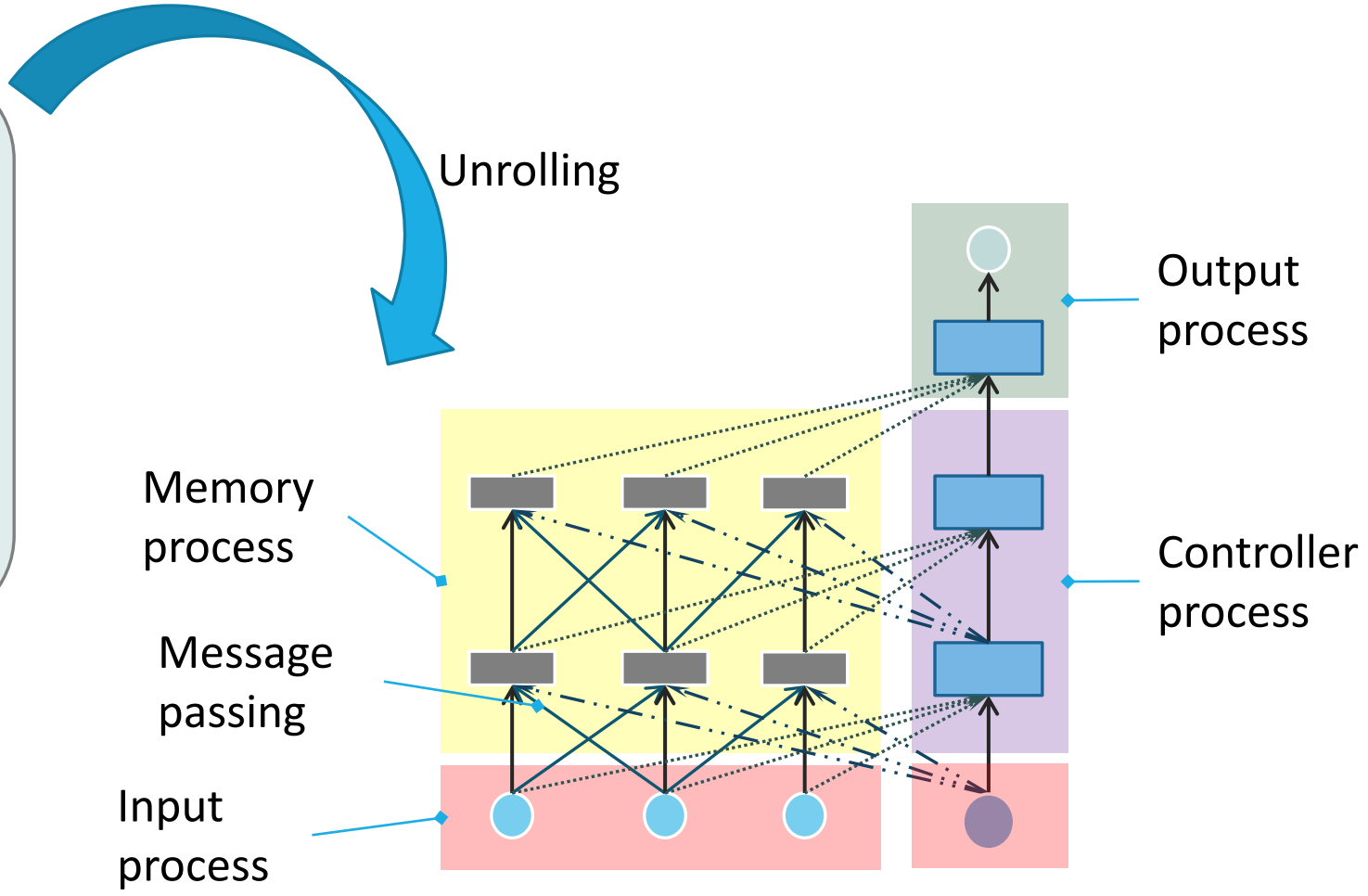
- #permutations are exponential function of #nodes
- The probability of a generated graph G need to be marginalized over all possible permutations

Multiple objectives:

- **Diversity** of generated graphs
- **Smoothness** of latent space
- Agreement with or optimization of multiple "**drug-like**" objectives

RDMN: A graph processing machine





#REF: Pham, T., Tran, T., & Venkatesh, S. (2018). Relational dynamic memory networks. *arXiv preprint arXiv:1808.04247*.

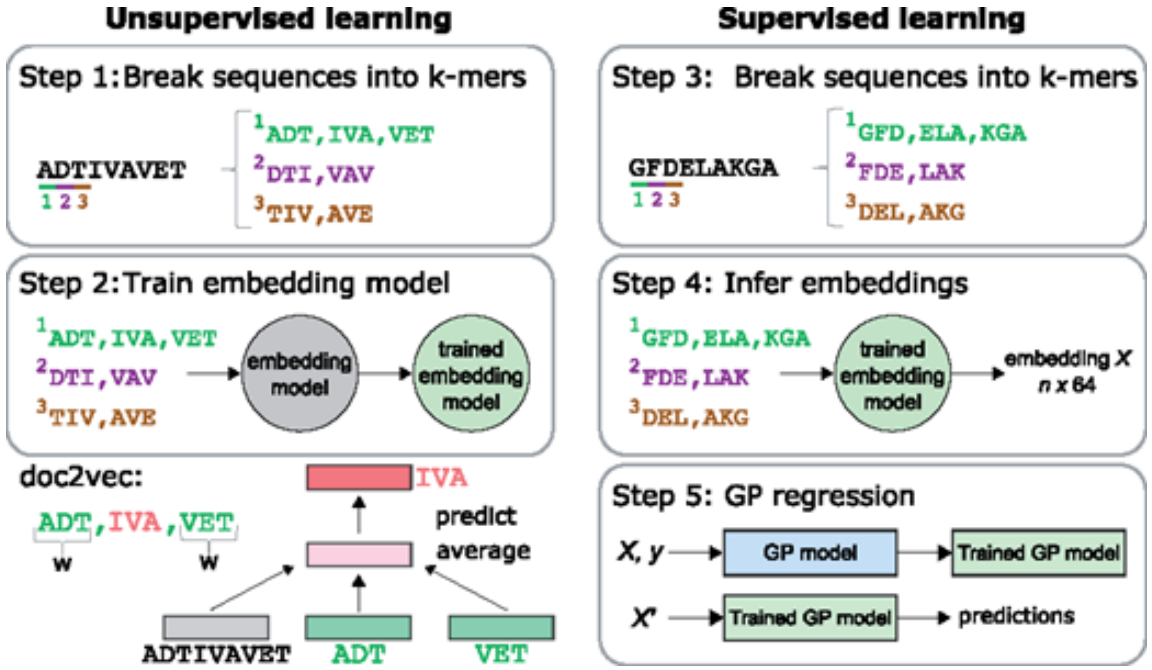
Representing proteins

1D sequence (vocab of size 20) – hundreds to thousands in length

2D contact map – requires prediction

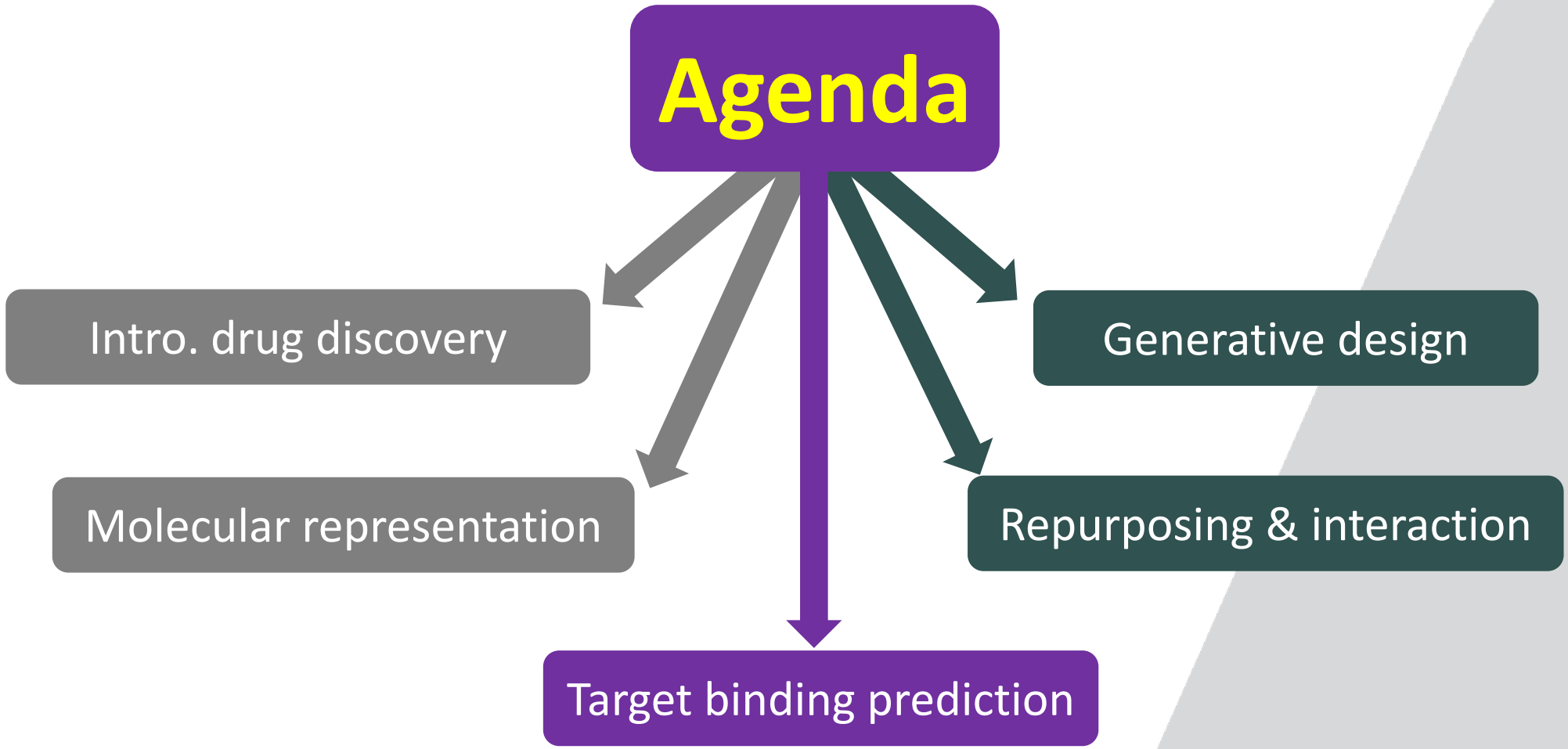
3D structure – requires folding information, either observed or predicted. Only a limited number of 3D structures are known.

