technische universität dortmund

fakultät für chemie und chemische biologie

Man, machine, molecules: A journey from face recognition to drug discovery

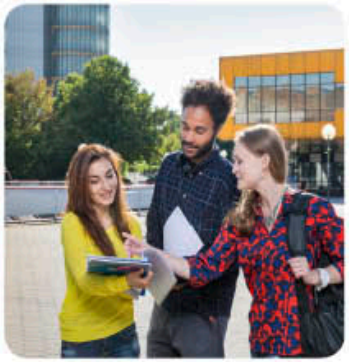
Prof. Dr. Paul Czodrowski

www.czodrowskilab.org

twitter.com/czodrowskipaul

www.czodrowskilab.org/music

github.com/czodrowskilab



Drug discovery

JOURNAL OF CHEMICAL INFORMATION AND MODELING

OCEAN: Optimized Cross rEActivity estimatioN

Paul Czodrowski* and Wolf-Guido Bolick

JOURNAL OF CHEMICAL INFORMATION AND MODELING

hERG Me Out

Paul Czodrowski

Journal of Medicinal Chemistry

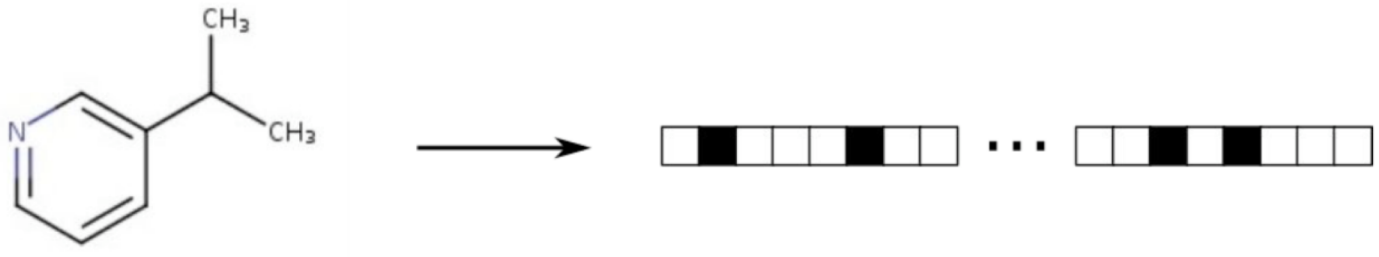
Selection of Fragments for Kinase Inhibitor Design: Decoration Is Key

Paul Czodrowski,* Günter Hölzemann, Gerhard Barnickel, Hartmut Greiner, and Djordje Musil*

dhd drug discovery hub dortmund

Computer-readability of molecules

Idea: Apply a transform to a molecule to generate a bit vector



Molecule → Fingerprint

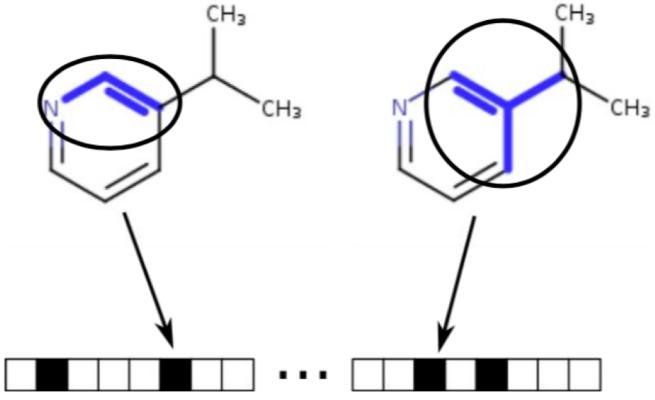
A typical **fingerprinting algorithm** can be generalized as

- extract features of the molecule
- Hash the features (= convert them to a number in a unique way)
- Use the hash to determine which bits should be set

Default fingerprint size is **4.096 bits**

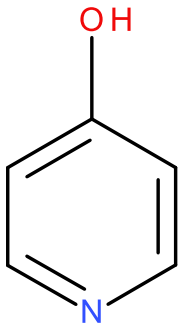
Fingerprint similarity == Molecule similarity?

