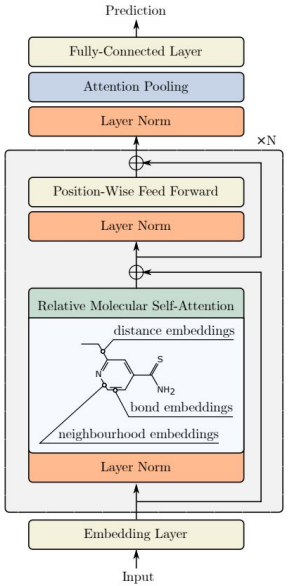
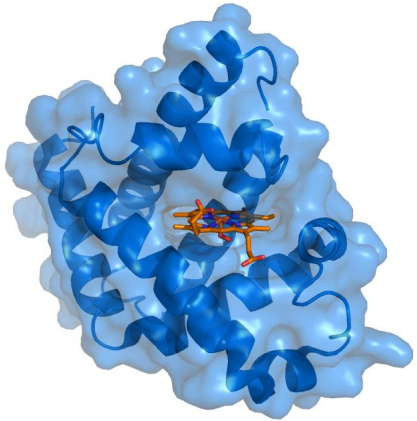
Drugsformer: Transformer for drug repurposing

Adam Kania Mateusz Pach Michał Wronka

June 12, 2023

Recap

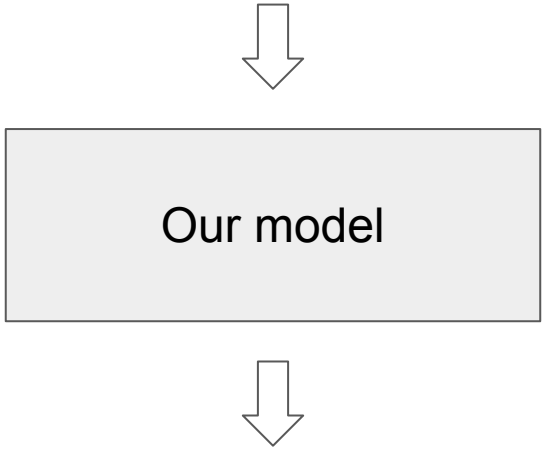




https://en.wikipedia.org/wiki/Ligand_%28biochemistry%29
Relative Molecule Attention Transformer (Maziarka et al. 2021)

Problem formulation

Ligand in SMILES + Protein with 3D representation and docking areas



Ligand activity and binding score for a given target protein

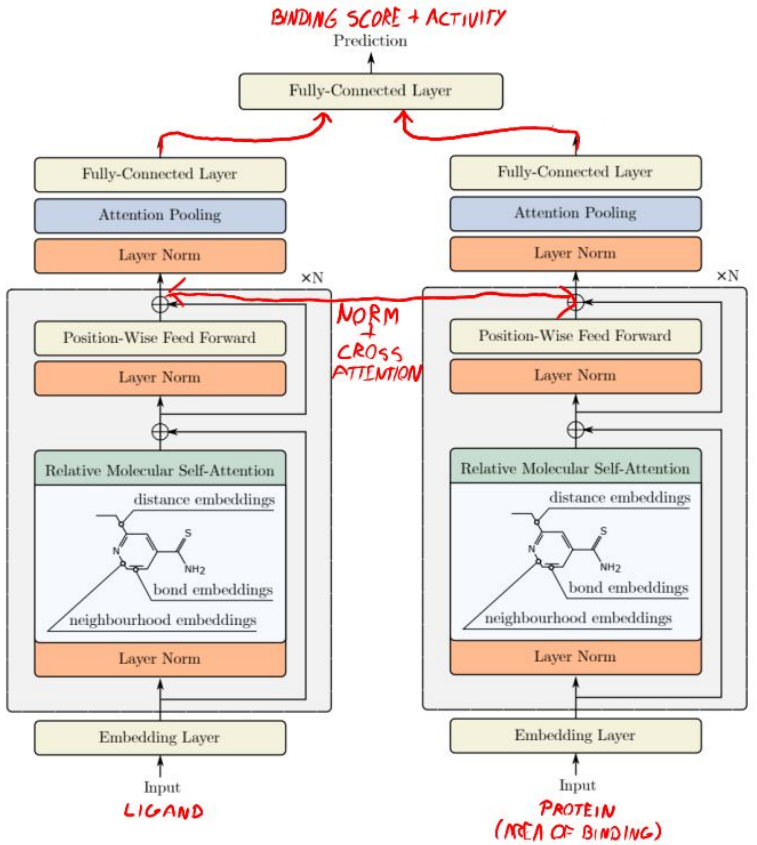
Architecture 1

We combine two RMATs and add cross-attentions between them.

Areas of bindings as inputs.

Variants:

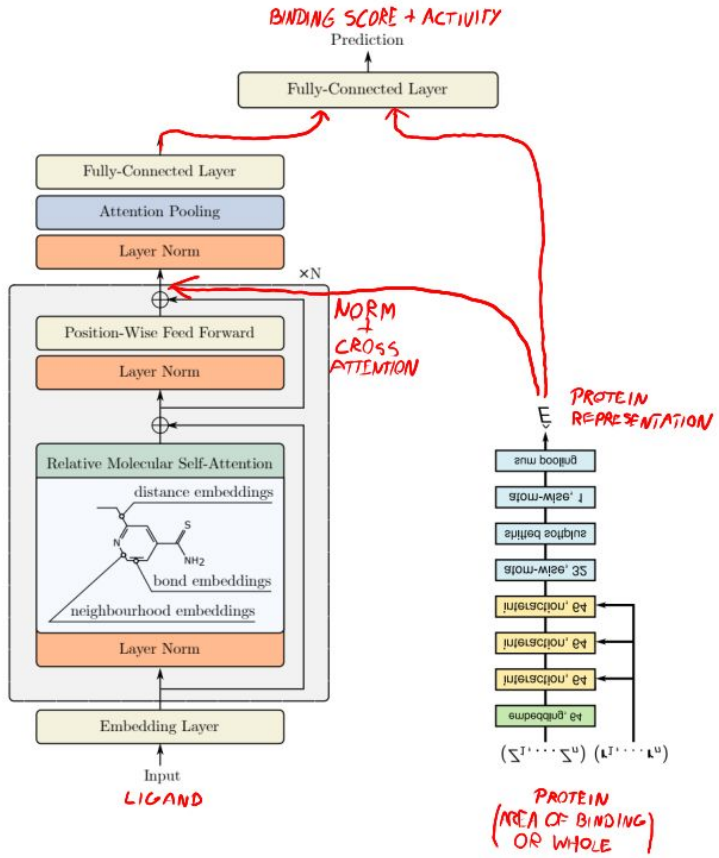
1. Only in ligand
2. Only in protein
3. In both



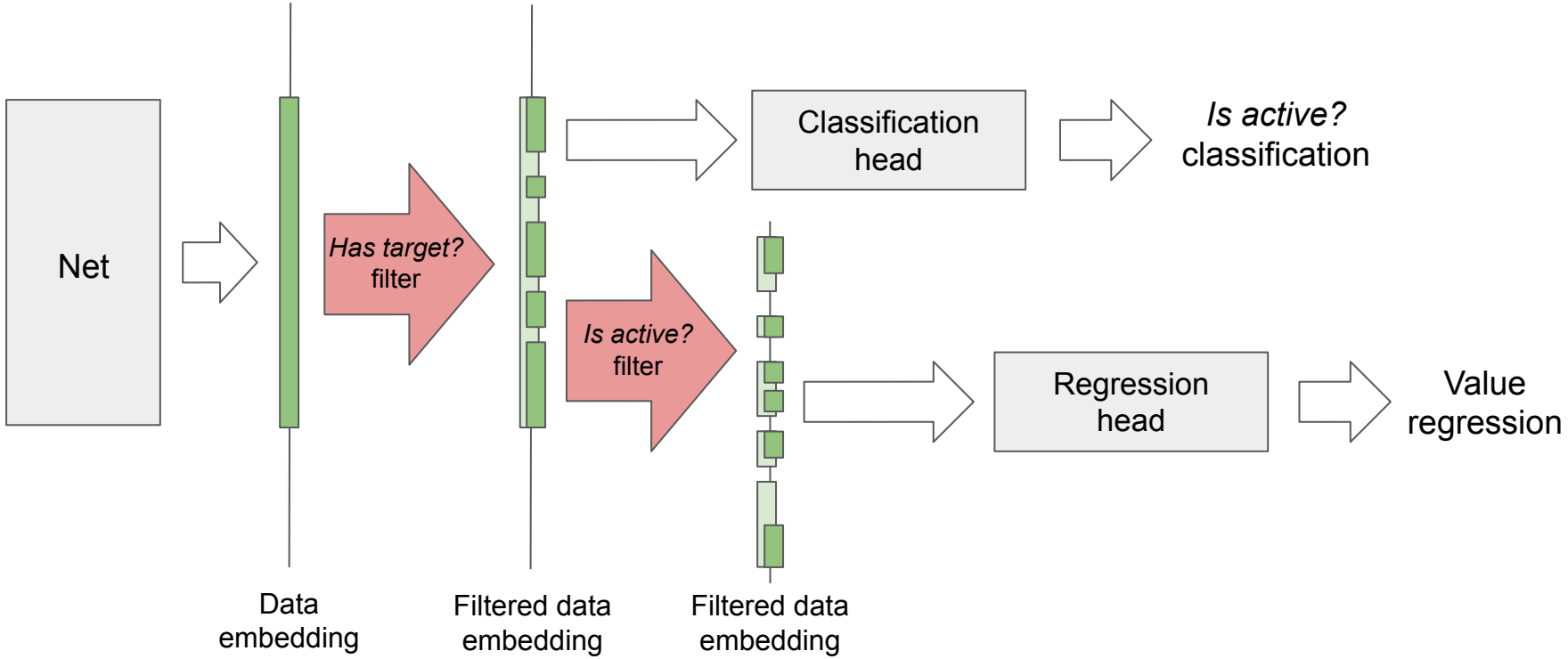
Architecture 2

We use RMATs and add cross-attentions with protein latent representation.

Areas of bindings or whole proteins as inputs.



Multiple task head



Dataset

- ~11.5k ligands for 7 targets including:
  - Cytochrome P450: CYP3A4, CYP2D6, CYP2C8,CYP2C9
  - Serotonin: 5ht7, 5ht1a
  - Dopamine: d2
- Activity (Ki/IC50) and docking scores
- Pockets coordinates

Implementation goals

Done

Milestone 1: Docking dataset

- [x] ~~Extract docking areas~~
- [x] ~~Build protein graph representation for SchNET~~
  - [x] ~~From whole protein~~
  - [x] ~~From docking area~~

Milestone 2: Architecture

- [x] ~~Build cross attending RMAT~~
  - [x] ~~Bidirectional~~
  - [x] ~~Directional into ligand~~
  - [x] ~~Directional into protein~~
- [x] ~~Produce latent from SchNET~~
- [x] ~~Build RMAT attending latent from SchNET~~

Postponed

Milestone 3: Big datasets

- [ ] Build 3D representations with AlphaFold
- [ ] Find docking areas (fpocket or Unet)
- [ ] Apply to the created models

H1: Model produces satisfying results on our dataset

Yes

Model	Binding score		Ki		IC50	
	MSE ↓	Acc ↑	MSE ↓	Acc ↑	MSE ↓	Acc ↑
RMat-RMat	**0.41**	**0.79**	0.94	0.58	**0.33**	**0.69**
RMat-SchNet	0.43	0.79	**0.86**	**0.69**	0.51	0.66

H2: Cross-attention outperforms a representations merge

No

Model	Cross-attention	Binding score		Ki		IC50	
		MSE ↓	Acc ↑	MSE ↓	Acc ↑	MSE ↓	Acc ↑
RMat-RMat		**0.41**	**0.79**	**0.94**	0.58	**0.33**	**0.69**
	✓	0.75	0.66	0.95	**0.81**	0.39	0.49

H3: Self-attention layers outperform graph layers

Inconclusive

Model	Binding score		Ki		IC50	
	MSE ↓	Acc ↑	MSE ↓	Acc ↑	MSE ↓	Acc ↑
RMat-RMat	**0.41**	**0.79**	0.94	0.58	**0.33**	**0.69**
RMat-SchNet	0.43	0.79	**0.86**	**0.69**	0.51	0.66

H4: General models are better than protein-specific ones

Inconclusive, but general are better when low on data

Model	Protein	Binding score		Ki		IC50	
		MSE ↓	Acc ↑	MSE ↓	Acc ↑	MSE ↓	Acc ↑
RMat D2	D2	**0.32**	**0.71**	0.63	**0.64**	-	-
RMat 5HT1A	5HT1A	**0.22**	**0.82**	0.83	**0.79**	-	-
RMat CYP2C8	CYP2C8	0.41	**0.84**	-	-	2.18	**0.94**
RMat-RMat	D2	0.41	0.68	**0.61**	0.59	-	-
	5HT1A	0.41	0.79	**0.70**	0.56	-	-
	CYP2C8	0.41	0.82	-	-	**0.35**	0.69

H5: Restricting the input to a pocket neighbourhood helps

Yes

Model	Only pocket	Binding score		Ki		IC50	
		MSE ↓	Acc ↑	MSE ↓	Acc ↑	MSE ↓	Acc ↑
RMat-RMat	✓	**0.41**	**0.79**	**0.94**	**0.58**	**0.33**	**0.69**
		-	-	-	-	-	-
RMat-SchNet	✓	**0.43**	**0.79**	0.86	**0.69**	0.51	**0.66**
		0.45	0.77	0.86	0.68	0.51	0.63

H6: Multiple tasks do not hurt the training

Yes

Binding score		Ki		IC50	
MSE ↓	Acc ↑	MSE ↓	Acc ↑	MSE ↓	Acc ↑
0.41	**0.79**	**0.94**	0.58	**0.33**	**0.69**
0.51	-	-	-	-	-
-	0.74	-	-	-	-
0.44	0.76	-	**0.65**	-	0.64

Questions