NLP-inspired embedding (word2vec, doc2vec, glove, seq2vec, ELMo, BERT, etc).



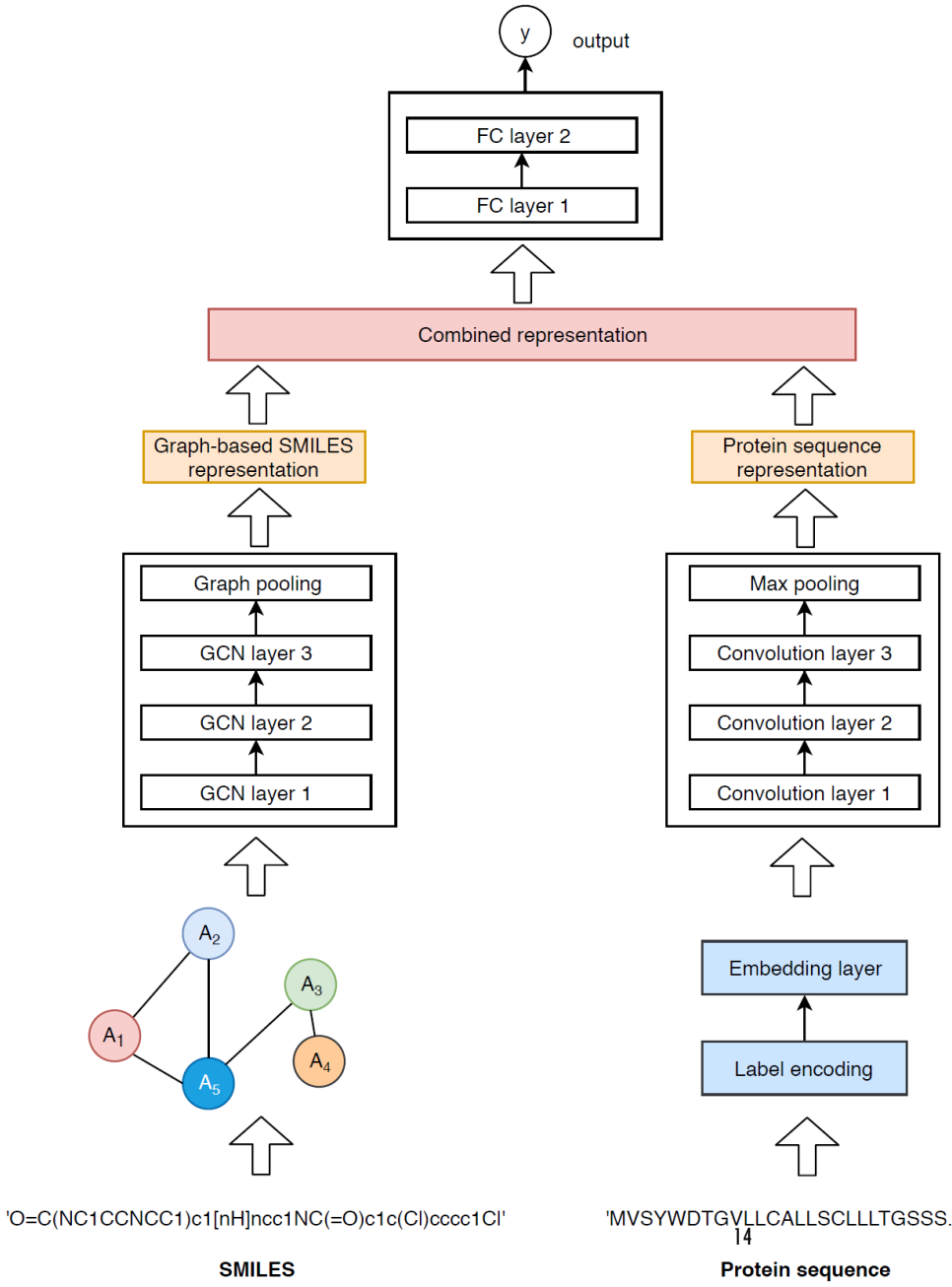
#REF: Yang, K. K., Wu, Z., Bedbrook, C. N., & Arnold, F. H. (2018). Learned protein embeddings for machine learning. *Bioinformatics*, *34*(15), 2642-2648.

Agenda


Intro. drug discovery
Molecular representation
Target binding prediction
Generative design
Repurposing & interaction

Drug-target binding as QA

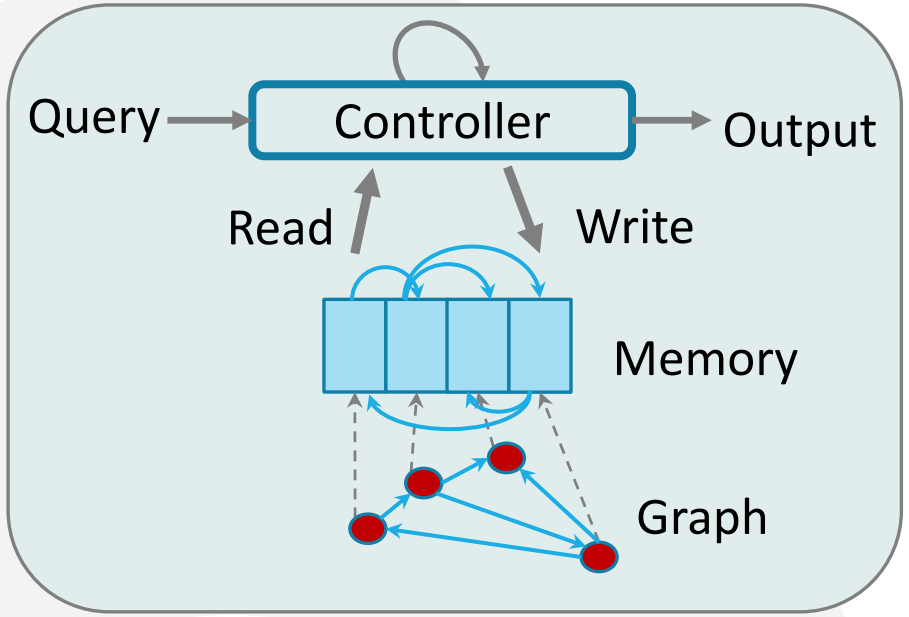
- Context: Binding targets (e.g., RNA/protein sequence, or 3D structures), as a set, sequence, or graph.
- Query: Drug (e.g., SMILES string, or molecular graph)
- Answer: Affinity, binding sites, modulating effects

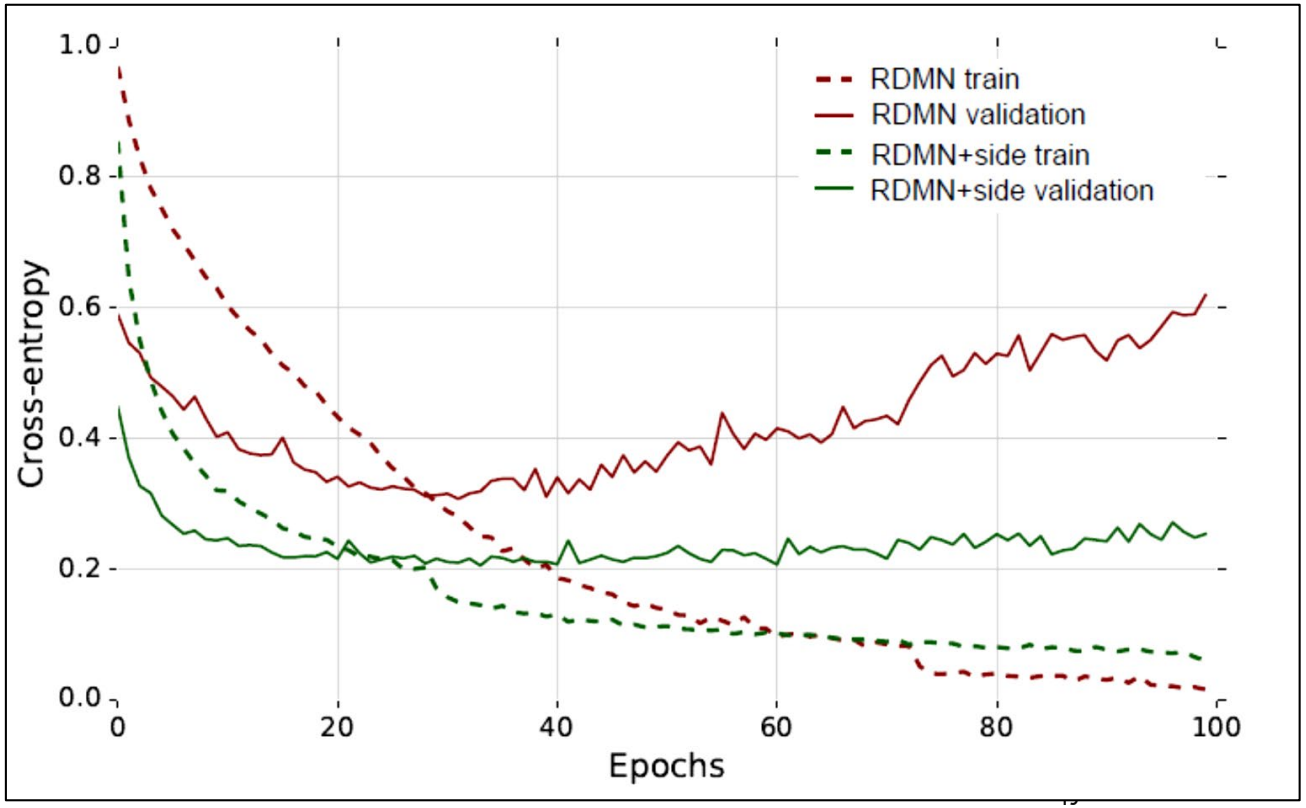


#REF: Nguyen, T., Le, H., & Venkatesh, S. (2019). GraphDTA: prediction of drug–target binding affinity using graph convolutional networks. *BioRxiv*, 684662.

More flexible drug-disease response with RDMN

Model	MicroF1	MacroF1	Average AUC
SVM	66.4	67.9	85.1
RF	65.6	66.4	84.7
GB	65.8	66.9	83.7
NeuralFP [19]	68.2	67.6	85.9
MT-NN [51]	75.5	78.6	90.4
RDMN	**77.8**	**80.3**	**92.1**

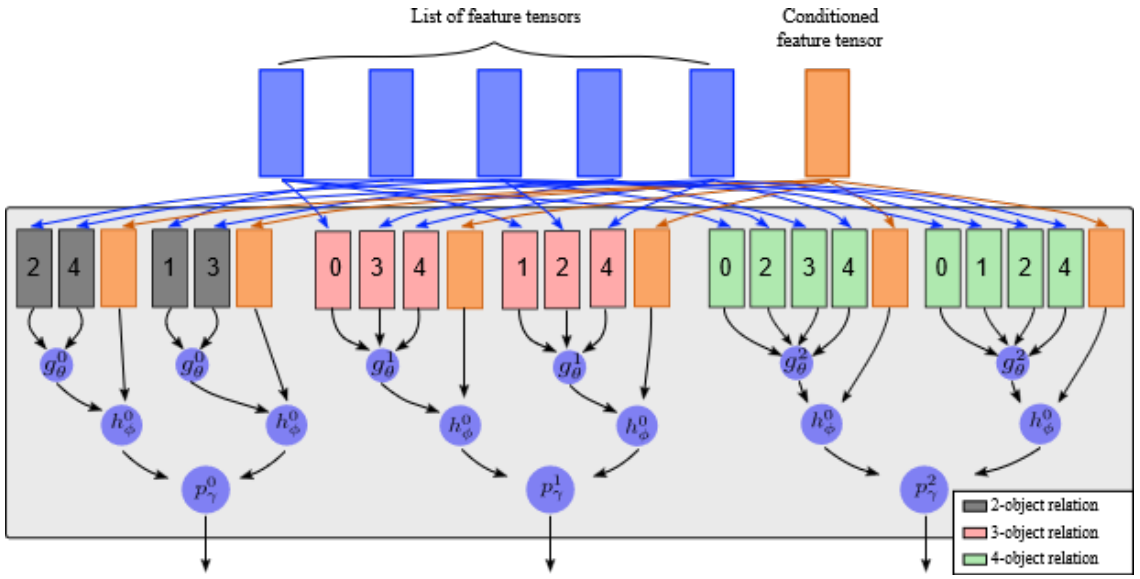


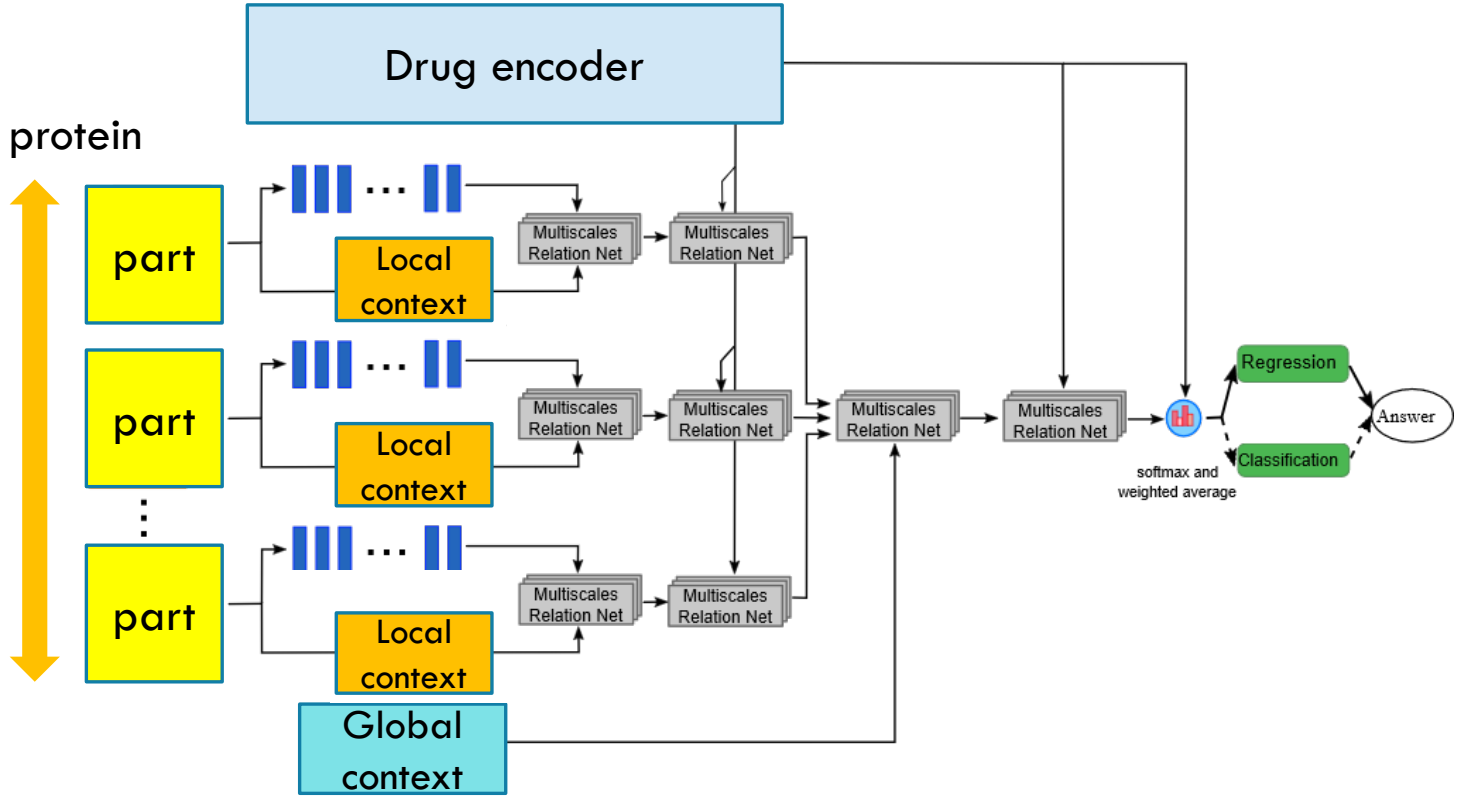


Drug-target binding as QA (2) - on-going work

Random relation unit

- Object-object interaction
- Objects-context interaction
- Shallow hierarchy



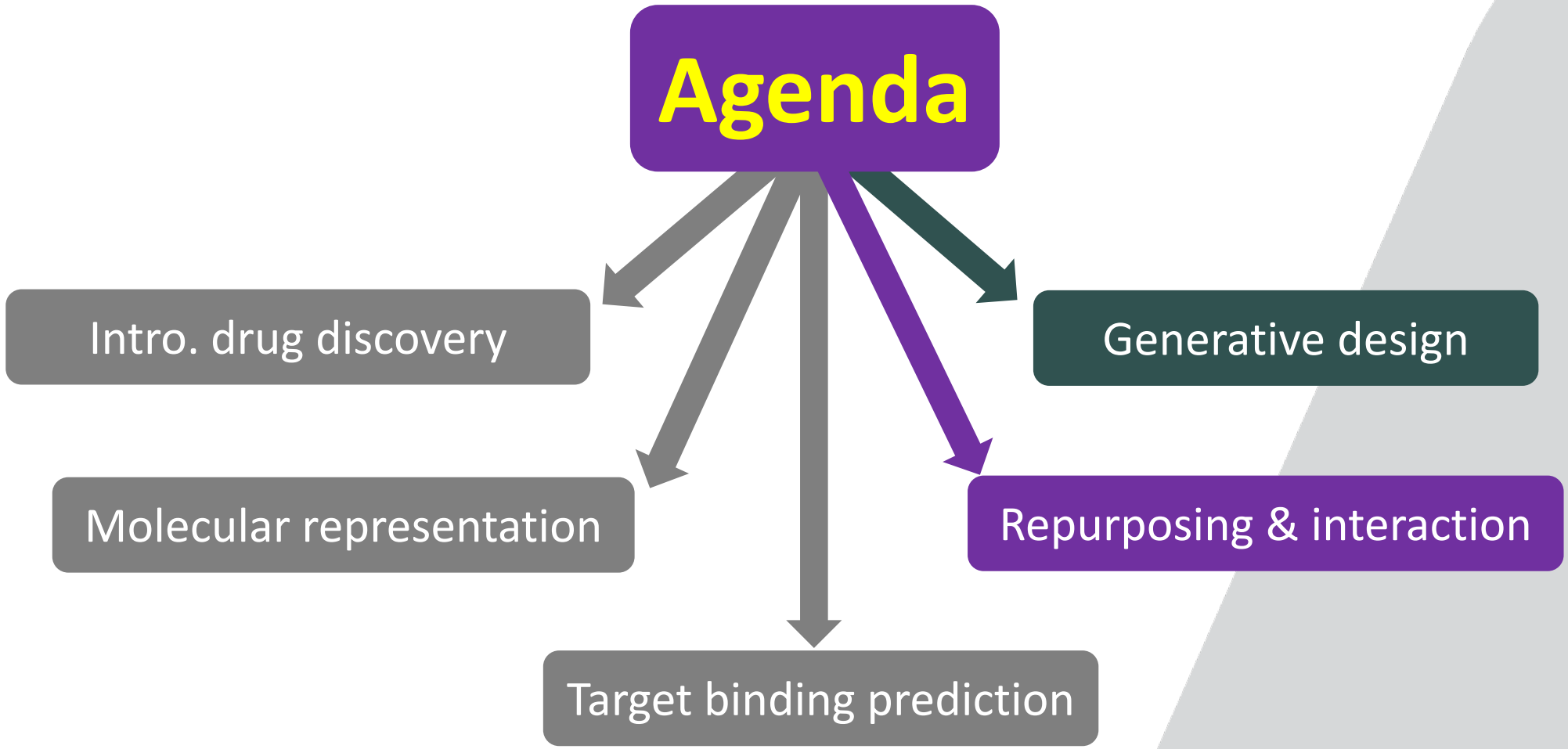


Protein as hierarchical random powerset

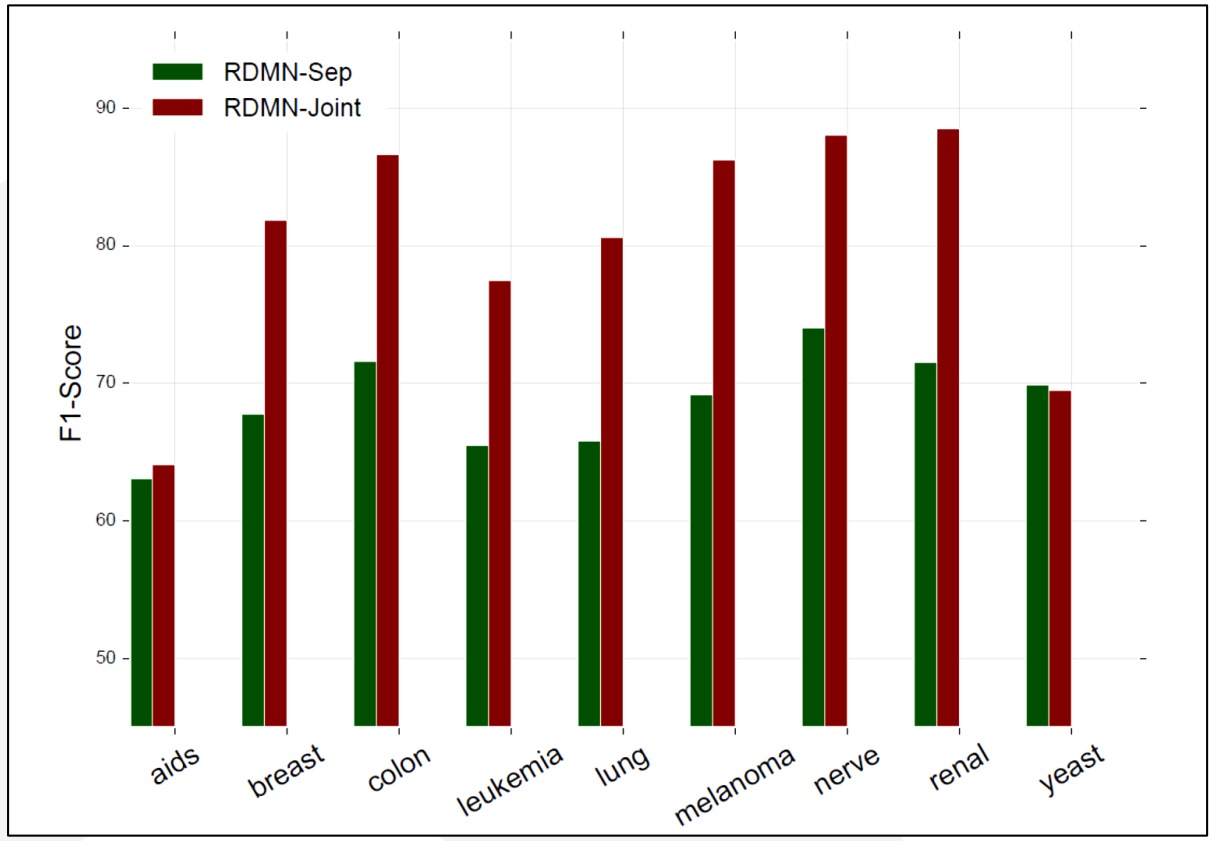
Bypassing:

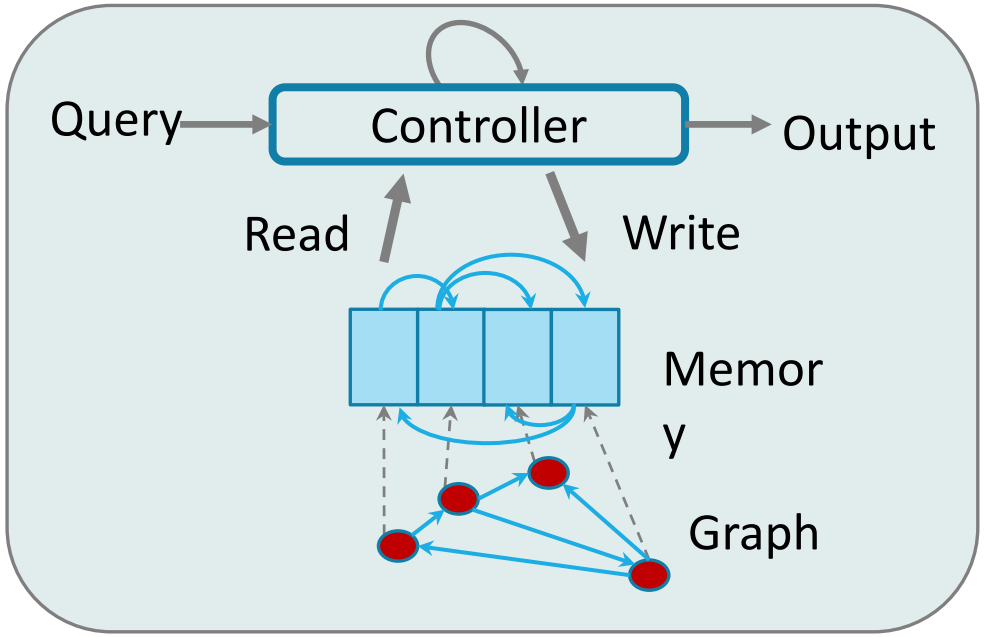
- Protein folding estimation
- Binding site estimation

Agenda

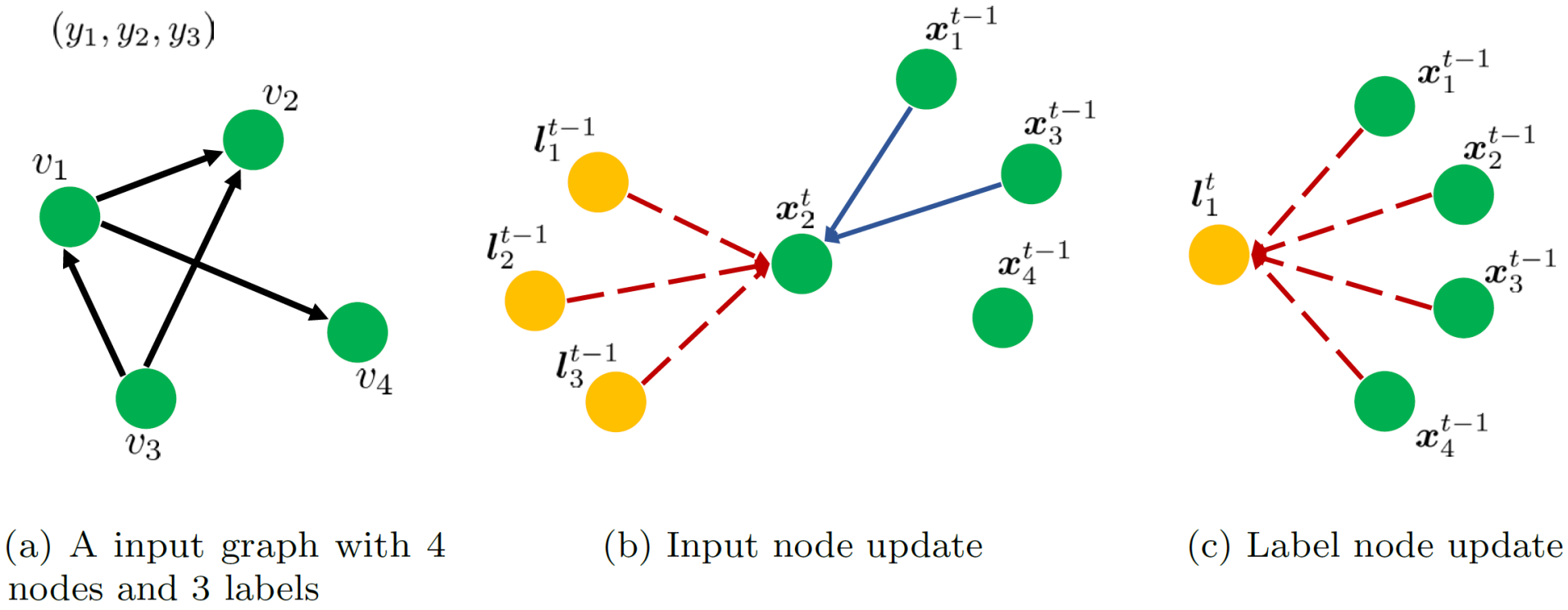

Intro. drug discovery
Molecular representation
Target binding prediction
Generative design
Repurposing & interaction

Tying param helps multiple diseases response with RDMN





GAML: Repurposing as multi-target prediction



(a) A input graph with 4 nodes and 3 labels

(b) Input node update

(c) Label node update

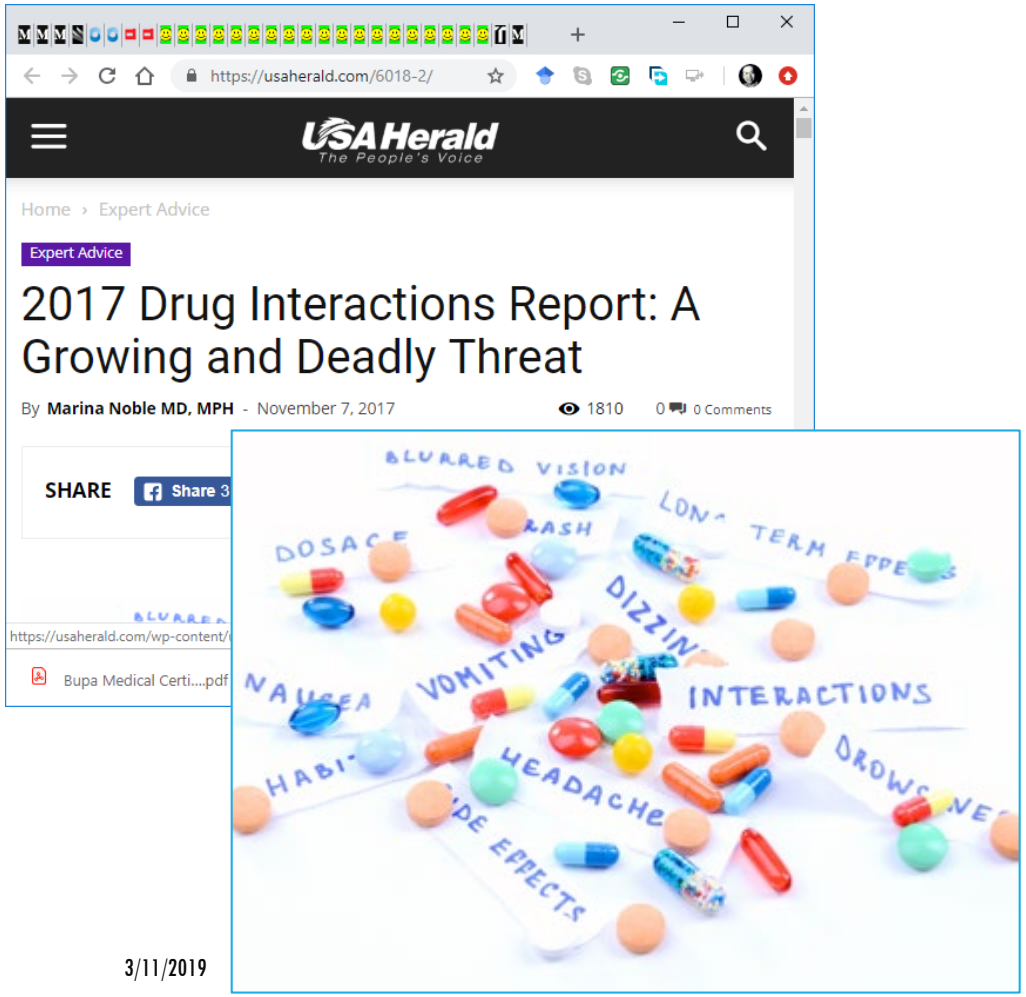
#REF: Do, Kien, et al. "Attentional Multilabel Learning over Graphs-A message passing approach." *Machine Learning, 2019*.

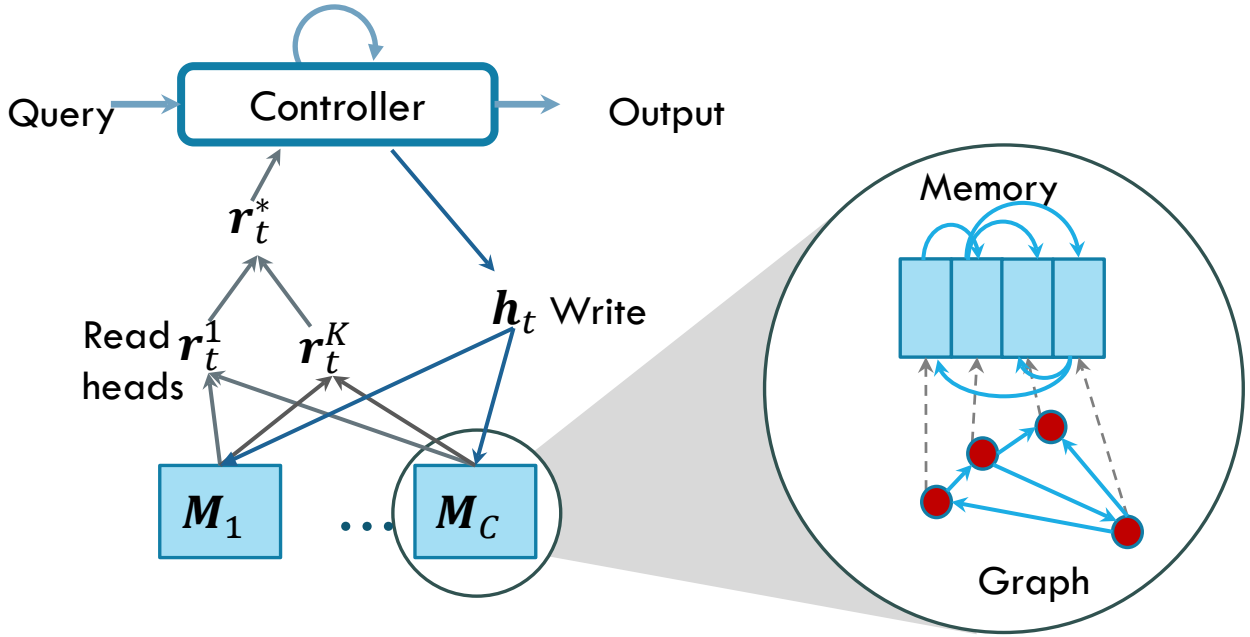
Dataset	Metrics	Fingerprint		SMILES	Molecular Graph		
		SVM	HWN	GRU	WL+SVM	CLN	GAML
9cancers	m-AUC	81.94	85.95	83.29	86.06	88.35	**88.78**
	M-AUC	81.37	85.85	82.74	85.74	88.23	**88.50**
	m-F1	50.63	57.44	55.97	54.55	59.48	**62.03***
	M-F1	50.71	57.29	55.99	54.54	59.50	**62.14***
50proteins	m-AUC	79.85	77.46	79.11	81.62	82.08	**82.82**
	M-AUC	74.77	73.78	75.25	77.60	78.36	**79.35***
	m-F1	17.21	16.37	16.08	17.04	18.37	**20.47***
	M-F1	18.40	15.87	14.96	18.66	17.72	**19.83***

Table 4: The performance in the multi-label classification with graph-structured input (m-X: micro average of X; M-X: macro average). SVM and HWN work on fingerprint representation; GRU works on string representation of molecule known as SMILES; WL+BR and CLN work directly on graph representation. Bold indicates better values. (*) $p < 0.05$.

#REF: Do, Kien, et al. "Attentional Multilabel Learning over Graphs-A message passing approach." *arXiv preprint arXiv:1804.00293*(2018).

