Bioinformatics, 34(15), 2018, 2642–2648
doi: 10.1093/bioinformatics/bty178
Advance Access Publication Date: 23 March 2018
Original Paper

Data and text mining

Learned protein embeddings for machine learning

Kevin K. Yang[1], Zachary Wu[1], Claire N. Bedbrook[2] and Frances H. Arnold[1,2,*]

[1]Division of Chemistry and Chemical Engineering and [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA

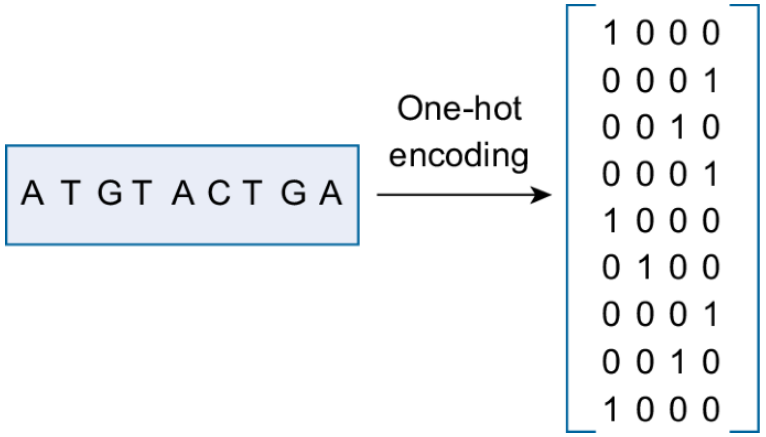
*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 30, 2017; revised on March 20, 2018; editorial decision on March 21, 2018; accepted on March 22, 2018

Exemplary protein encoding methods

- Physicochemical properties of amino acids
- One-hot encoding
- String mismatch kernels



Modeling scheme

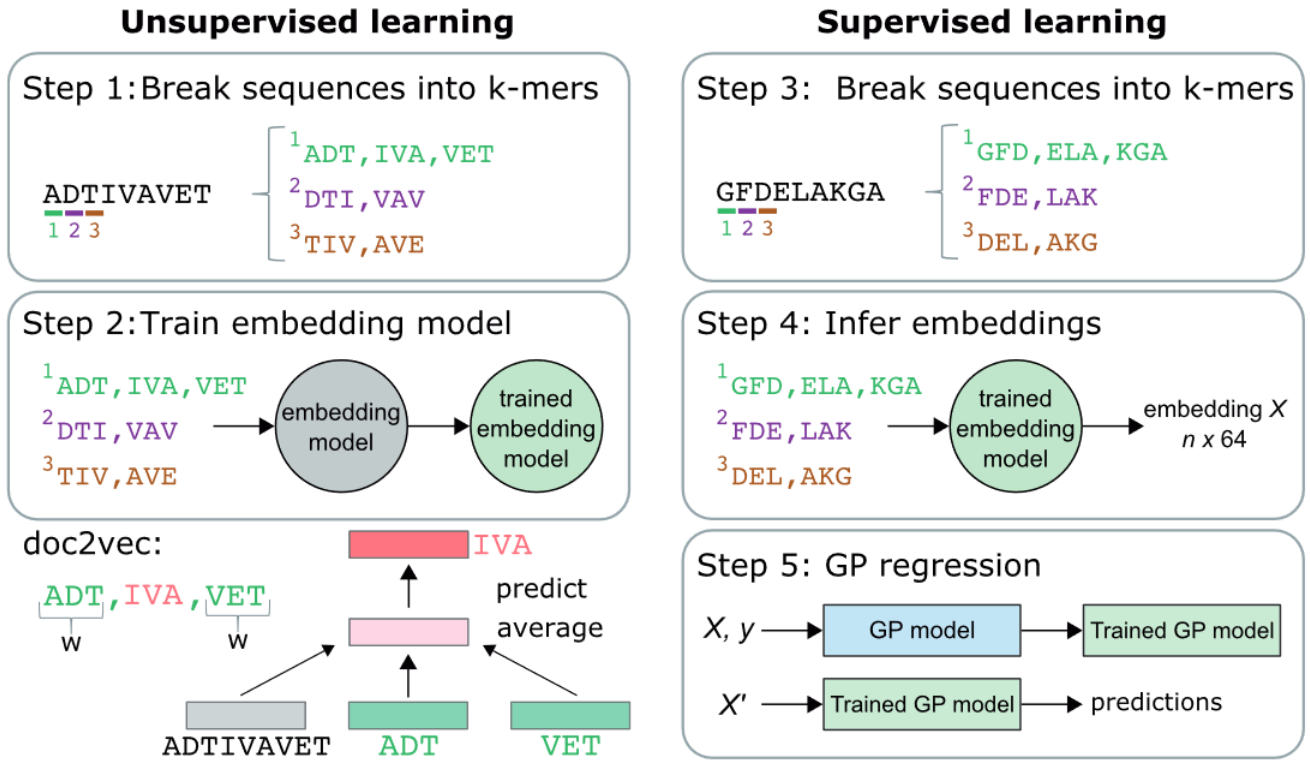
k = [1, 5]

w = [1, 7]

20-fold CV

25 epochs

d = 64



GP models were trained on:

- Embedded representations
- One-hot representation
- Mismatch string kernels (k = 5, m = 1)
- ProFET
- Subset of AAIndex (64 properties)
- One-hot representation with structural information

Table 1. Summary of tasks used to evaluate embedded representations

Task	*n*	Protein	Library	Property	Citation
Localization	248	Channelrhodopsin	Recombination	Plasma membrane localization	Bedbrook *et al.* (2017a)
T50	261	Cytochrome P450	Recombination	Thermostability	Li *et al.* (2007) and Romero *et al.* (2013)
Absorption	81	Bacterial rhodopsin	Site-saturation	Peak absorption wavelength	Engqvist *et al.* (2015)
Enantioselectivity	152	Epoxide hydrolase	Site-saturation	Enantioselectivity	Zaugg *et al.* (2017)

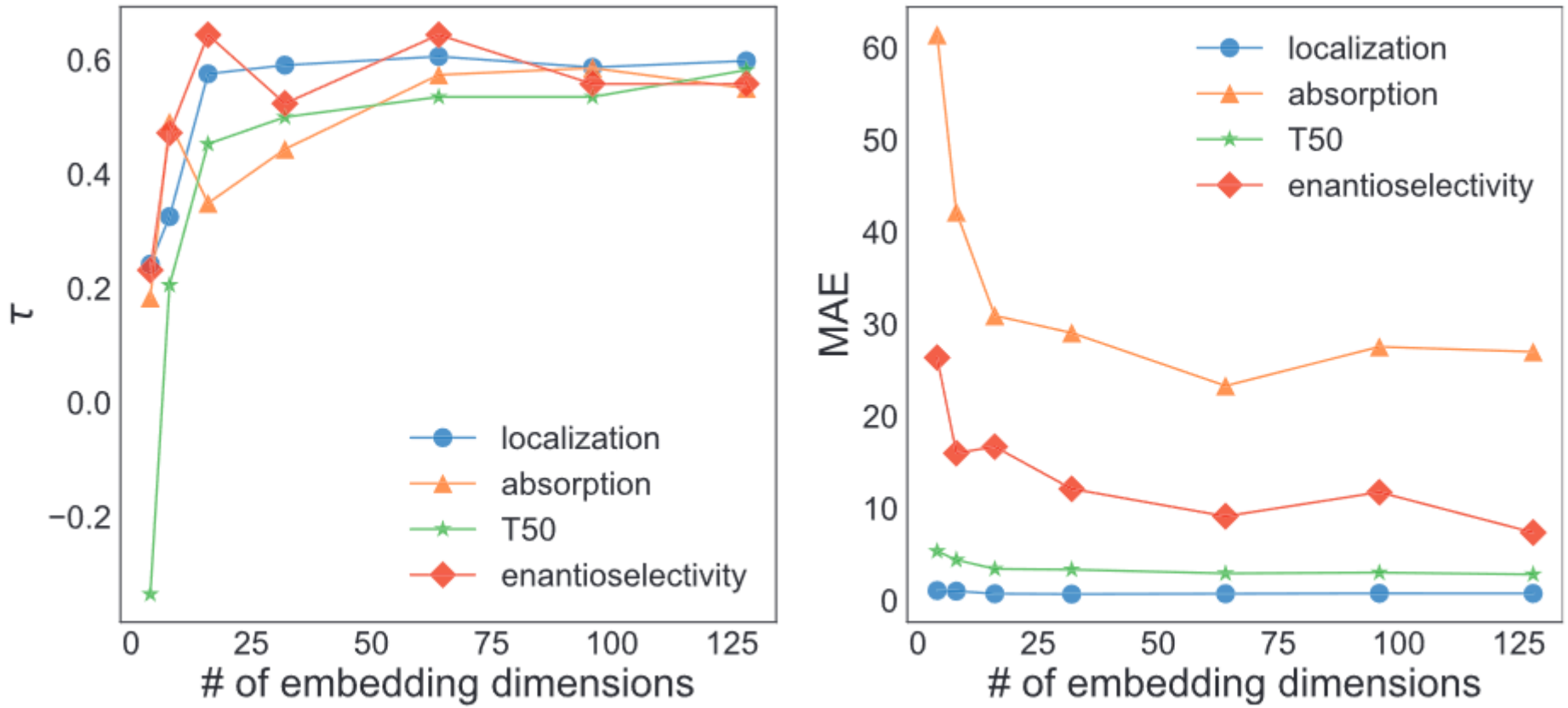
Table 2. Comparison of learned, dense, embedded representations, ProFET, AAIndex properties, mismatch string kernels and one-hot representations of sequence and structure for predicting protein properties using GP regression

Task	n_{train}	n_{test}	Representation	d	MAE	τ	$\log P$
Localization	215	33	Embedding	**64**	**0.73**	**0.60**	−43.5
			One-hot seq. and struct.	600 747	0.76	**0.60**	**−43.2**
			One-hot sequence	7161	0.76	0.59	−43.7
			Mismatch kernel	–	0.86	0.55	−54.6
			ProFET	1173	1.03	0.32	−54.9
			AAIndex properties	21 824	0.76	0.55	−44.3
T50	242	19	Embedding	**64**	**2.91**	**0.61**	−59.5
			One-hot seq. and struct.	994 980	2.98	0.53	−57.3
			One-hot sequence	9786	2.94	0.57	−57.2
			Mismatch kernel	–	4.03	0.38	−58.5
			ProFET	1173	4.93	0.43	−63.7
			AAIndex properties	29 824	2.95	0.51	**−56.2**
Absorption	62	19	Embedding	**64**	23.3	0.57	−109.2
			One-hot sequence	6258	22.1	0.63	−111.0
			Mismatch kernel	–	**17.8**	**0.68**	**−103.9**
			ProFET	1173	53.5	0.32	−174.7
			AAIndex properties	19 072	30.1	0.35	−116.4
Enantioselectivity	136	16	Embedding	**64**	9.14	**0.64**	−64.5
			One-hot sequence	8358	8.16	0.50	**−63.3**
			Mismatch kernel	–	**7.50**	0.46	−65.1
			ProFET	1173	27.9	0.27	−76.7
			AAIndex properties	25 472	12.5	0.25	−65.7

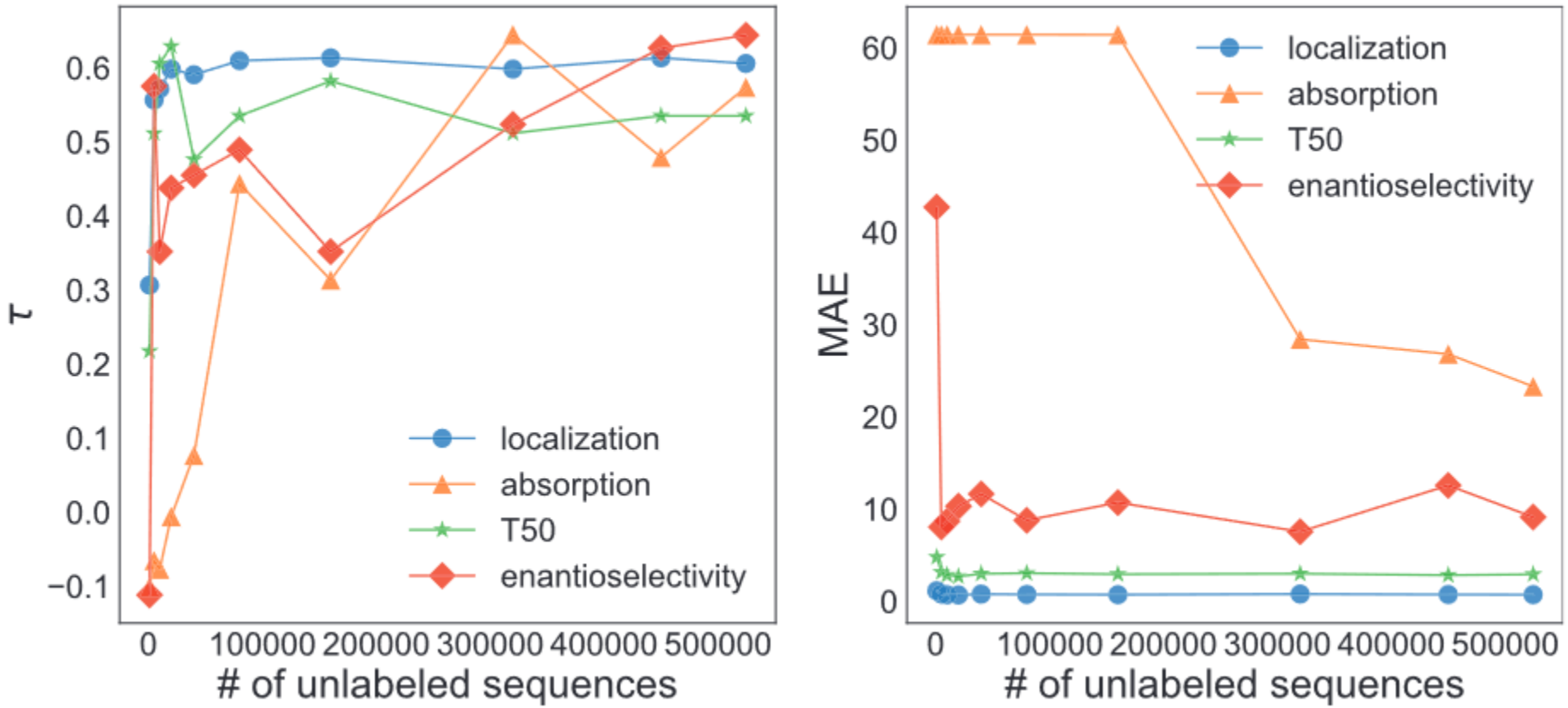
Table 3. Negative controls

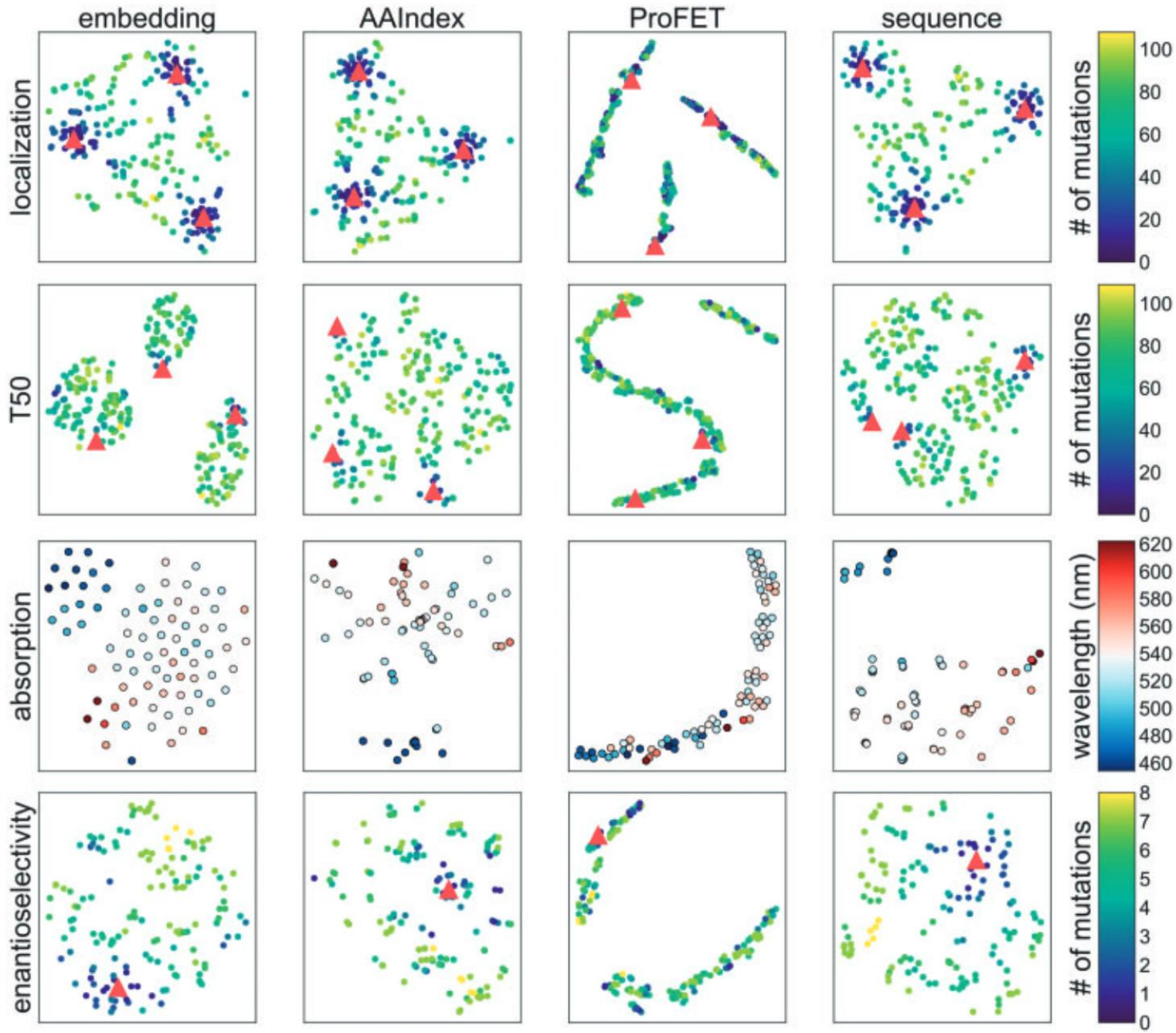
Task	Control	*MAE*	τ	log *P*
Localization	–	0.73	0.60	−43.5
	Task sequences only	0.86	0.50	−50.0
	Shuffled task sequences	1.21	0.16	−57.4
	Shuffled training labels	1.16	−0.39	−58.3
T50	–	2.91	0.61	−59.5
	Task sequences only	5.02	0.45	−63.3
	Shuffled task sequences	4.49	0.31	−61.8
	Shuffled training labels	5.72	−0.35	−67.1
Absorption	–	23.3	0.57	−109.2
	Task sequences only	61.4	0.34	−162.1
	Shuffled task sequences	61.4	−0.03	−162.0
	Shuffled training labels	61.4	−0.43	−162.0
Enantioselectivity	–	9.14	0.64	−64.5
	Task sequences only	41.3	−0.06	−85.2
	Shuffled task sequences	42.7	0.27	−84.7
	Shuffled training labels	42.8	0.06	−84.8

Effect of embedding dimension on accuracy

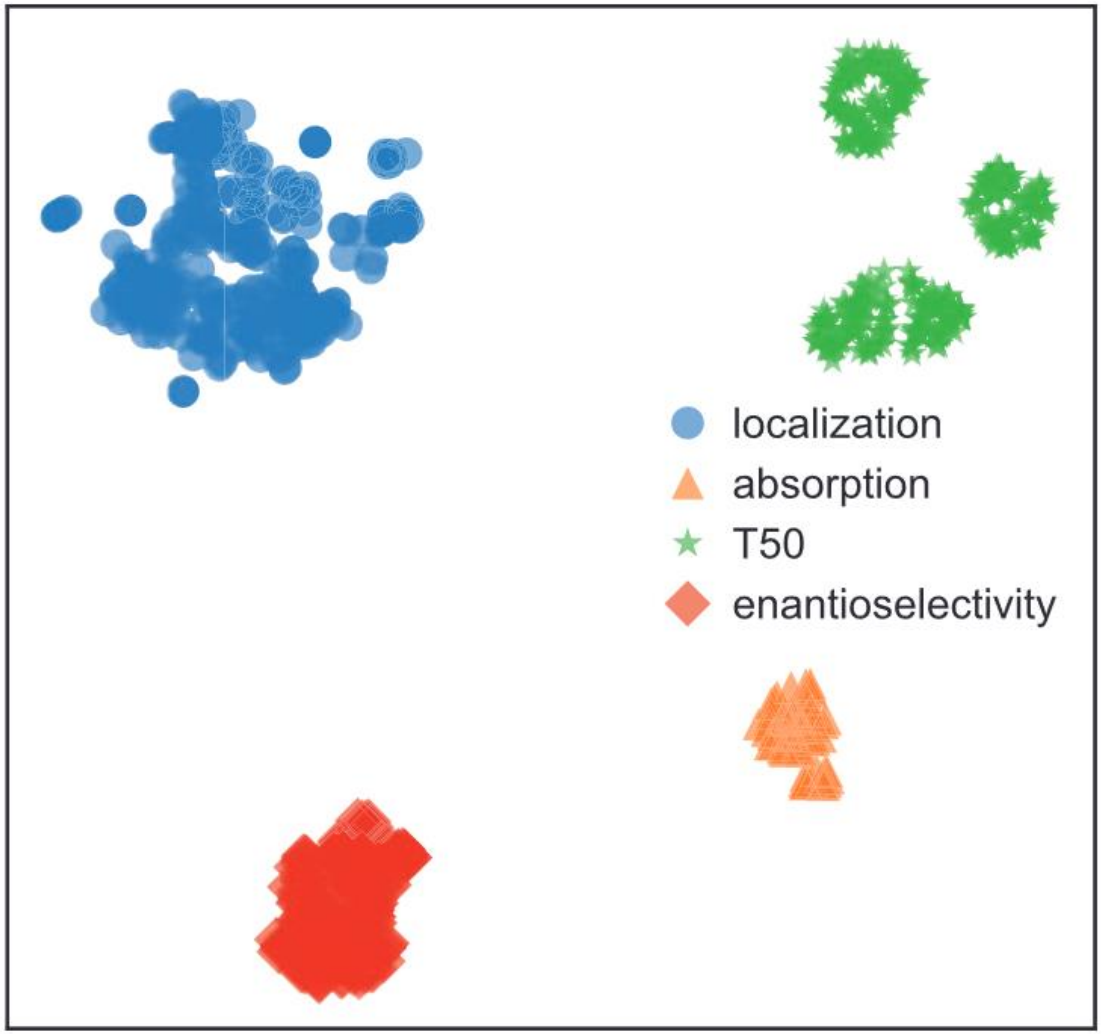


Effect of number of unlabeled sequences on accuracy




embedding
AAIndex
ProFET
sequence
localization
T50
absorption
enantioselectivity
# of mutations
wavelength (nm)
100
80
60
40
20
0
620
600
580
560
540
520
500
480
460
8
7
6
5
4
3
2
1
0

Projection of embeddings for all tasks



Conclusions

- Embedding models can be applied to predict the functional properties of a small number of related proteins
- Embeddings generalize across protein families, library designs and protein properties
- 32 dimensions are sufficient to achieve competitive model performance
- The unsupervised embedding model incorporates information from the unlabeled sequences
- We are able to transfer information encoded in these unlabeled sequences to a specific task