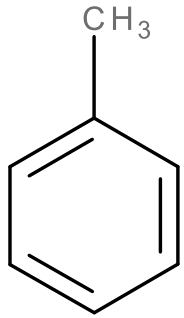
- Each fingerprint bit corresponds to a fragment of the molecule



- Assumption: molecules that are similar have a lot of fragments in common
- No "right" answer for defining similarity

Computer scientist's guide to chemical similarity





1	0	1	1	Compound A

1	1	0	1	Compound B

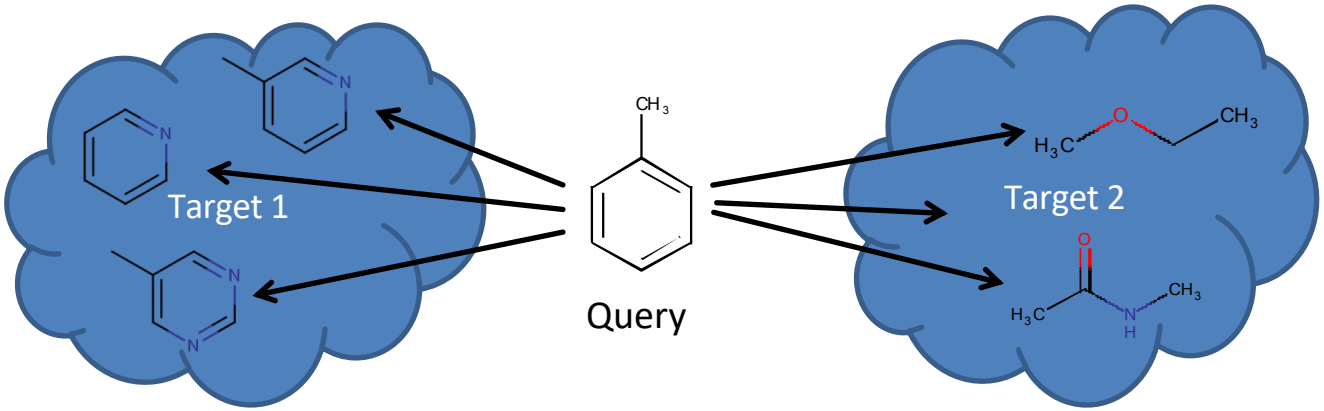
1	0	0	1	Bits set in A & B

1	1	1	1	Bits set in A or B

Tanimoto Similarity = (# bits set in A & B)/(# bits set in A or B)

= 2/4 = 0.5

What is OCEAN?



TC:	0.7			
	0.6			T1:
	0.4	z-Score	p-Value	e-Value
Σ	1.7 →	2.5 →	0.01 →	0.02

			TC:	0.6
T2:				0.4
e-Value	p-Value	z-Score		0.3
0.22 ←	0.11 ←	1.2 ←	Σ	1.3

Prediction of biological target profile

Heuristics approach to chemical similarity

1,885 drugs on the market extracted from ChEMBL17

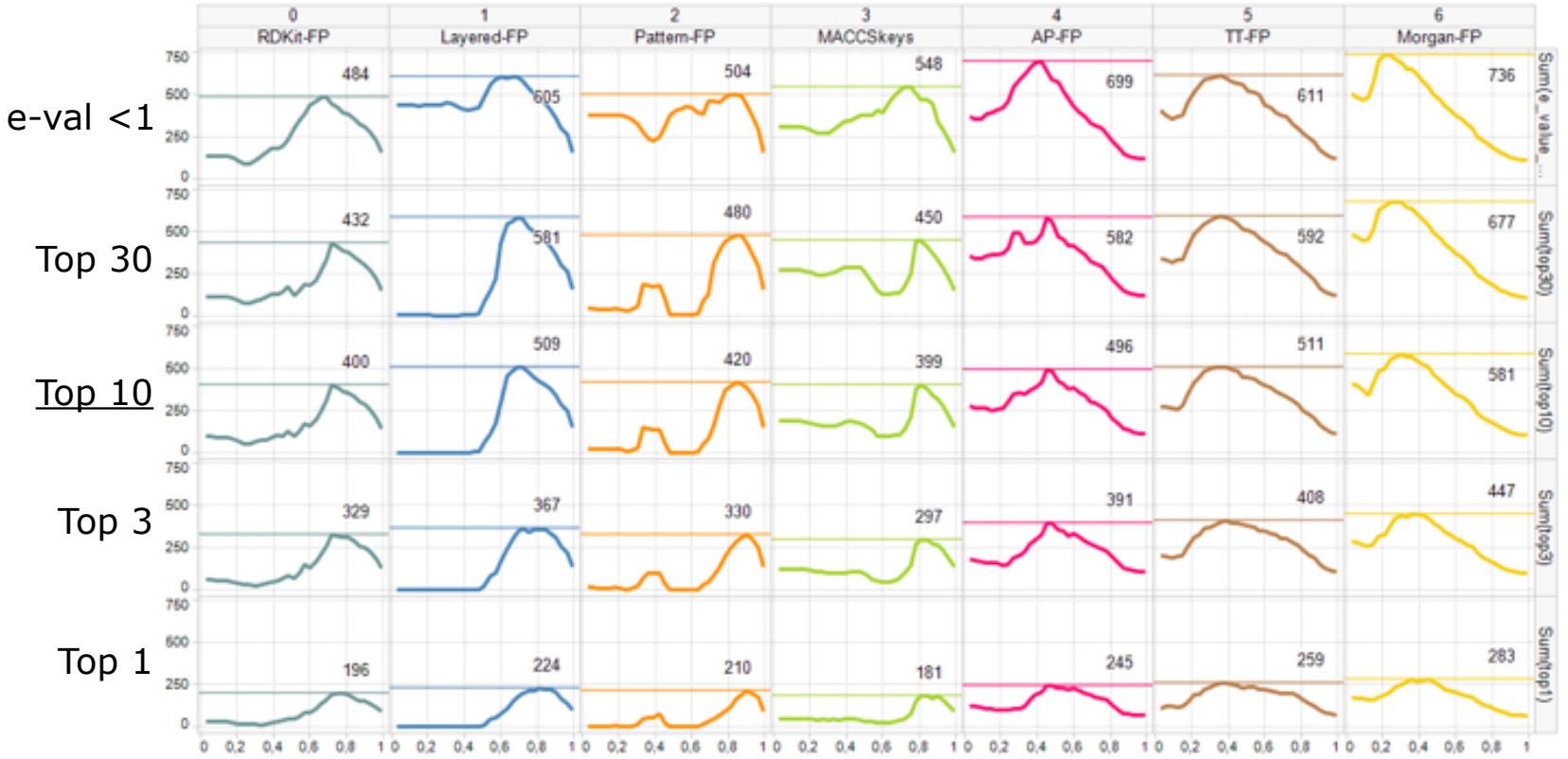
no Target in Ocean-DB:

928

Target in Ocean-DB:

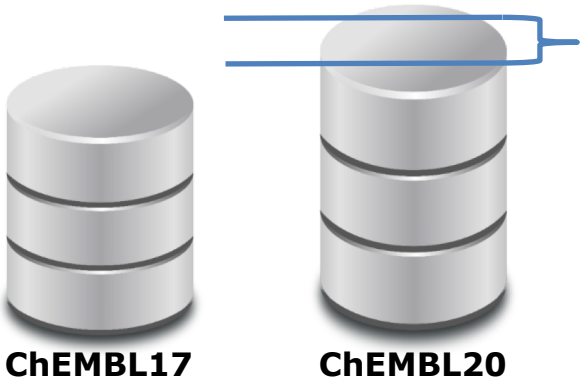
957

Search for the optimal Tanimoto coefficient



Morgan performs best!

External test set based on time stamp

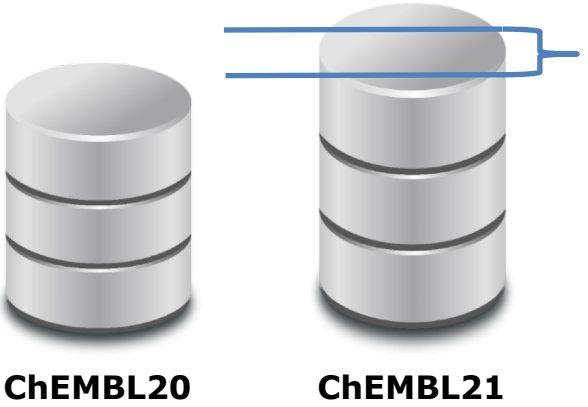


ΔChEMBL20_ChEMBL17

- new drugs (53 drug ← → target relationships)
- new compounds (43,126 act // 6,884 inact drug ← → target relationships)

- calculate statistics (raw score mean, raw score stdDev) for scoring
- find best FP/TC-cutoff with approved drugs

Time-Split Cross-Validation



ΔChEMBL21_ChEMBL20

- new compounds (19,901 act // 2,761 inact drug ← → target relationships)

Polypharmacology Prediction

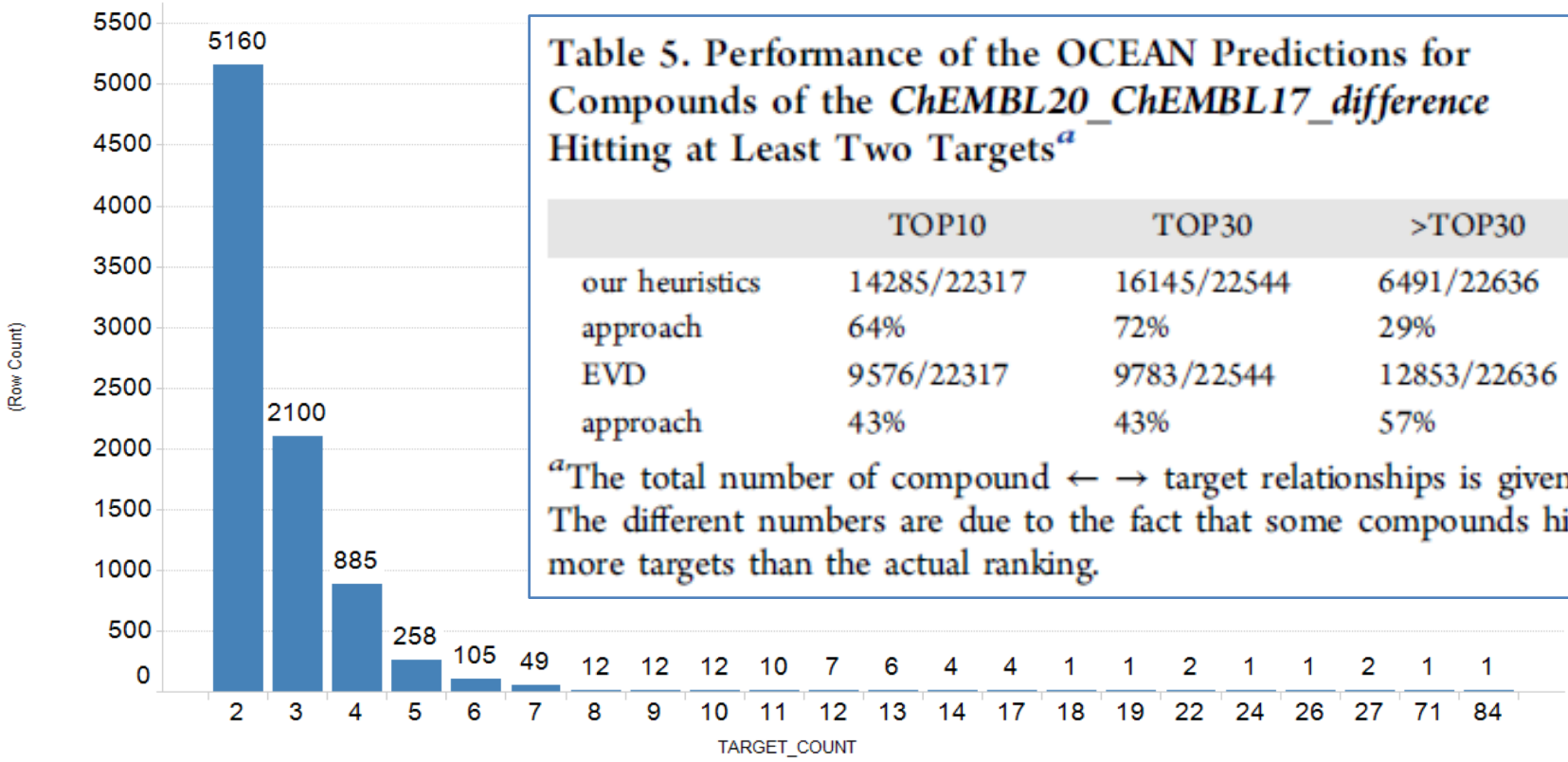


Table 5. Performance of the OCEAN Predictions for Compounds of the *ChEMBL20_ChEMBL17_difference* Hitting at Least Two Targets[a]

	TOP10	TOP30	>TOP30
our heuristics	14285/22317	16145/22544	6491/22636
approach	64%	72%	29%
EVD	9576/22317	9783/22544	12853/22636
approach	43%	43%	57%

[a]The total number of compound ← → target relationships is given. The different numbers are due to the fact that some compounds hit more targets than the actual ranking.

Distribution of the annotated number of targets per compound in OCEAN-DB (at least, 2 targets hit by one compound) based on ChEMBL20_ChEMBL17 difference.

OCEAN is able to identify multiple hitted targets for 64% of all cases in best 10 predicted targets and 72% within best 30 predicted targets

„One more thing"

Teaching Data Science for Chemistry/ChemicalBiology

SoSe

- *Einführung Data Science im Bereich Chemie und Chemische Biologie*
- 2 VL + 1 Ü [komplett von Czodrowski gehalten]

WiSe

- *Deskriptive Statistik*
- 2 VL [Statistik] + 1 Ü [„Python&Chemie-lastig“ durch Czodrowski]

SoSe

- *Statistisches Lernen*
- 4 VL + 2 Ü