Drug-drug interaction via RDMN





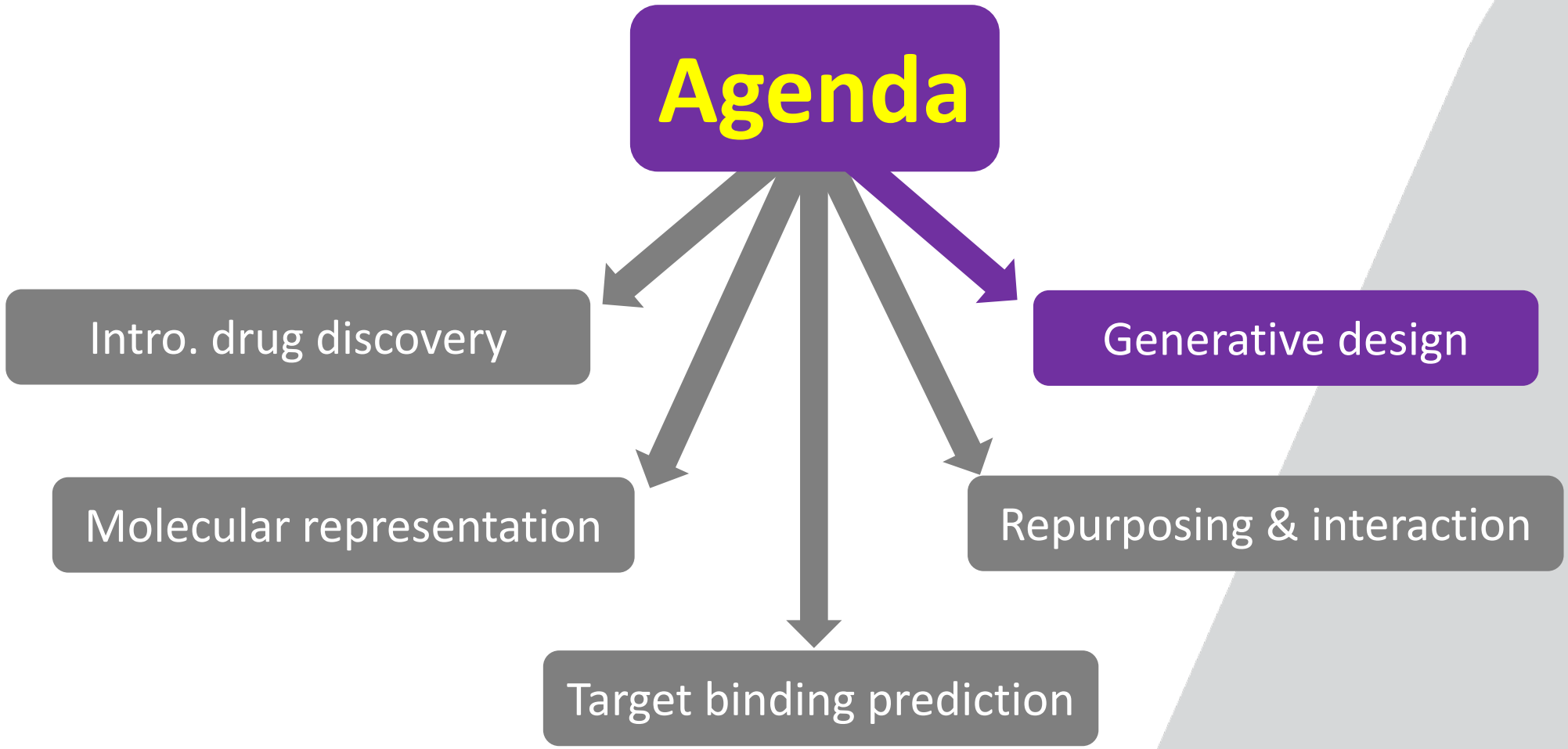
#REF: Pham, Trang, Truyen Tran, and Svetha Venkatesh. "Relational dynamic memory networks." *arXiv preprint arXiv:1808.04247*(2018).

Results on STITCH database

	CCI900		CCI800	
	AUC	F1-score	AUC	F1-score
Random Forests	94.3	86.4	98.2	94.1
Highway Networks	94.7	88.4	98.5	94.7
DeepCCI [31]	96.5	92.2	99.1	97.3
RDMN	96.6	92.6	99.1	97.4
RDMN+multiAtt	97.3	93.4	99.1	97.8
RDMN+FP	97.8	93.3	99.4	98.0
RDMN+multiAtt+FP	98.0	94.1	99.5	98.1
RDMN+SMILES	98.1	94.3	99.7	97.8
RDMN+multiAtt+SMILES	**98.1**	**94.6**	**99.8**	**98.3**

Table 3 The performance on the CCI datasets reported in AUC and F1-score. *FP* stands for fingerprint and *multiAtt* stands for multiple attentions.

Agenda


Intro. drug discovery
Generative design
Molecular representation
Repurposing & interaction
Target binding prediction

Drug design as structured machine translation, aka conditional generation

Can be formulated as structured machine translation:

- Inverse mapping of (knowledge base + binding properties) to (query) → One to many relationship.

Representing graph as string (e.g., SMILES), and use sequence VAEs or GANs.

Generative graph models

- Model nodes & interactions
- Model cliques

Sequences

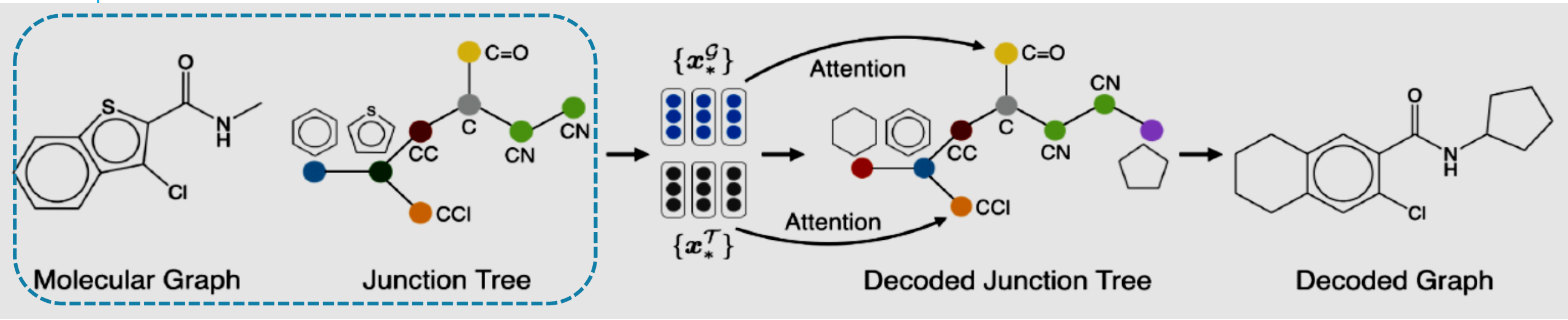
- Iterative methods

Reinforcement learning

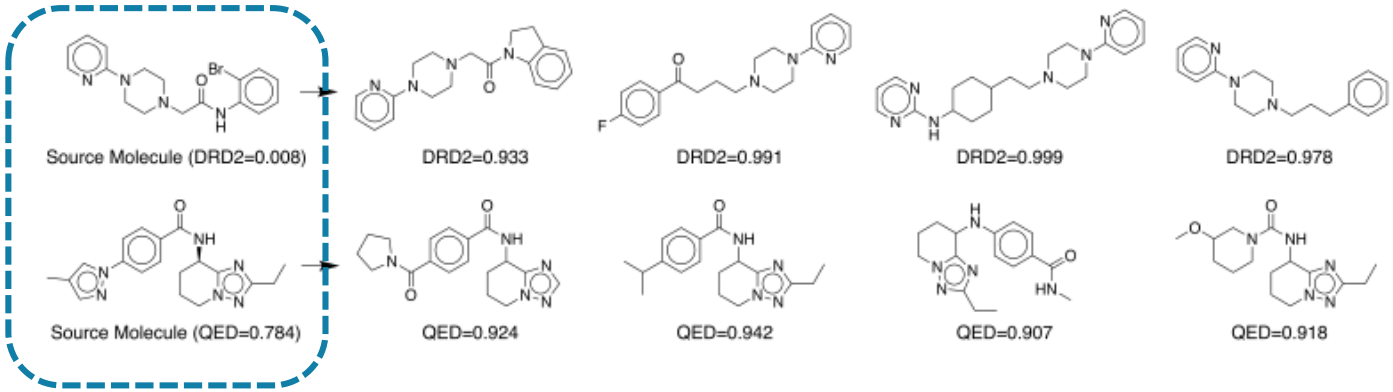
- Discrete objectives

Any combination of these + memory.

