



Chemoinformatics Strasbourg Summer School 2018
University of Strasbourg, 25 June - 29 June 2018

HTS-likeness: physicochemical parameters to create libraries

Pavel Polishchuk, Mariia Matveieva

Institute of Molecular and Translational Medicine
Faculty of Medicine and Dentistry
Palacky University

pavlo.polishchuk@upol.cz

Design steps of general purpose HTS library creation

Physicochemical filters

drug-/lead-likeness to select favorable compounds

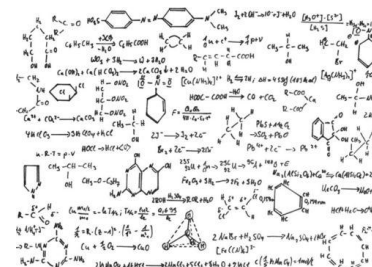
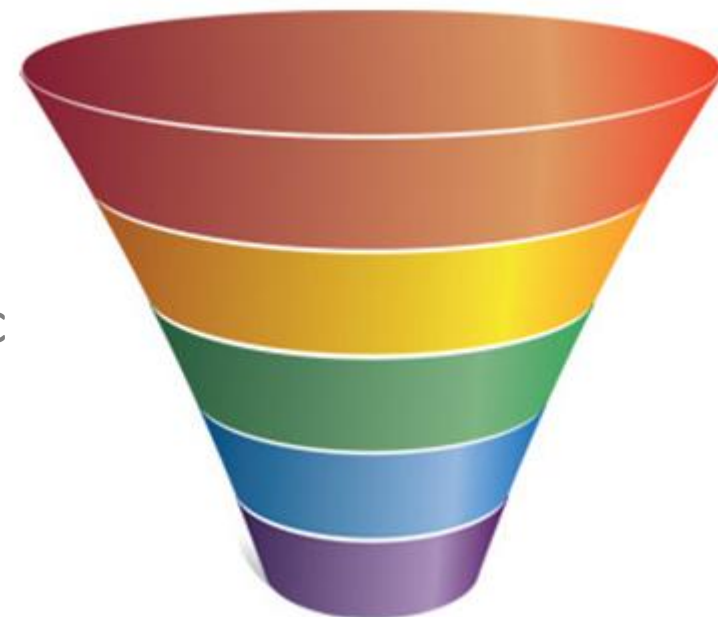


Structural filters/predictive models

to remove compounds potentially toxic, unstable, reactive, false positives, etc

Diversity selection

to better cover available chemical space



Physicochemical rules/filters/predictive models

	Lipinski	Oprea drug-like	Oprea lead-like	Walters
acceptor count	≤ 10	2-9	0-8	≤ 10
donor count	≤ 5	0-2	0-5	≤ 5
logP	≤ 5		-3.5 - 4.5	-5 - 5
molecular weight	≤ 500		≤ 450	200 - 500
RTB		2-8		≤ 8




drugs/drug candidates
oral bioavailability

leads
capacity for optimization

Ideal general purpose HTS library

- small
- high chances to find hits
- return true hits for variety of assays
- no promiscuous compounds
- soluble
- stable



minimum set of
requirements

PubChem data set

94 PubChem assays

assay type	number of assays	
	training set	test set
cell-based	27	20
biochemical	22	21
other	0	3
total	49	45
compounds	230 325	72 760
hit rates, %	0.004-5.10	0.014-2.55

no PAINS, no frequent hitters

Compounds are from MLSMR library which was created using different strategies of compounds selection (including physicochemical filters and diversity selection)

Physicochemical parameters

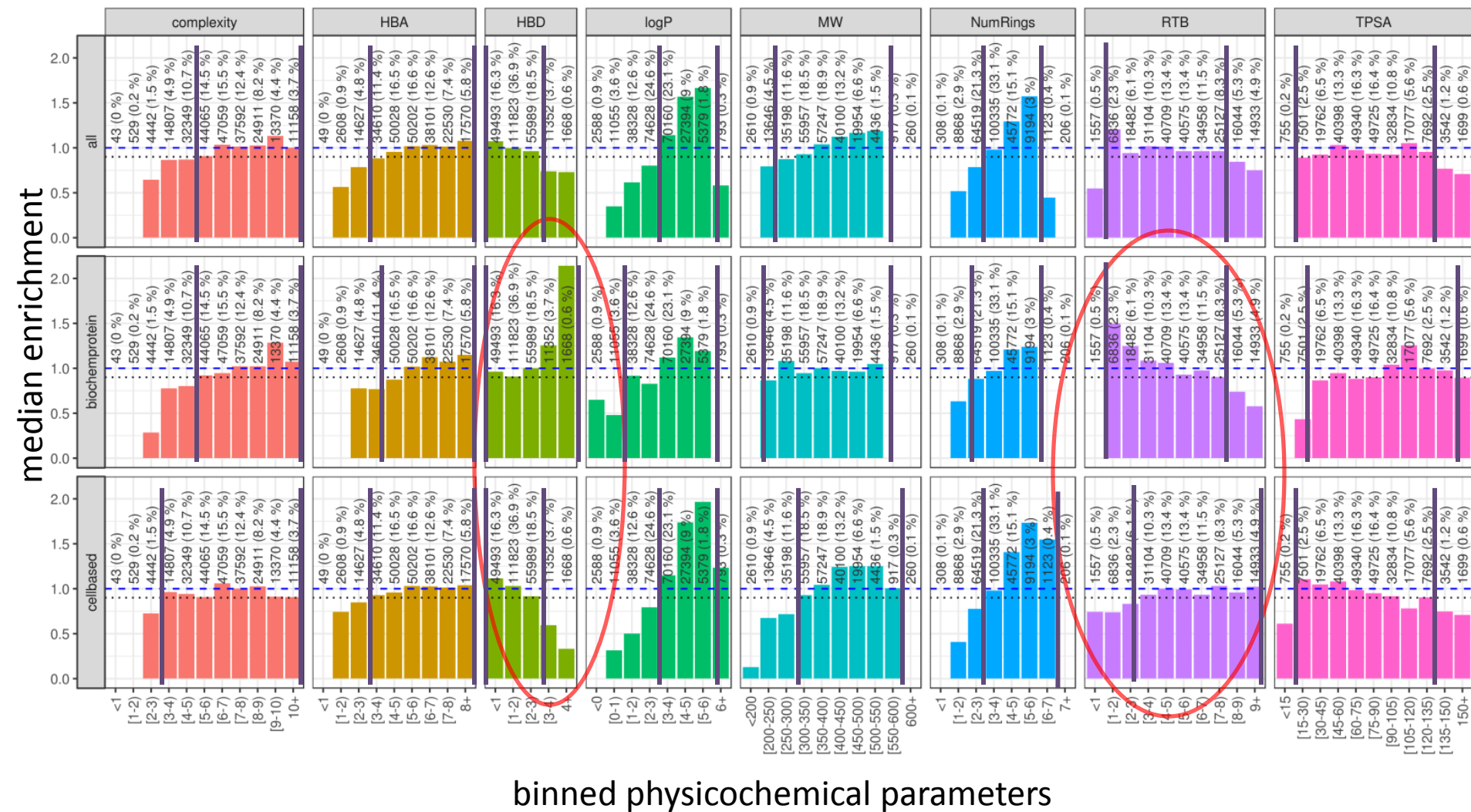
Physicochemical properties calculated with RDKit

- H-bond donors count (HBD)
- H-bond acceptors count (HBA)
- Complexity = HBD + HBA
- logP
- MW
- Topological polar surface area (TPSA)
- Rings count (NumRings)
- Rotatable bonds count (RTB)

$$\text{Enrichment} = \frac{\text{hit rate for selected compounds}}{\text{baseline hit rate}}$$

PubChem training set

Distribution of median enrichment vs. values of physicochemical parameters



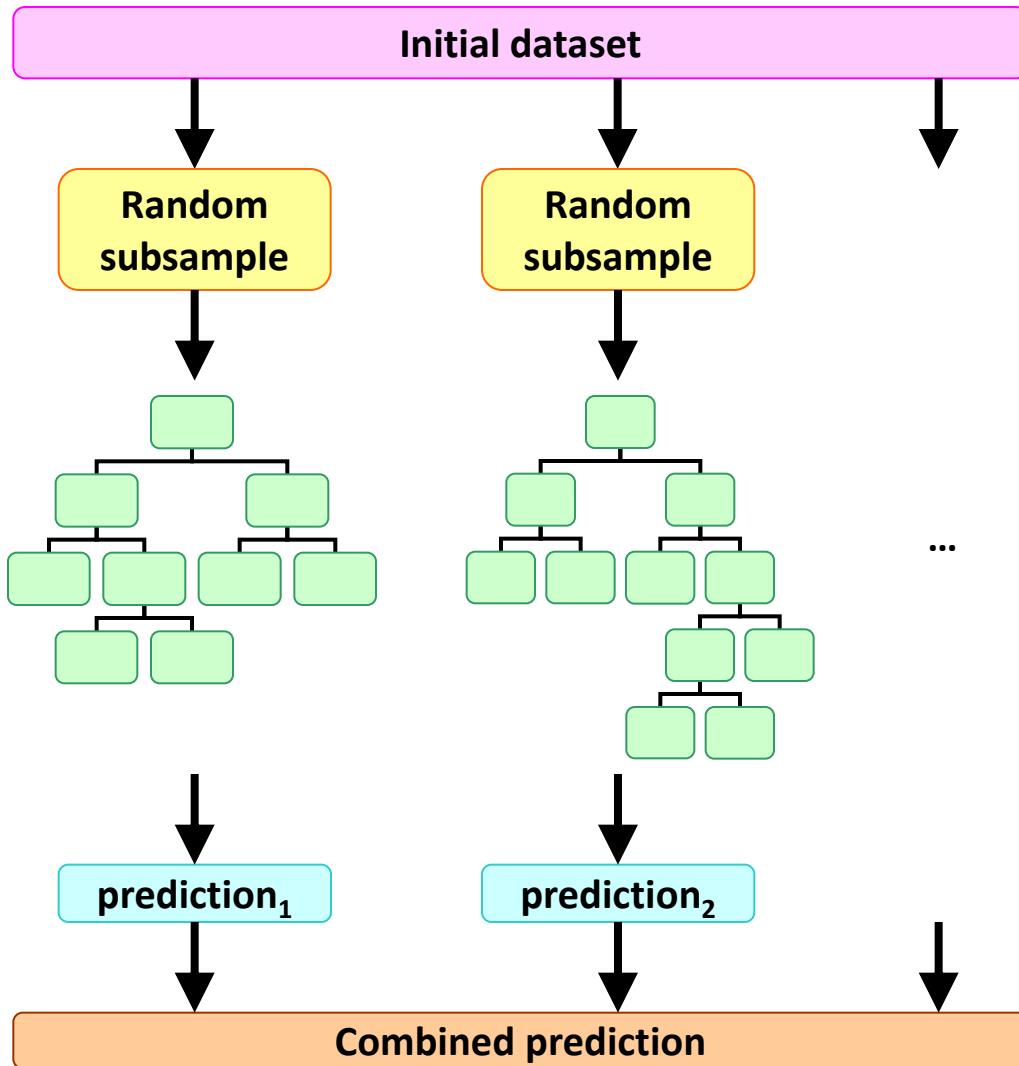
Manually derived rules from the PubChem training set

	all	biochemical	cell-based	Lipinski	Oprea DL	Oprea LL	Walters
complexity	5-10	5-10	3-10				
acceptor count	3-8	4-8	3-8	<= 10	2-9	<= 8	<= 10
donor count	0-2	0-4	0-2	<= 5	<= 2	<= 5	<= 5
logP	3-6	1-6	>3	<= 5		-3.5 - 4.5	-5 - 5
molecular weight	250-550	200-550	300-600	<= 500		<= 450	200 - 500
Ring count	3-5	2-5	3-6				
RTB	1-7	1-6	3-9		2-8		<= 8
TPSA, A ²	15-135	30-150	15-135				

Application of the derived rules to the test set

rule set	number of selected compounds	datasets median enrichment		
		all	biochemical	cell-based
all	14 852 (20.4%)	1.18	1.15	1.24
biochemical	26 407 (36.3%)	1.01	1.06	0.99
cell-based	21 941 (30.2%)	1.00	1.05	0.93

Random forest model



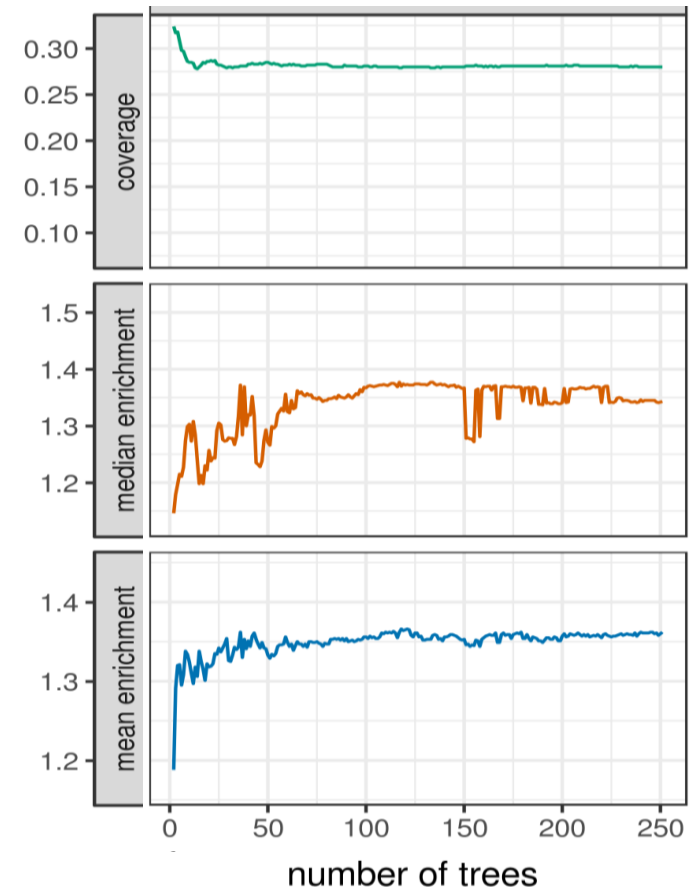
Random subsample = 2/3

Nvars = 3

Ntrees = 250

min_parent_samples = 3000

min_child_samples = 1000



PubChem test set prediction

Manually derived rules

rule set	number of selected compounds	datasets median enrichment		
		all	biochemical	cell-based
all	14 852 (20.4%)	1.18	1.15	1.24
biochemical	26 407 (36.3%)	1.01	1.06	0.99
cell-based	21 941 (30.2%)	1.00	1.05	0.93

Random Forest prediction

model	number of selected compounds	dataset median enrichment		
		all	biochemical	cell-based
all assays	20 337 (28.0%)	1.34	1.15	1.45
biochemical assays	12 528 (17.2%)	1.36	1.38	1.27
cell-based assays	29 179 (40.2%)	1.16	1.08	1.36

Common physicochemical filters

rule set	number of selected compounds	datasets median enrichment		
		all	biochemical	cell-based
Lipinski	61 624 (84.8%)	0.98	1.00	0.98
Oprea drug-like	55 984 (77.0%)	0.95	0.92	0.97
Oprea lead-like	50 566 (69.5%)	0.99	1.01	0.89
Walters	57 533 (79.1%)	1.02	1.05	1.03

NCI60

	NCI60 data set (-logGI ₅₀)
inactive threshold	≤ 5
active threshold	> 7
number of assays with > 9000 compounds tested	68
number of compounds in the data set	46 982
hit rates	1.4% - 6.1%

NCI60 prediction

Common physicochemical filters

rule set	number of selected compounds	median enrichment
Lipinski	34 497 (73.4%)	0.97
Oprea drug-like	26 951 (57.4%)	1.03
Oprea lead-like	29 295 (62.4%)	0.98
Walters	32 824 (69.9%)	1.00

Manually derived rules

rule set	number of selected compounds	median enrichment
all	7 043 (15.0%)	1.07
biochemical	13 232 (28.2%)	1.03
cell-based	9 080 (19.3%)	1.03

Random Forest prediction

model	number of selected compounds	median enrichment
all assays	18 525 (39.4%)	1.29
biochemical assays	16 412 (34.9%)	1.65
cell-based assays	23 258 (49.5%)	1.52

Quantitative estimate of drug-likeness (QED)

ARTICLES

PUBLISHED ONLINE: 24 JANUARY 2012 | DOI: 10.1038/NCHEM.1243

nature
chemistry

Quantifying the chemical beauty of drugs

G. Richard Bickerton¹, Gaia V. Paolini², Jérémy Besnard¹, Sorel Muresan³ and Andrew L. Hopkins^{1*}

Table 1 | Optimized desirability function weightings by Shannon entropy.

	Shannon entropy	Rank	M_r	ALOGP	HBD	HBA	PSA	ROTB*	AROM*	ALERTS
QED _{w,max}	293.42	1	0.50	0.25	0.50	0.00	0.00	0.50	0.25	1.00
QED _{w,mo}	293.03	1-1000	0.66	0.46	0.61	0.05	0.06	0.65	0.48	0.95
QED _{w,u}	283.08	81,657	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

data set	median enrichment at QED / coverage				
	>= 0.5	>= 0.6	>= 0.7	>= 0.8	>= 0.9
PubChem training set	0.90 / 0.82	0.83 / 0.67	0.76 / 0.46	0.66 / 0.24	0.59 / 0.04
PubChem test set	0.93 / 0.74	0.87 / 0.58	0.76 / 0.39	0.77 / 0.19	0.53 / 0.04
NCI60	0.60 / 0.55	0.53 / 0.38	0.56 / 0.23	0.38 / 0.09	0.42 / 0.01

Conclusion

- HTS-like chemical space is partially overlapped with a drug-like chemical space
- HTS-likeness rules may reduce the size of a library and improve hit rates relatively to drug-likeness filters
- Random Forest models more accurately estimate HTS-likeness than manually derived rules

Acknowledgments



Finton Sirockin
Petr Ertl



This work is supported by the grant of the Ministry of education, youth and sport of Czech Republic (MSMT-5727/2018-2)

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY