ChemBoost: A Chemical Language-based Approach for Drug-Target Affinity Prediction

Rıza Özçelik+, Hakime Öztürk+, Arzucan Özgür, & Elif Özkırımlı

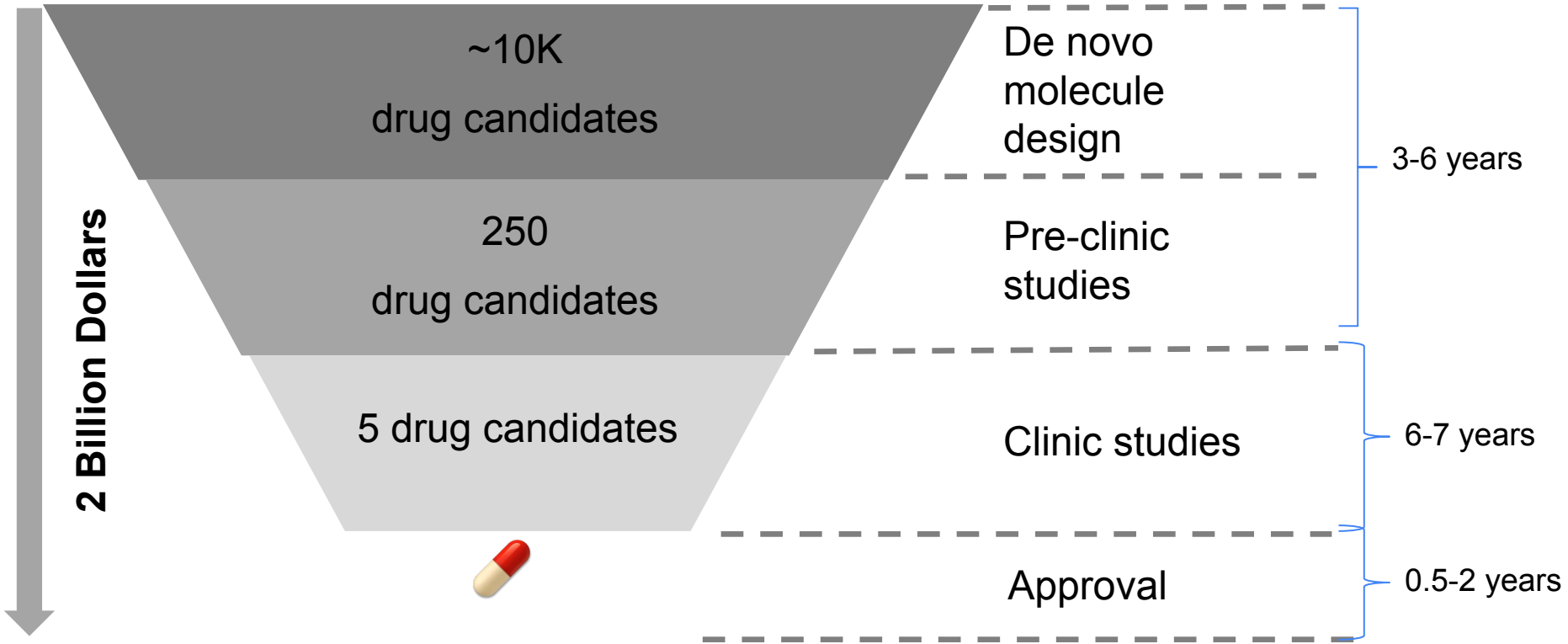
(2021). *Molecular Informatics,* 40(5), 2000212.

+Equal contribution

ISMB/ECCB 2021

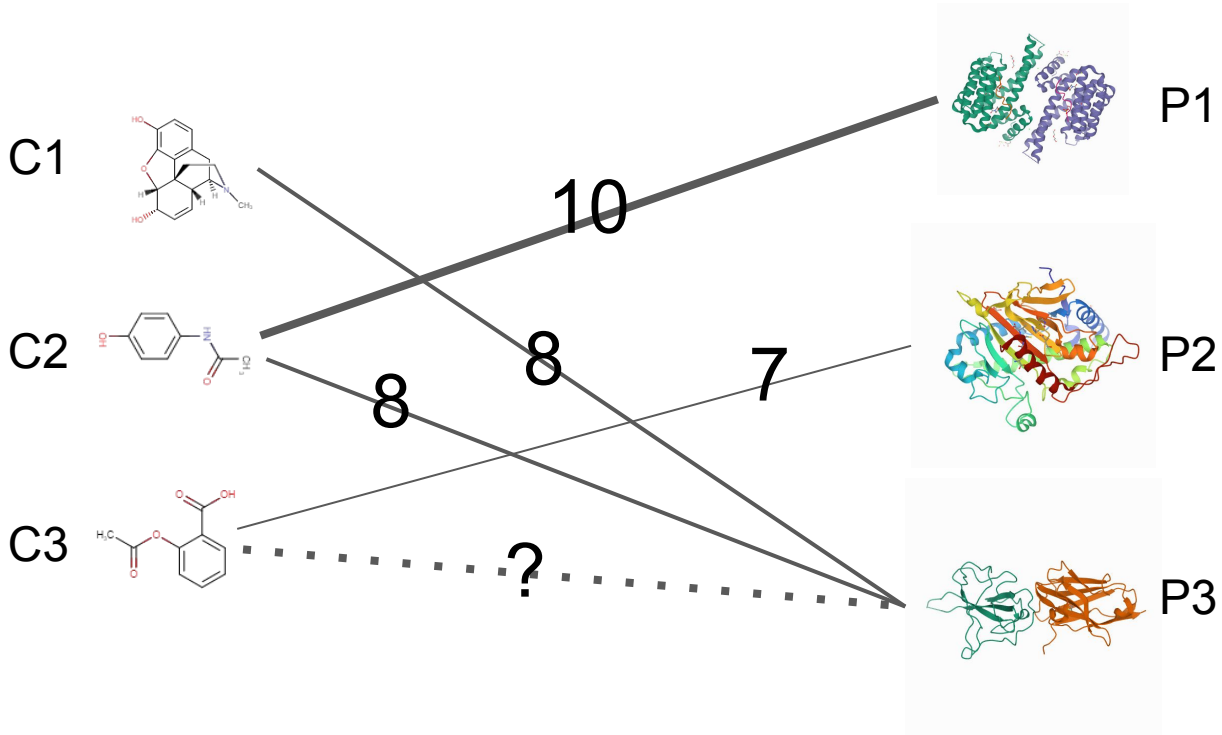
July 29

Drug Discovery is a Long and Expensive Process

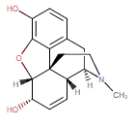


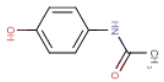
http://www.ogjc.osaka-gu.ac.jp/php/nakagawa/TRIZ/eTRIZ/

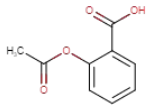
Searching High-Affinity Pairs



Story of a Needle in the Haystack

C1 

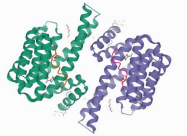
C2 

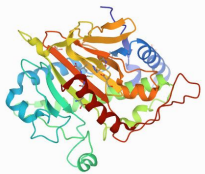
CN 

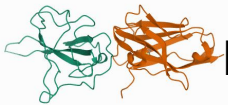
~2.1M compounds in ChEMBL

Too many combinations!

Computational Methods!

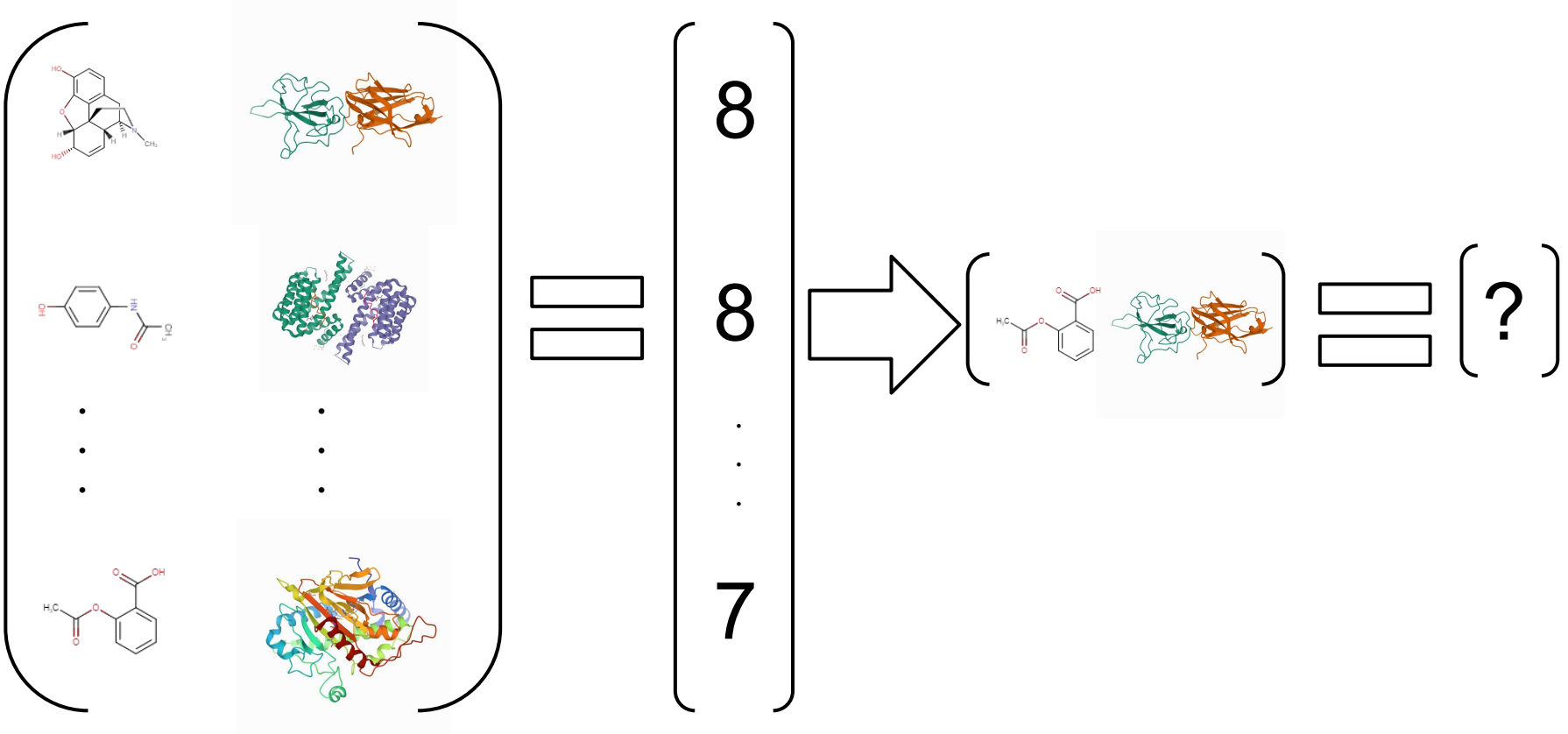
 P1

 P2

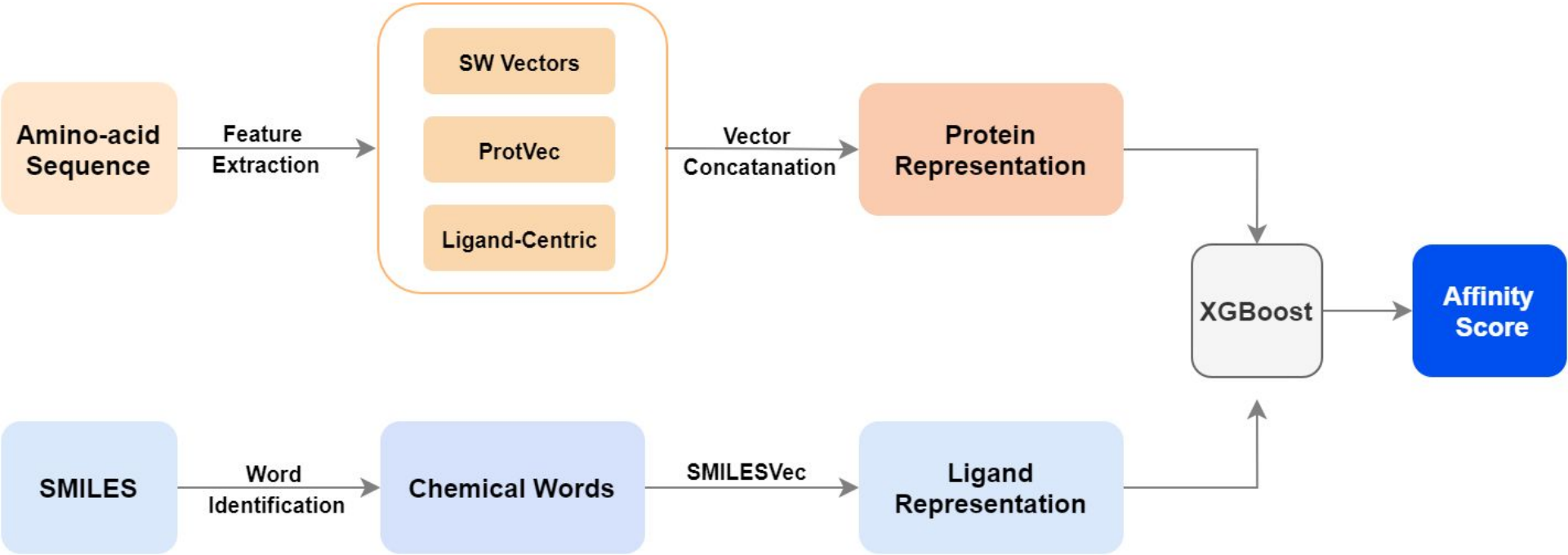
 PN

~560K proteins in SwissProt

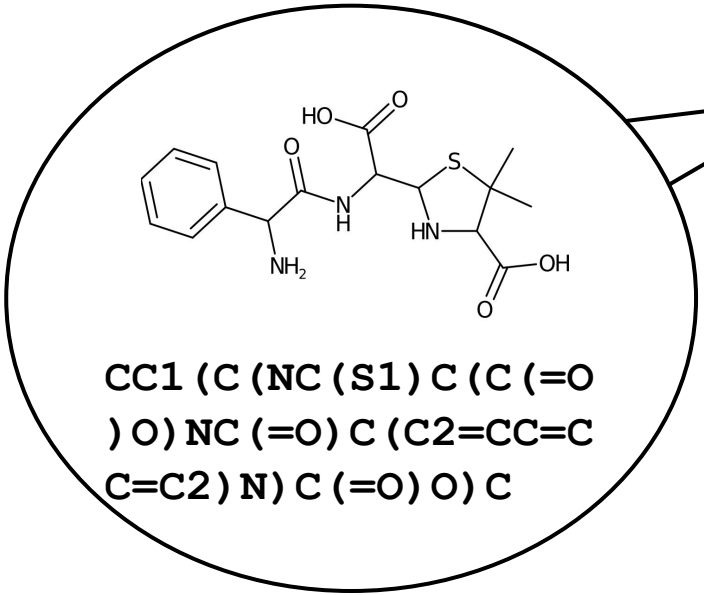
Machine Learning for Affinity Prediction



ChemBoost: A Chemical Language-based Approach for Drug-Target Affinity Prediction

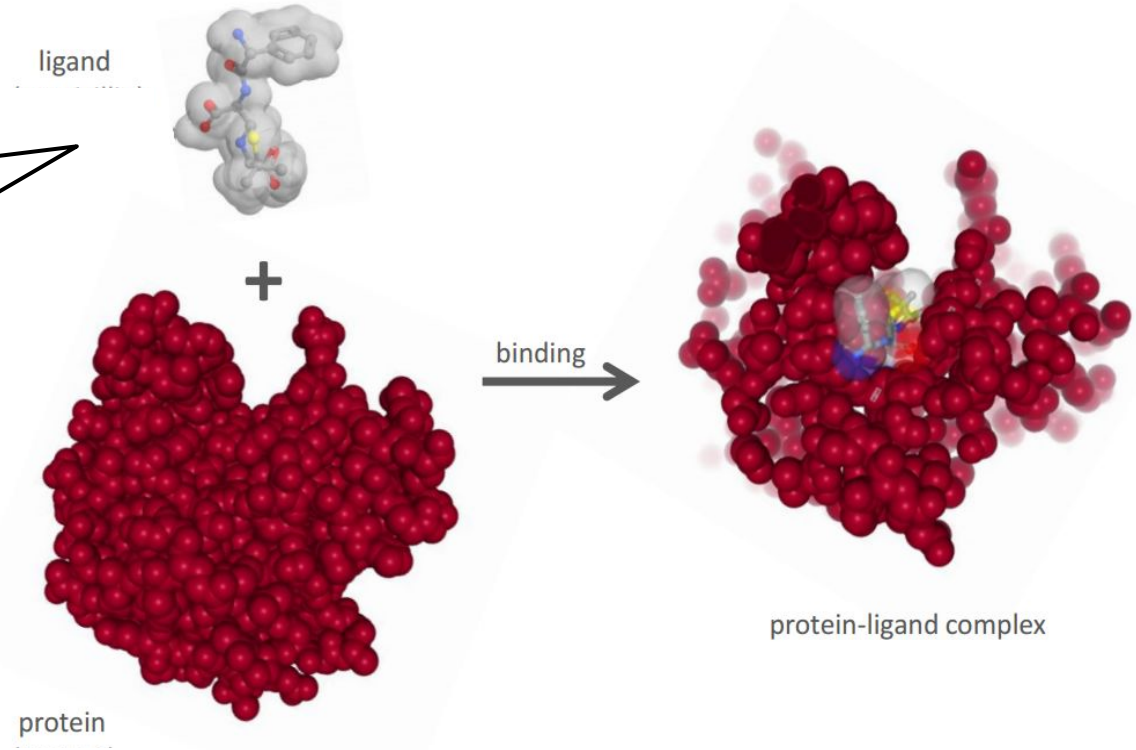


Chemical Language



CC1(C(NC(S1)C(C(=O)O)NC(=O)C(C2=CC=CC=C2)N)C(=O)O)C

Simplified Molecular Input Line Entry System (SMILES)



If **SMILES** is a document...

SMILES: `CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)C`

Where are the words?

If **SMILES** is a document, 8-mers are the words

SMILES: `CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)C`

Chemical Words:

`CC1(C(N2`

If **SMILES** is a document, 8-mers are the words

SMILES: `CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)C`

Chemical Words:

`CC1(C(N2`

`C1(C(N2C`

If **SMILES** is a document, 8-mers are the words

SMILES: `CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)C`

Chemical Words:

`CC1(C(N2`

`C1(C(N2C`

`1(C(N2C(`

If **SMILES** is a document, 8-mers are the words

SMILES: `CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)C`

Chemical Words:

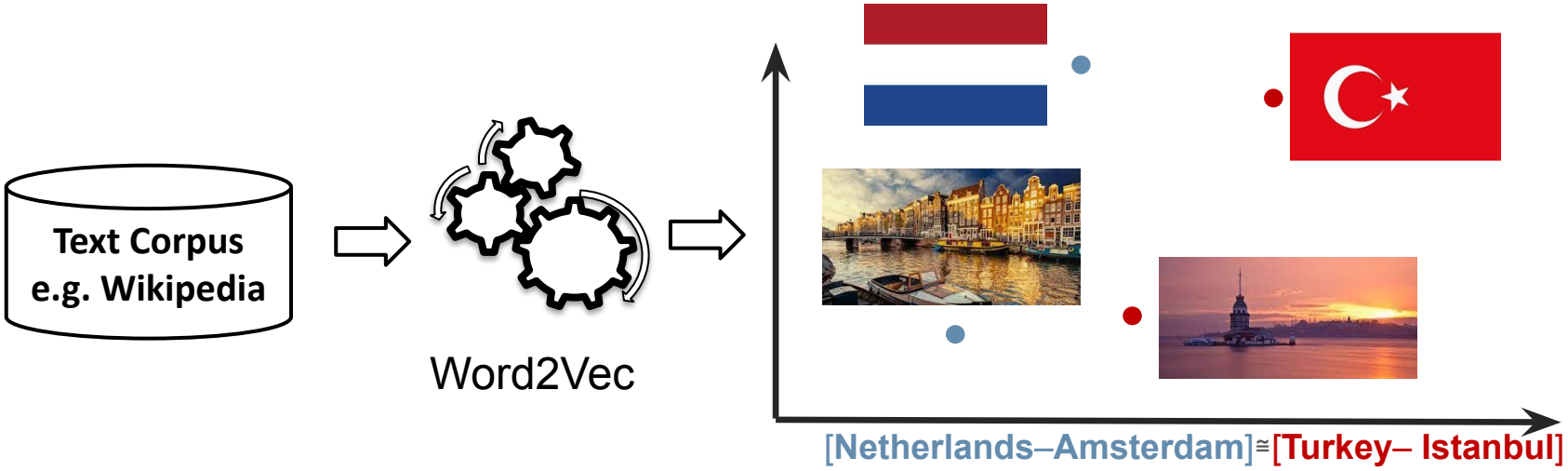
`CC1(C(N2`

`C1(C(N2C`

`1(C(N2C(`

`C(=O)O)C`

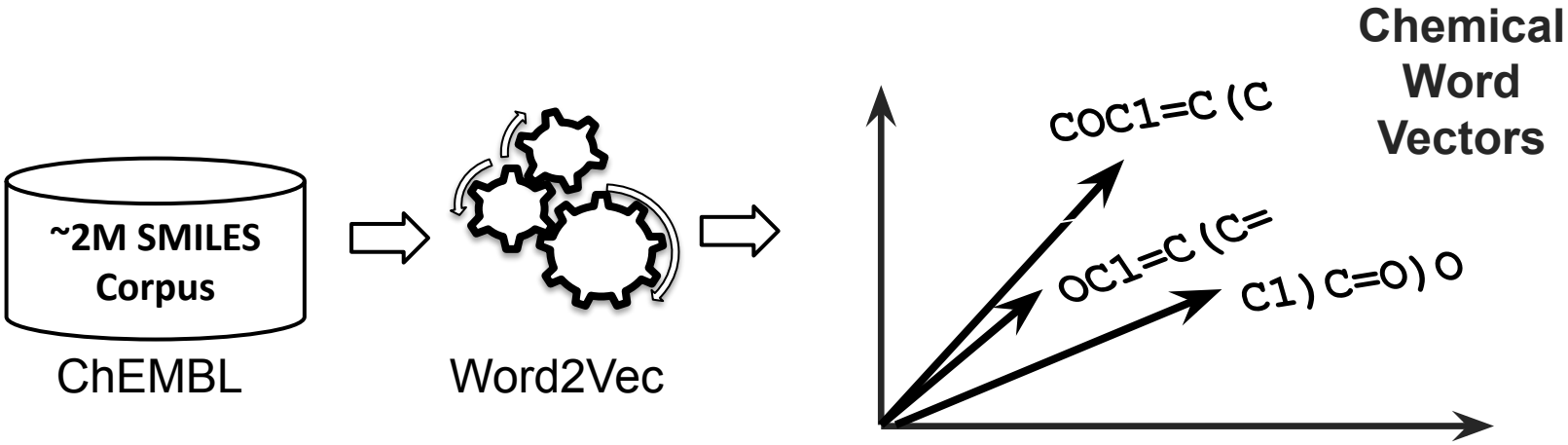
Word Vectors with Word2Vec



Words appear in the similar contexts are semantically similar.

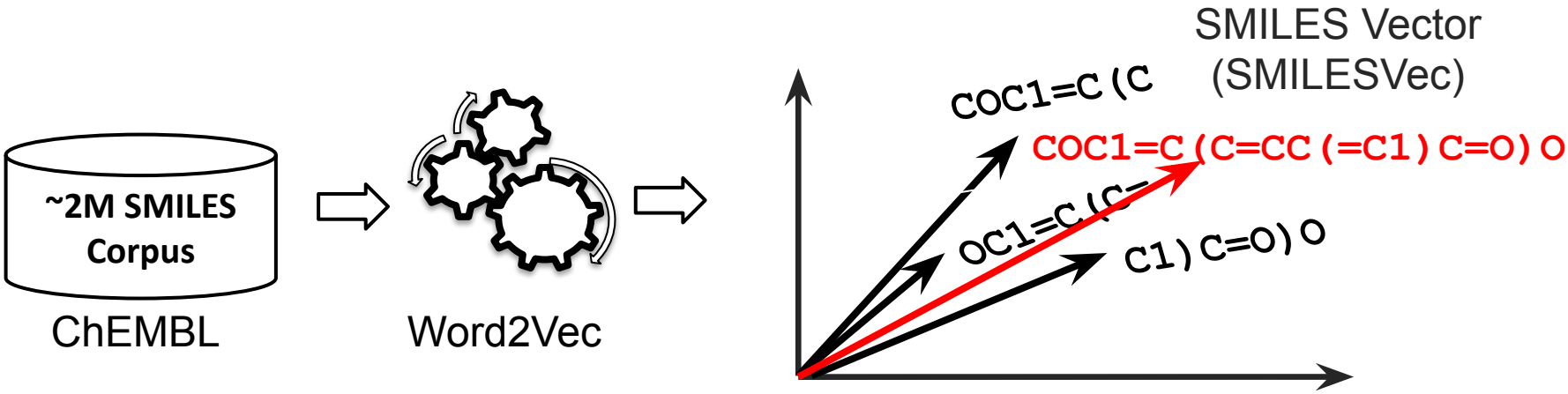
Mikolov, Tomas, et al. 2013.

SMILESVec: Ligand Vectors



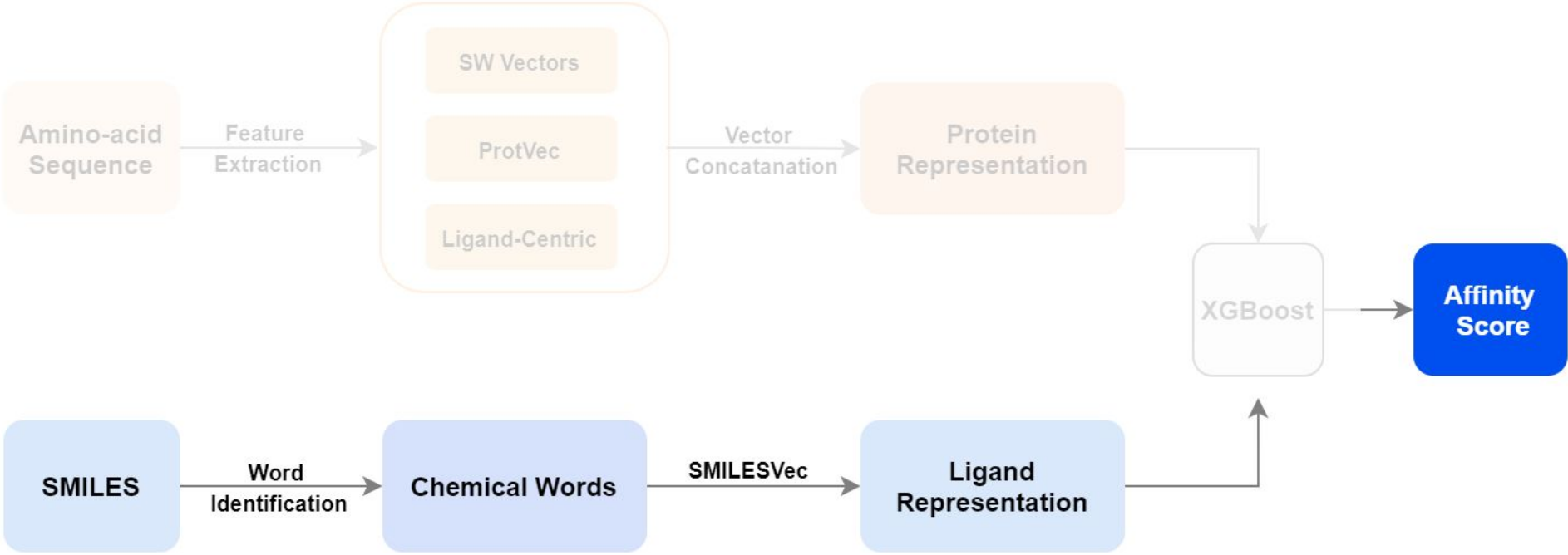
H. Ozturk, E. Ozkirimli, and A. Ozgur. ***A novel methodology on distributed representations of proteins using their interacting ligands.*** *Bioinformatics,* Volume 34, Issue 13, Pages i295-i303, 2018.

SMILESVec: Ligand Vectors

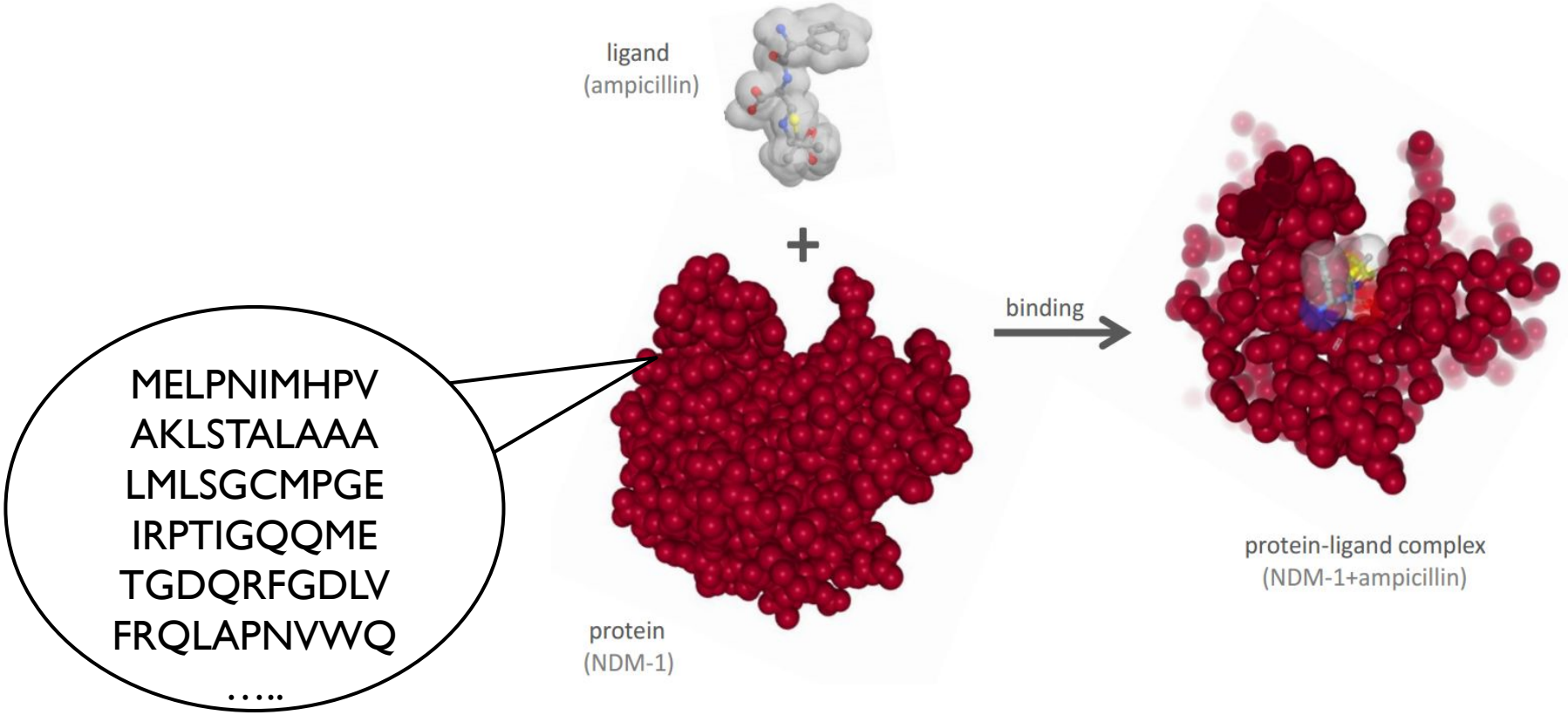


H. Ozturk, E. Ozkirimli, and A. Ozgur. ***A novel methodology on distributed representations of proteins using their interacting ligands.*** *Bioinformatics,* Volume 34, Issue 13, Pages i295-i303, 2018.

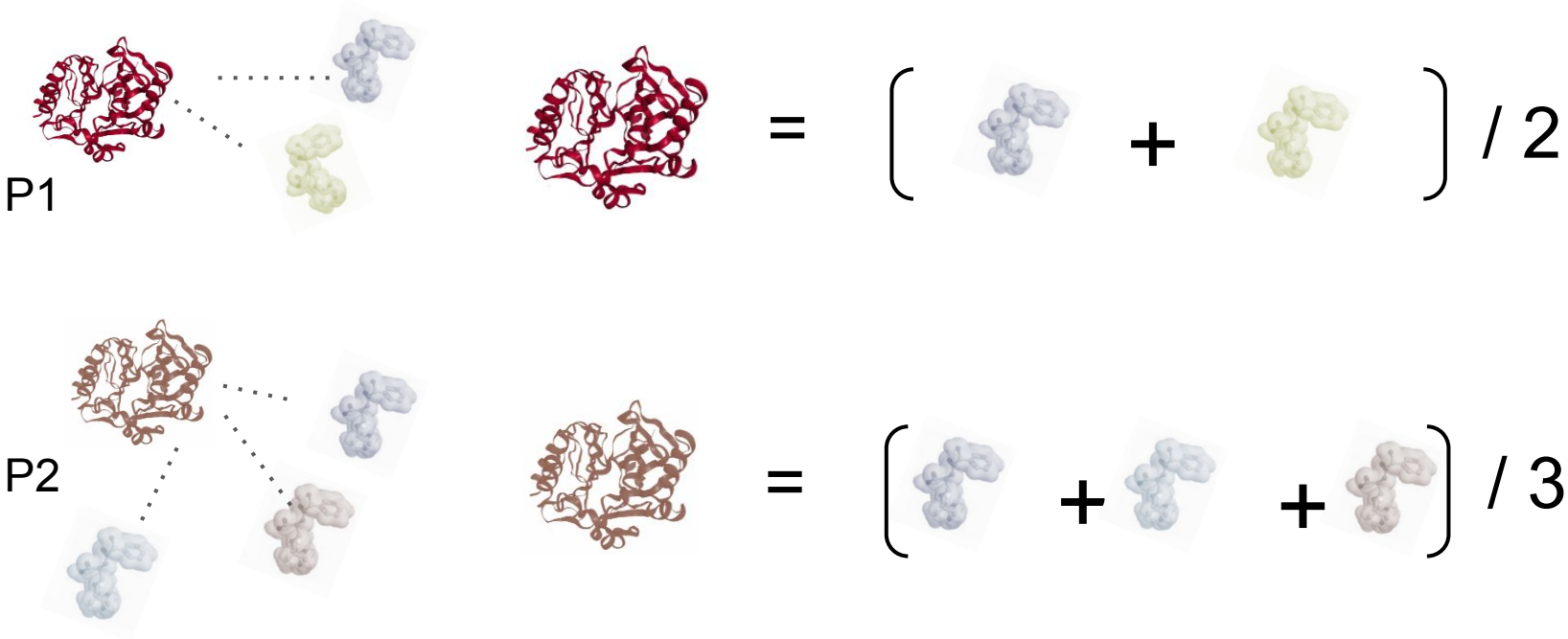
ChemBoost Ligand Representation



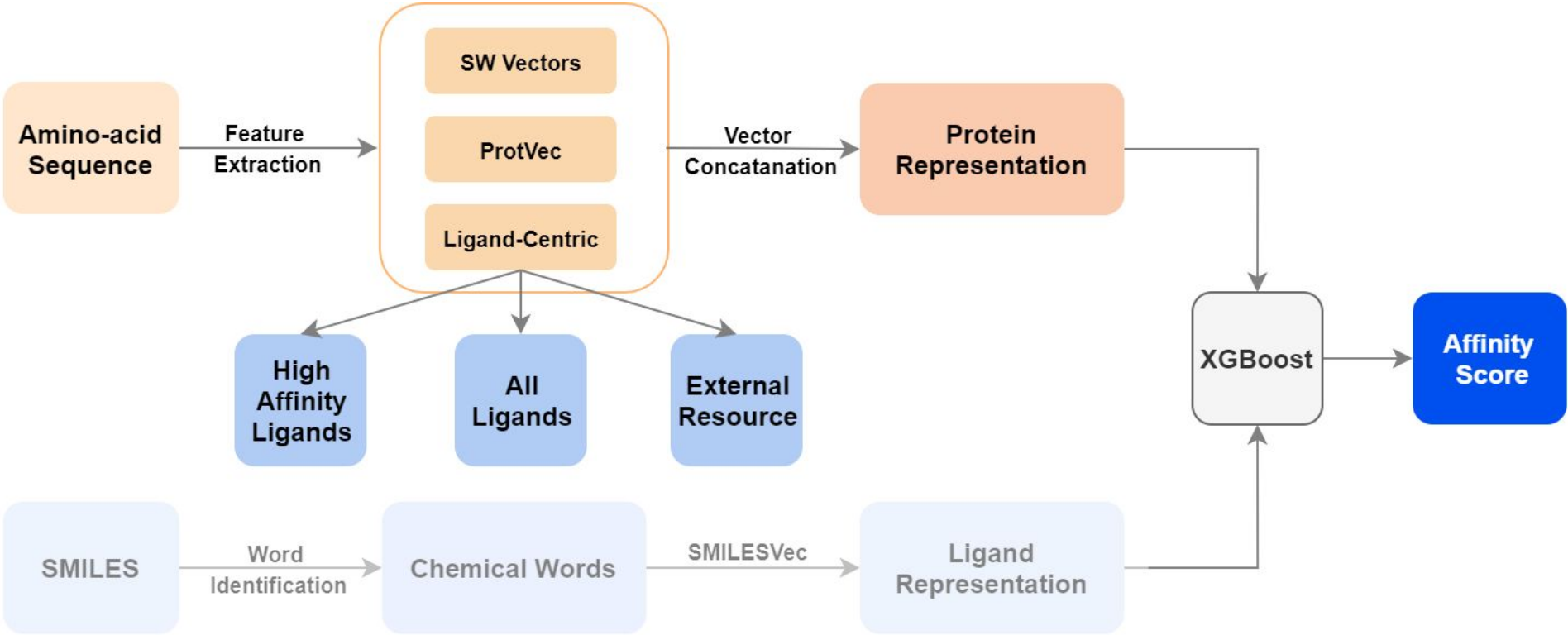
Proteins as Amino-acid Sequences



Idea: Ligand-centric protein representation



ChemBoost Protein Representation



Experiments

- **BDB:** 490 proteins, 924 ligands, ~31K interactions
- **KIBA:** Kinase dataset, 229 proteins, 2111 ligands, ~118K interactions
- 5-fold cross-validation
- Evaluation: Mean squared error (MSE) and concordance index (CI)

Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. Journal of Chemical Information and Modeling, 54(3):735–743, 2014.

∴ High affinity ligands of a protein are more informative than all known ligands.

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

∴ Incorporating an external database improves performance on BDB.

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

∴ Hybrid models are more reliable than single-representation.

Model			BDB Scores		KIBA Scores	
Name	Protein Representation	Ligand Representation	CI	MSE	CI	MSE
Model (1)	SW	SMILESVec (8-mer)	0.873	0.439	0.837	0.203
Model (2)	ProtVec	SMILESVec (8-mer)	0.854	0.512	0.818	0.244
Model (3)	ProtVec	SMILESVec (BPE)	0.849	0.548	0.814	0.252
Model (4)	SMILESVec (all, 8-mer)	SMILESVec (8-mer)	0.847	0.524	0.823	0.243
Model (5)	SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.845	0.478	0.829	0.221
Model (6)	SMILESVec (SB, BPE)	SMILESVec (BPE)	0.842	0.497	0.825	0.227
Model (7)	SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.856	0.454	0.829	0.223
Model (8)	SW & SMILESVec (SB, 8-mer)	SMILESVec (8-mer)	0.873	0.420	0.837	0.206
Model (9)	SW & SMILESVec (BindingDB SB, 8-mer)	SMILESVec (8-mer)	0.871	0.420	0.836	0.207

ChemBoost in the Wild

	BDB Scores		KIBA Scores	
Model	**CI**	**MSE**	**CI**	**MSE**
KronRLS	0.814 (0.002)	0.939 (0.004)	0.782 (0.001)	0.411
SimBoost	0.853 (0.003)	0.485 (0.043)	0.836 (0.001)	0.223 (0.003)
DeepDTA	0.863 (0.007)	0.397 (0.011)	0.846 (0.002)	0.215 (0.005)
ChemBoost	0.871 (0.002)	0.420 (0.007)	0.836 (0.001)	0.207 (0.002)

- Pahikkala et al. "Toward more realistic drug–target interaction predictions." *Briefings in bioinformatics* 16.2 (2014): 325-337.
- Tong, et al. "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines." *Journal of cheminformatics* 9.1 (2017): 24.
- Öztürk et al. "DeepDTA: deep drug–target binding affinity prediction." Bioinformatics 34.17 (2018): i821-i829.

∴ Novel biomolecule representation is challenging for all models.

Model	Warm		Cold Ligand		Cold Protein		Cold	
	MSE	CI	MSE	CI	MSE	CI	MSE	CI
Model (1)	0.373	0.885	1.178	0.736	0.720	0.799	1.393	0.657
Model (7)	0.404	0.863	1.185	0.700	1.156	0.749	1.576	0.596
Model (9)	0.361	0.880	1.157	0.730	0.800	0.786	1.358	0.665
DeepDTA	0.345	0.879	1.350	0.672	0.810	0.778	1.522	0.614
Model (1)	0.185	0.845	0.450	0.732	0.298	0.762	0.588	0.646
Model (7)	0.202	0.839	0.445	0.736	0.453	0.734	0.667	0.638
Model (9)	0.183	0.847	0.442	0.735	0.340	0.748	0.614	0.640
DeepDTA	0.199	0.853	0.456	0.754	0.400	0.747	0.655	0.652

Thanks for listening!