Molecular optimization as machine translation

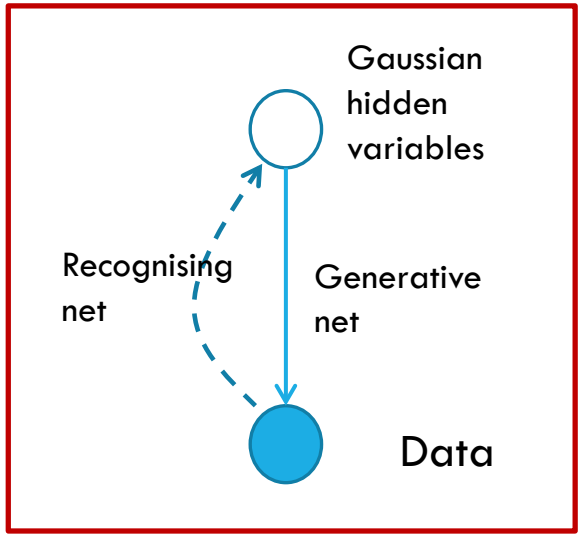


- The molecular space: up to 10^{60}
- It is easier to modify existing molecules, aka "*molecular paraphrases*"
- Molecular optimization as graph-to-graph translation



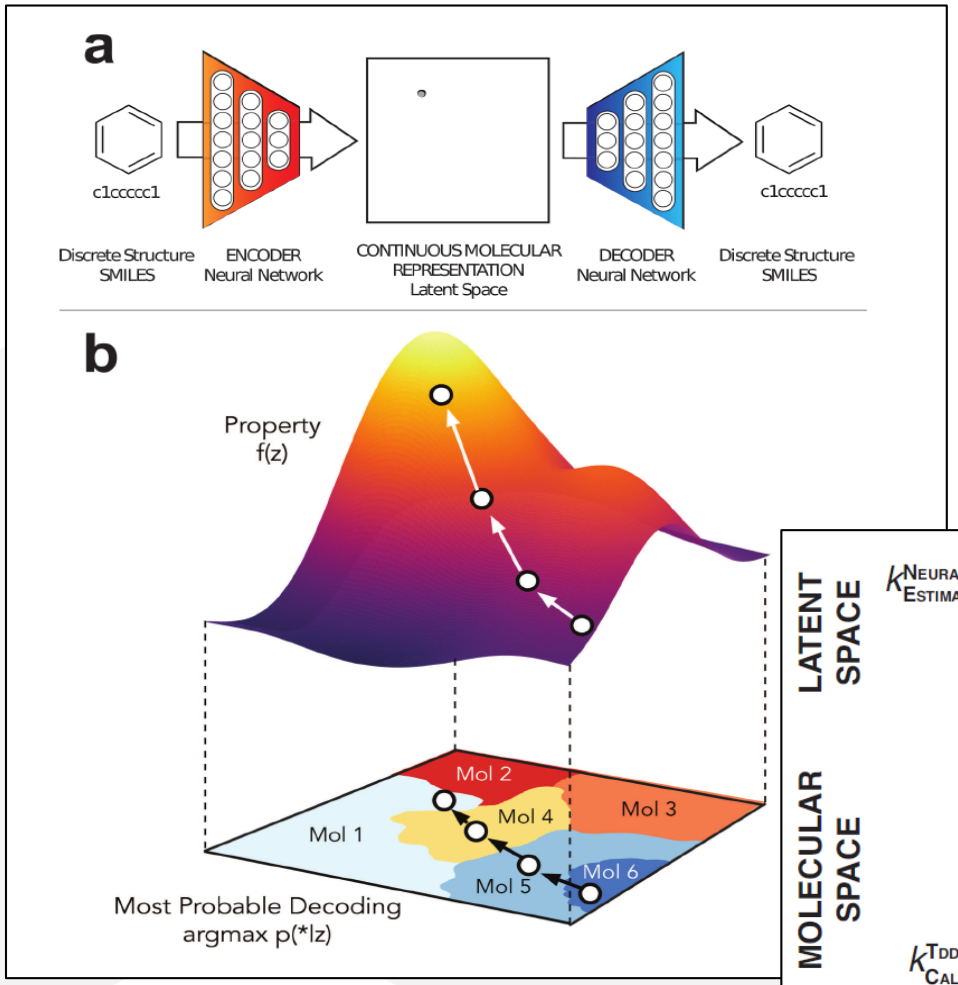
#REF: Jin, W., Yang, K., Barzilay, R., & Jaakkola, T. (2019). Learning multimodal graph-to-graph translation for molecular optimization. *ICLR*.

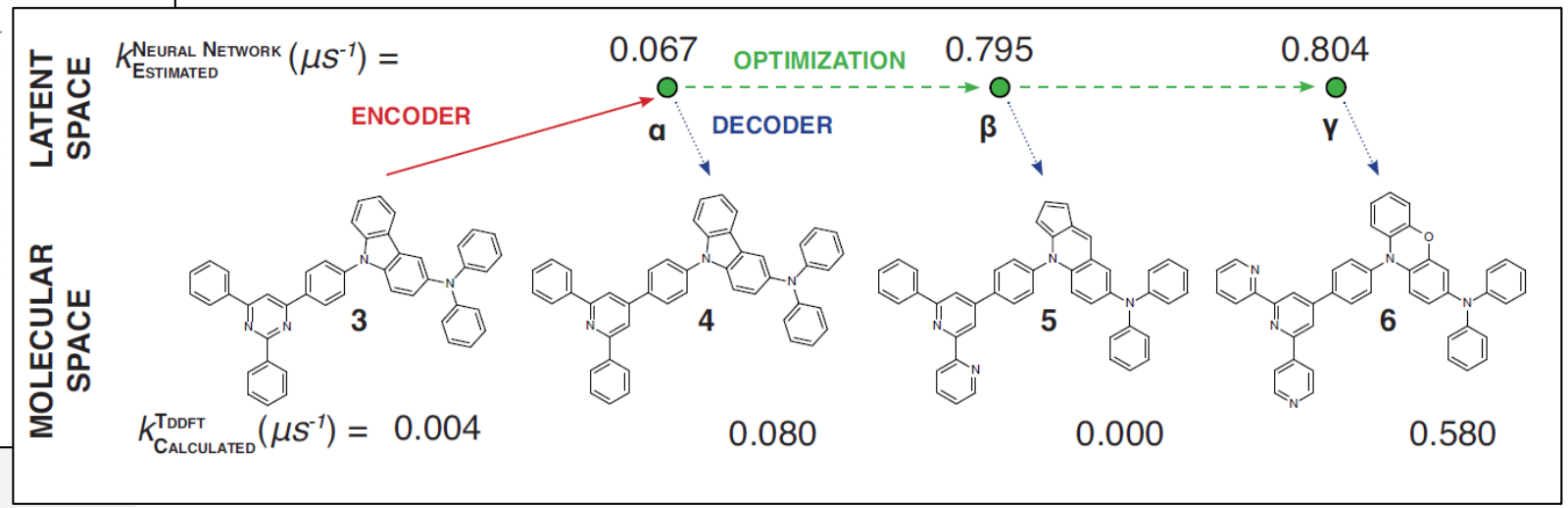
VAE for drug space modelling



Model: SMILES → VAE+RNN

#REF: Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS Central Science* (2016).



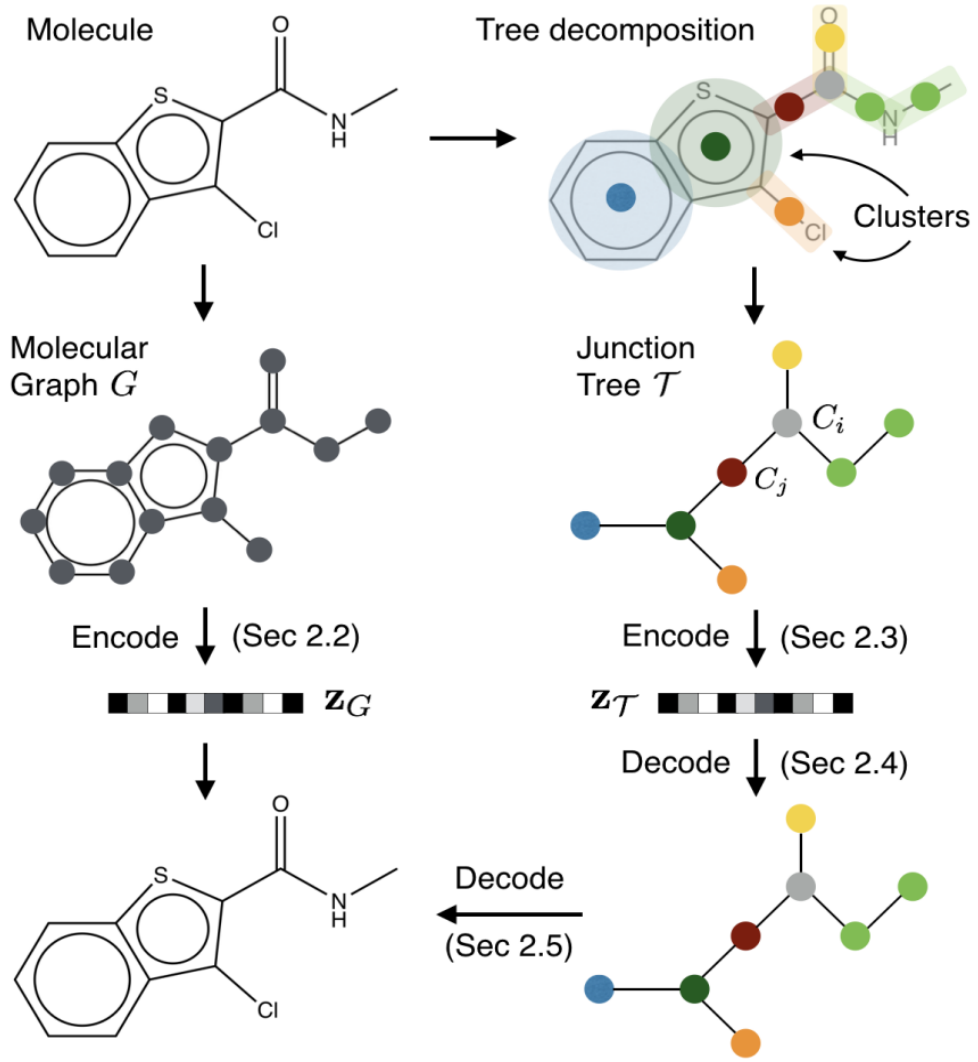


Junction tree VAE

Junction tree is a way to build a “thick-tree” out of a graph

Cluster vocab:

- rings
- bonds
- atoms



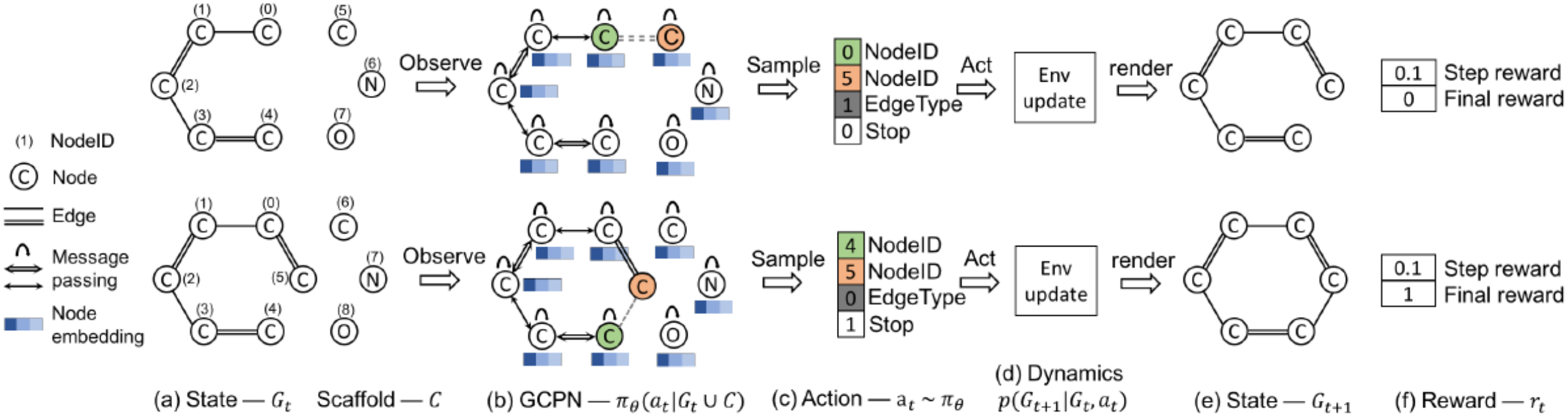
Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. *ICML'18*.

Graphs + Reinforcement learning

Generative graphs are very hard to get it right: The space is too large!

Reinforcement learning offers step-wise construction: one piece at a time

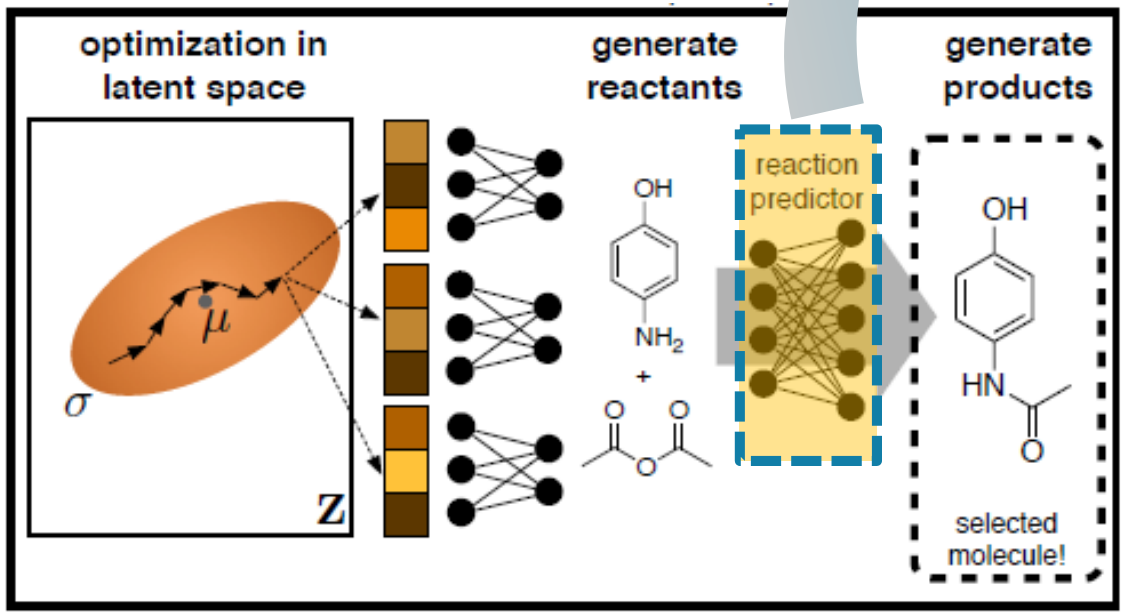
- A.k.a. Markov decision processes
- As before: Graphs offer properties estimation



You, Jiaxuan, et al. "Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation." *NeurIPS* (2018).

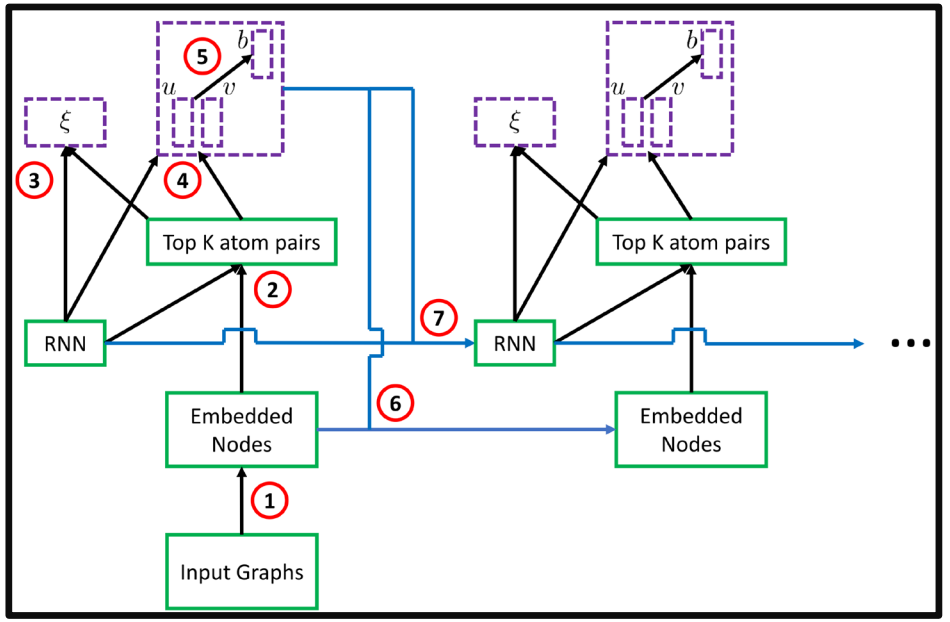
Searching for synthesizable molecules

MoleculeChef



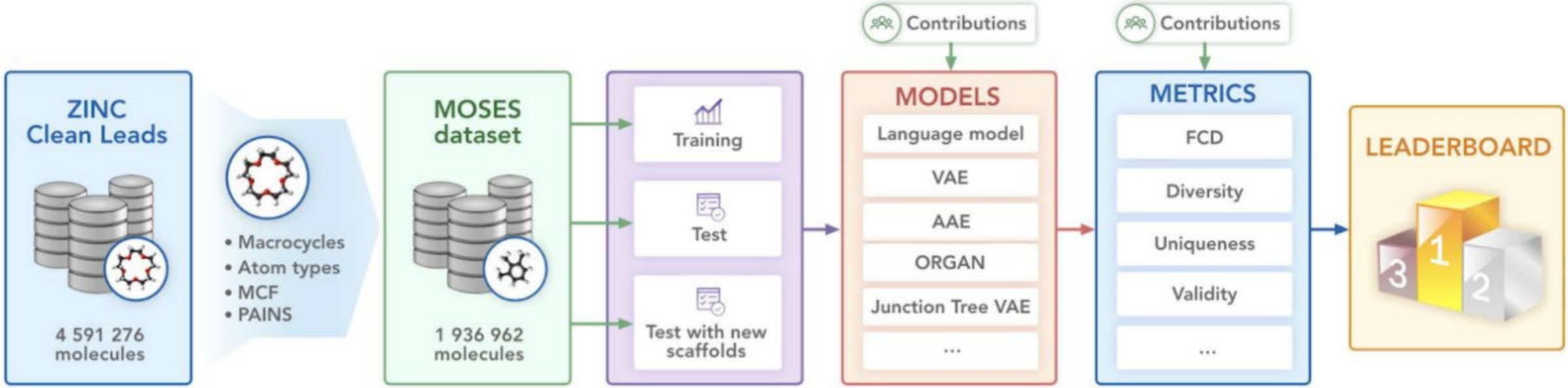
#REF: Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H., & Hernández-Lobato, J. M. (2019). A Model to Search for Synthesizable Molecules. *arXiv preprint arXiv:1906.05221*.

GTPN – reaction predictor



#REF: Do, K., Tran, T., & Venkatesh, S. (2019, July). Graph transformation policy network for chemical reaction prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 750-760). ACM.

Play ground: MOSES



https://medium.com/neuromation-io-blog/moses-a-40-week-journey-to-the-promised-land-of-molecular-generation-78b29453f75c

Thank you

Truyen Tran

truyen.tran@deakin.edu.au

truyentran.github.io

@truyenoz

letdataspeak.blogspot.com

goo.gl/3jJ1O0

APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY