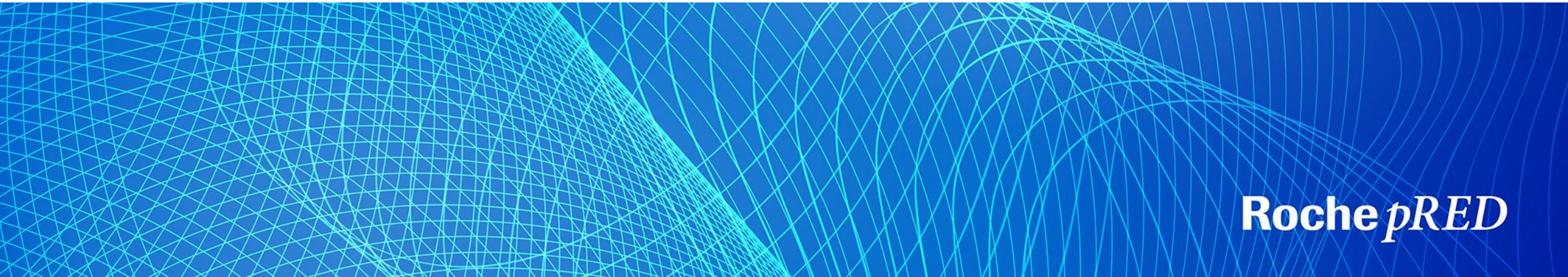
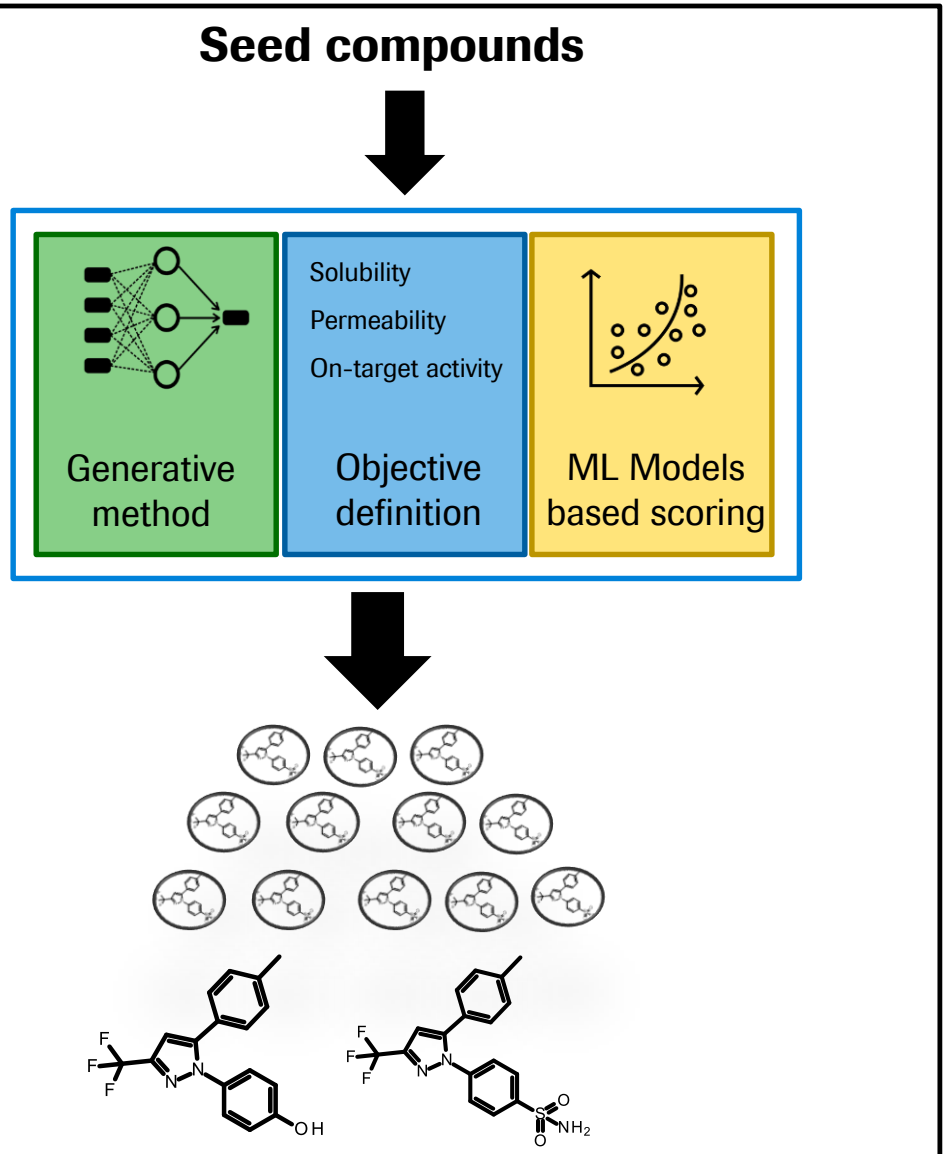
Modifying the Synthetic Accessibility Score to Identify Undesirable Virtual Compounds

14.10.21

Mahendra Awale



De-novo compound optimization



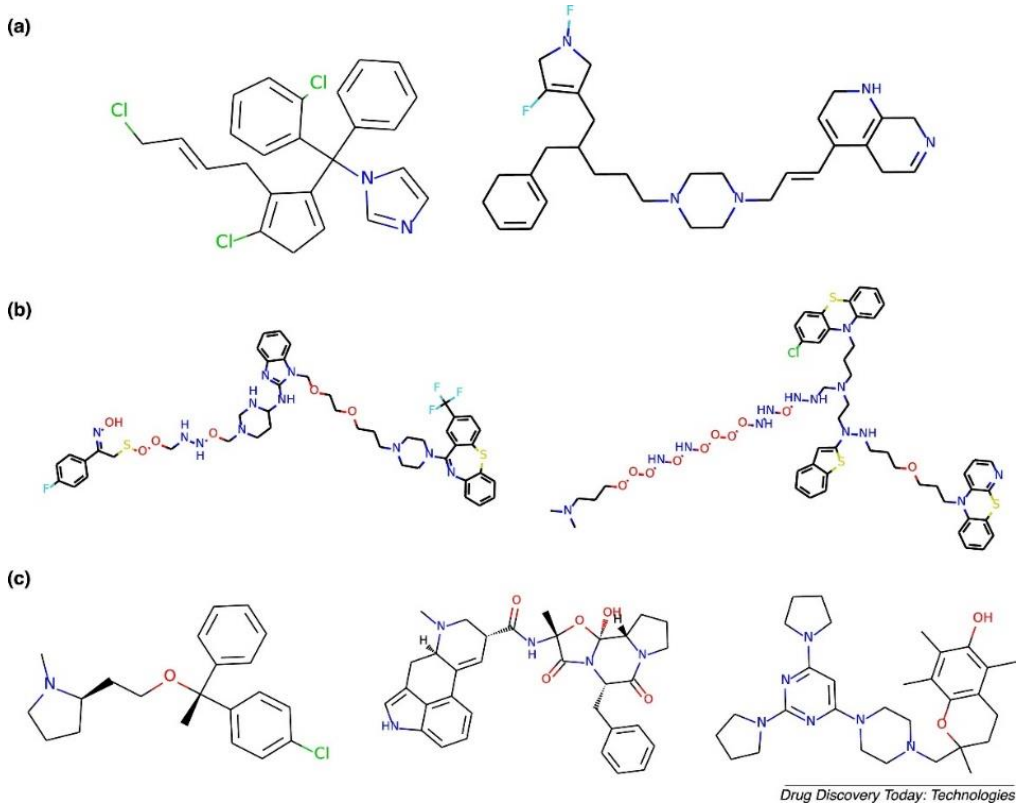
	Atom based	Fragment based	Reaction based
Gradient free	GB-GA[21]	MOARF[6]	SYNOPSIS[28]
	ChemGE[22]	CReM[27]	AutoGrow4[29]
	MSO[23]	CoG[19]	
Gradient based	Segler *et al.*[32]	Pogány *et al.*[41]	DINGOS[44]
	REINVENT[33]	JT-VAE[42]	Molecule Chef[45]
	GraphVAE[34]	DeepFMPO[43]	ChemBO[46]
	MolGAN[35]		REACTOR[47]
	CG-VAE[36]		PGFS[48]
	Li *et al.*[37]		DoGs[49]
	MolDQN[38]		
	GraphINVENT[39]		
	Optimol[51]		

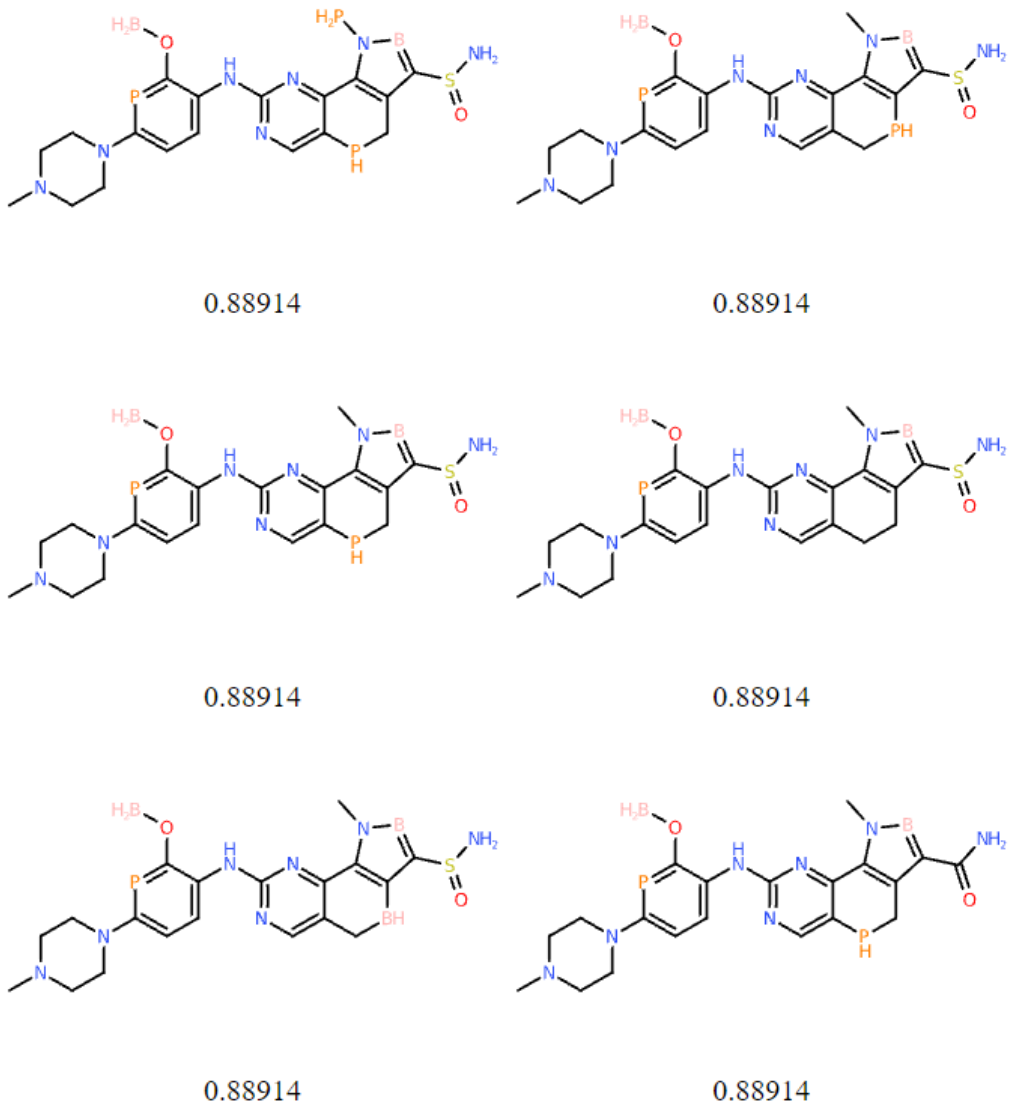
* ***De novo* molecular design and generative models**, Joshua Meyers *etal,* **2021**, Drug Discovery today

Drug Discovery Today: Technologies
Volumes 32–33, December 2019, Pages 55-63

On failure modes in molecule generation and optimization

Philipp Renz [1], Dries Van Rompaey [2], Jörg Kurt Wegner [2], Sepp Hochreiter [1], Günter Klambauer [1]





*https://www.benevolent.com/guacamol

Weeding out undesirable compounds

During Optimization: Introduce additional parameters in objective

- SA score (synthetic accessibility score)
- QED score (Quantitative estimation of druglikness)
- Descriptors
- Functional groups/Smarts filters
- RAscore

References:

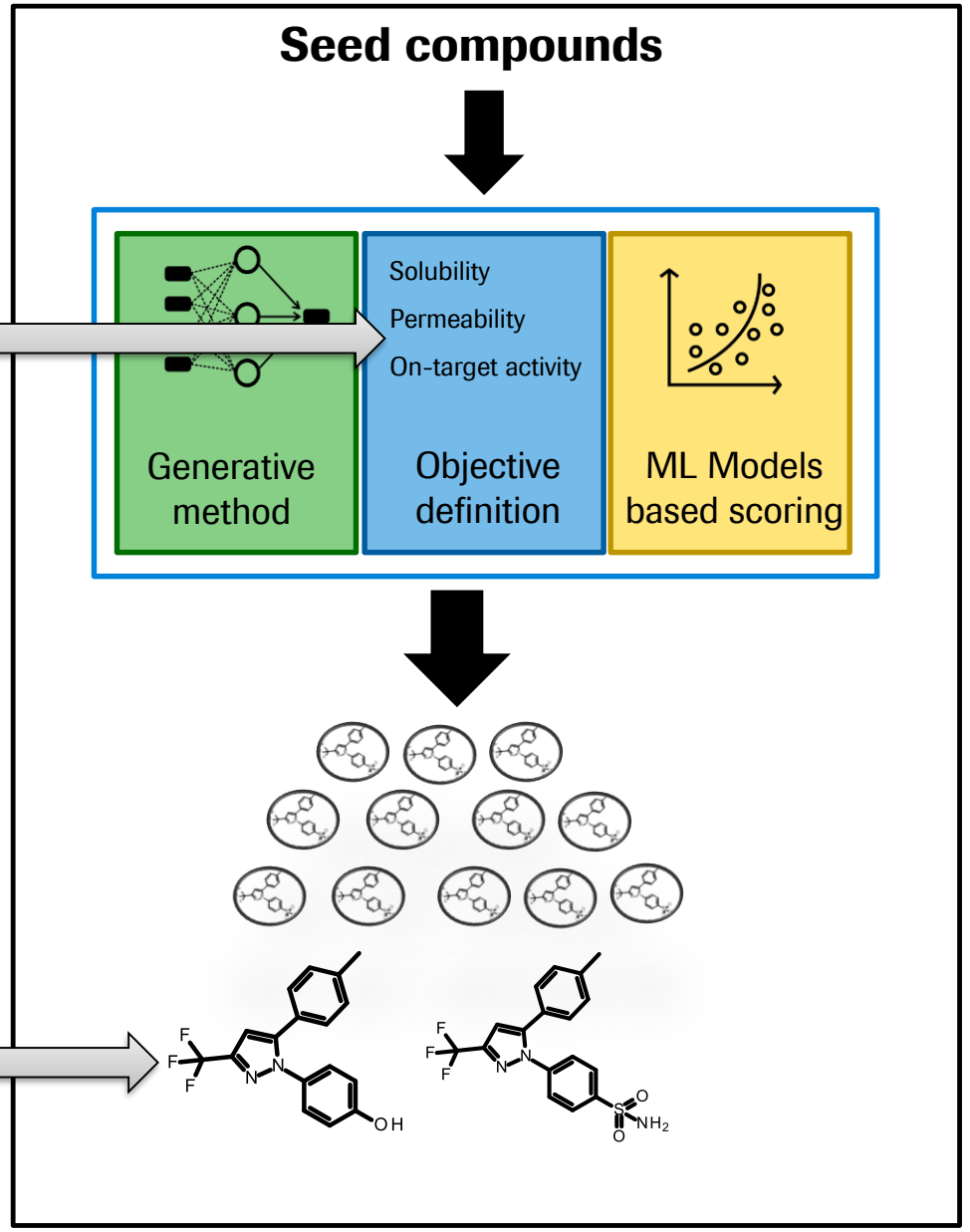
Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, Peter ertl, *J. Cheminfo*, **2009**

Quantifying the chemical beauty of drugs, Andrew Hopkins, Nat. Chem. **2012**

Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning, Chem. Sci. Amol Thakkar, **2021**

After Optimization/ filter or rank compounds

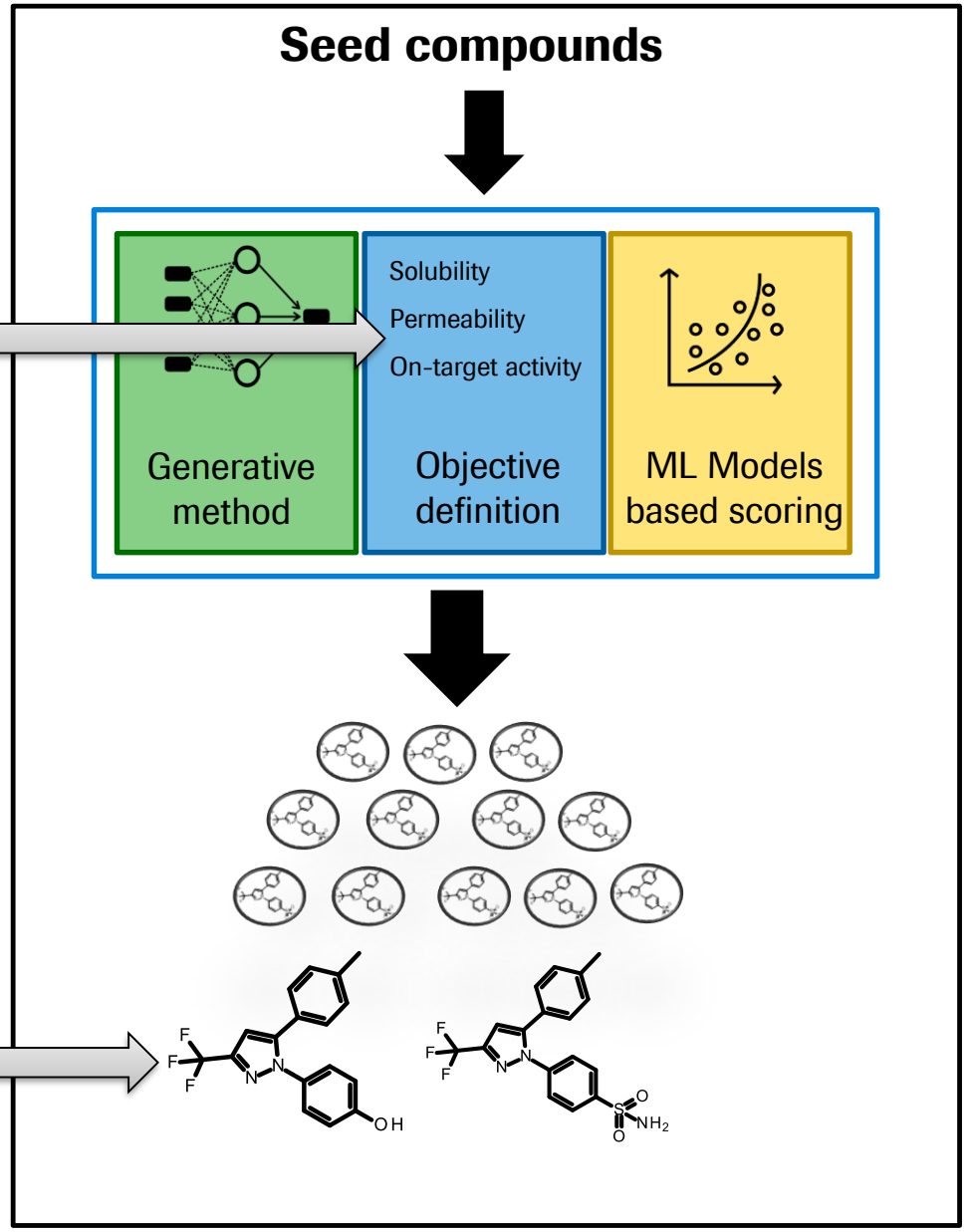
- SA score (synthetic accessibility score)
- QED score (Quantitative estimation of druglikness)
- Descriptors
- Functional groups/Smarts filters
- RAscore



Weeding out undesirable compounds

During Optimization: Introduce additional parameters in objective

- SA score (synthetic accessibility score)
- QED score (Quantitative estimation of druglikness)
- Descriptors
- Functional groups/Smarts filters
- RAscore



References:

Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, Peter ertl, *J. Cheminfo*, **2009**

Quantifying the chemical beauty of drugs, Andrew Hopkins, Nat. Chem. **2012**

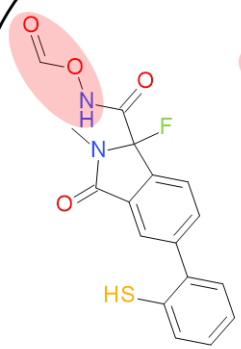
Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning, Chem. Sci. Amol Thakkar, **2021**

After Optimization/ filter or rank compounds

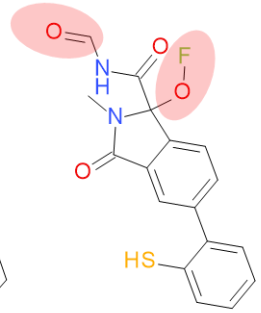
- SA score (synthetic accessibility score)
- QED score (Quantitative estimation of druglikness)
- Descriptors
- Functional groups/Smarts filters
- RAscore

Many undesirable compounds are not captured by SA-score and QED

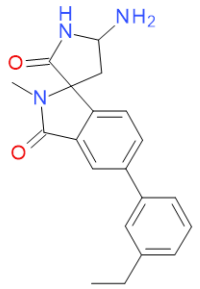
a) Valsartan SMARTS MPO



SA score: 3.79
QED: 0.38

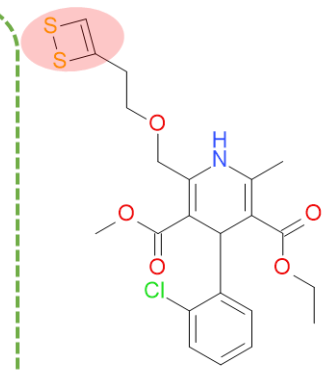


SA score: 3.86
QED: 0.65

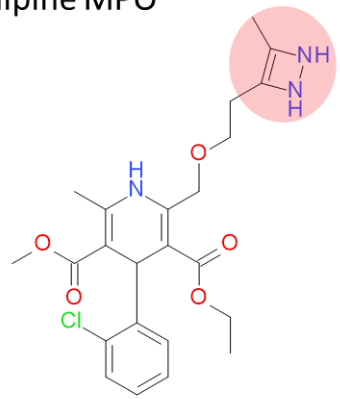


SA score: 4.03
QED: 0.88

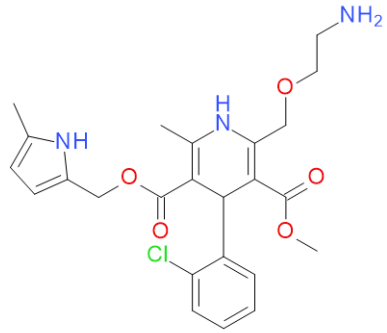
b) Amlodipine MPO



SA score: 3.34
QED: 0.32

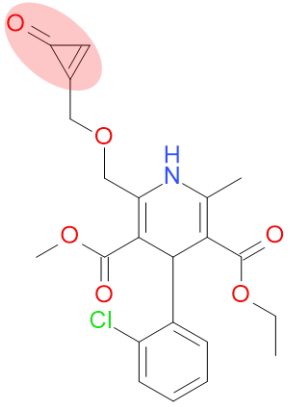


SA score: 3.37
QED: 0.39

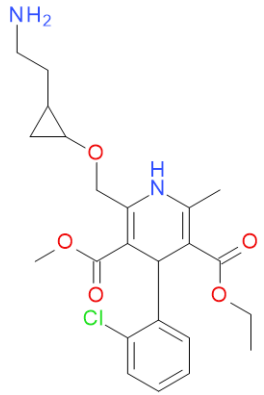


SA score: 3.35
QED: 0.38

b) Amlodipine MPO

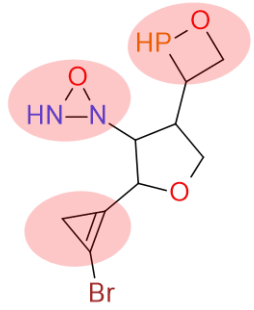


SA score: 3.04
QED: 0.64

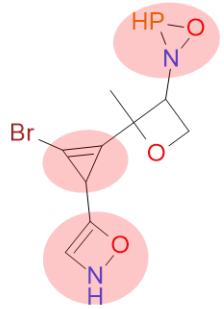


SA score: 3.71
QED: 0.56

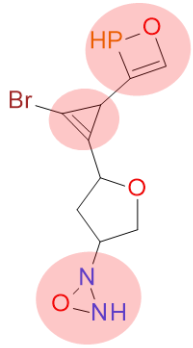
c) Very high QED for compounds with undesirable substructures*



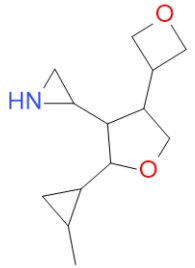
QED: 0.86
SA score: 6.18



QED: 0.94
SA score: 5.91



QED: 0.94
SA score: 6.06



QED: 0.72
SA score: 4.70

Low synthetic accessibility

10

1

High synthetic accessibility

SA score

High Quality compound

1

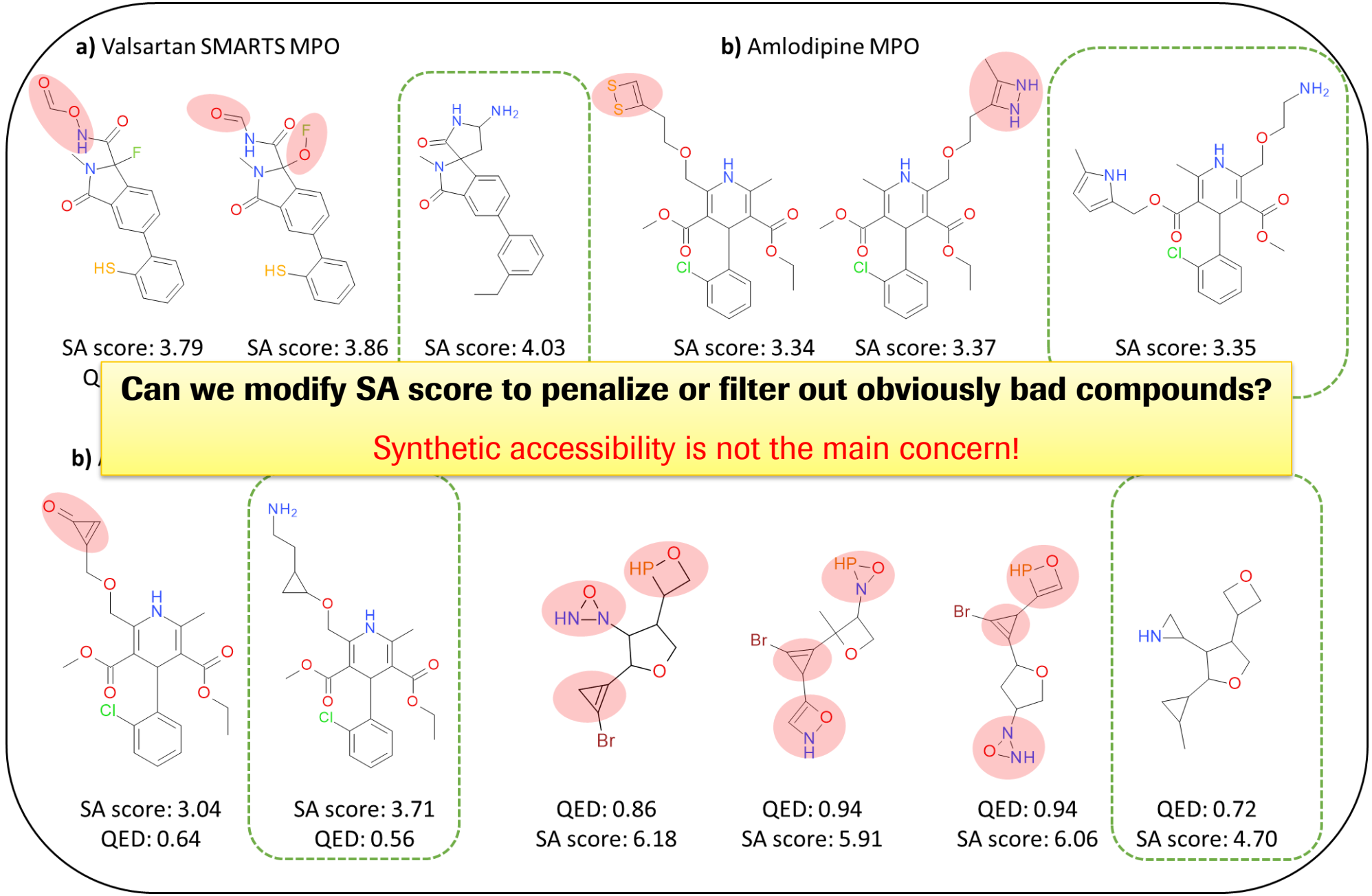
0

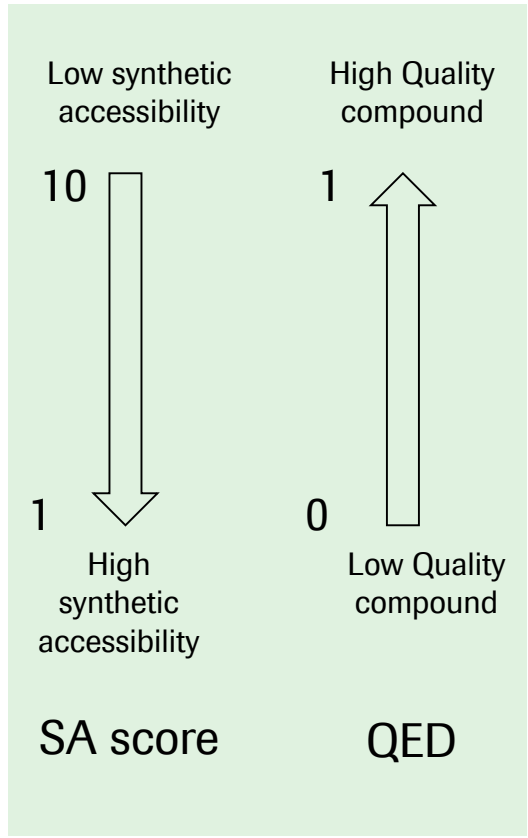
Low Quality compound

QED

*https://www.benevolent.com/guacamol

Many undesirable compounds are not captured by SA-score and QED



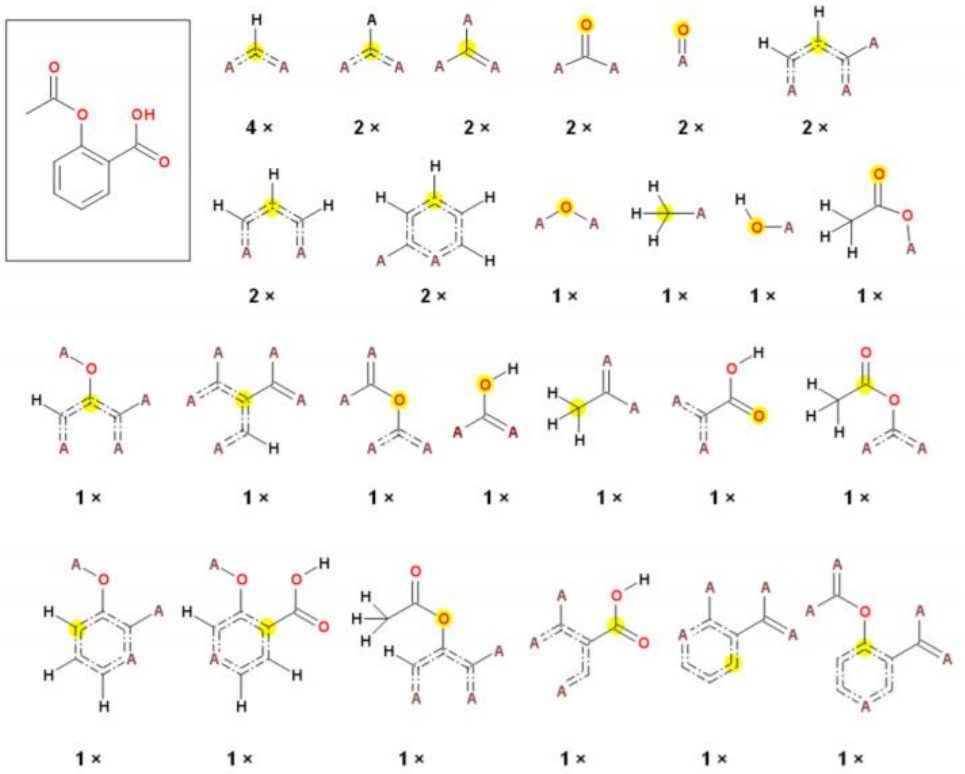


*https://www.benevolent.com/guacamol

SA score: How does it work?

SAscore = fragmentScore − complexityPenalty

Sum of contributions from different fragments present in a molecule



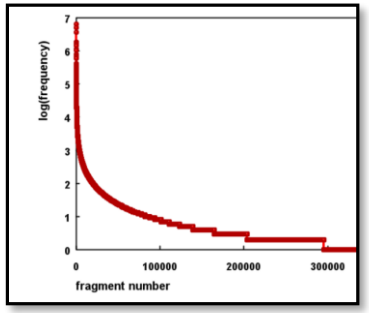
Fragments are defined as morgan fingerprint bits

Up to radius 2

*If a fragment not present in FCD penalty = -4

934,046 compounds from PubChem

fragment compounds



Compute frequencies of fragments

$$f_i = \log\left(\frac{frequency\ of\ fragment\ i}{no.\ of\ unique\ fragments\ in\ database\ * 0.8}\right)$$

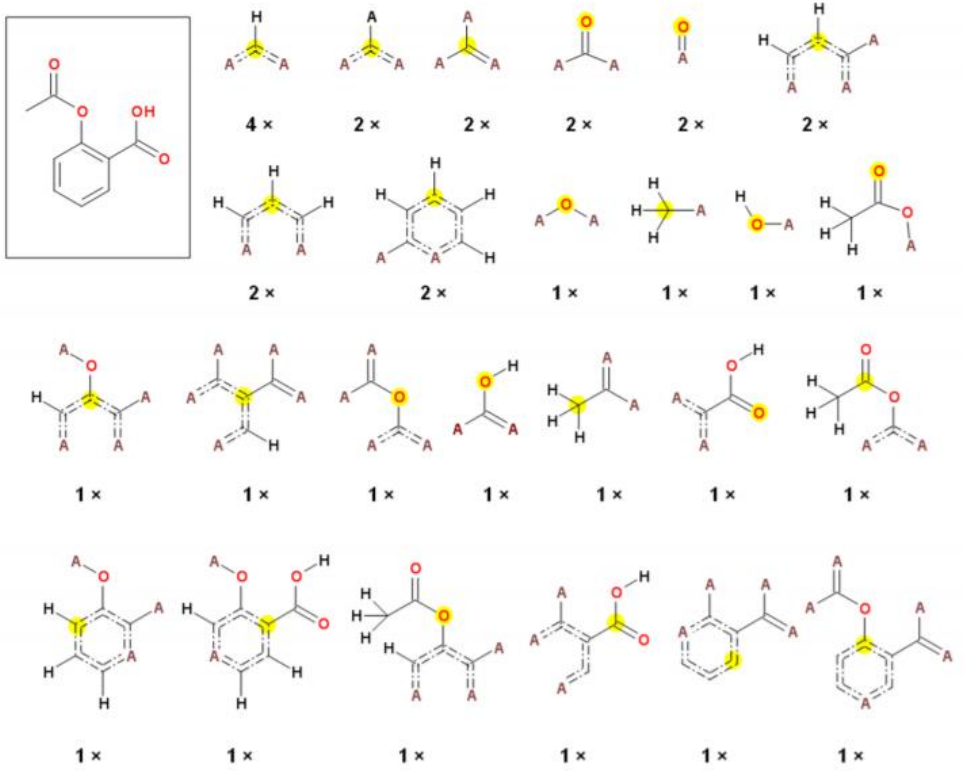
Fragment (morgan bit id)	contribution
68675643	-1.34
253103978	-3.56
485463469	0.49
882399112	0.82
---	--

fragment contribution database (FCD)

SA score: How does it work?

SAscore = fragmentScore − complexityPenalty

Sum of contributions from different fragments present in a molecule

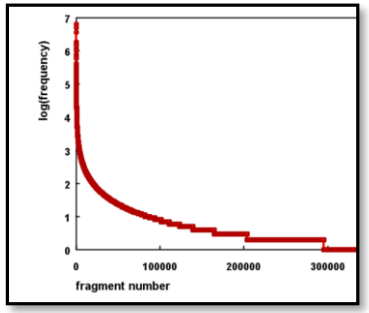


Fragments are defined as morgan fingerprint bits

Up to radius 2

934,046 compounds from PubChem

fragment compounds



Compute frequencies of fragments

$$f_i = \log\left(\frac{frequency\ of\ fragment\ i}{no.\ of\ unique\ fragments\ in\ database\ * 0.8}\right)$$

Fragment (morgan bit id)	contribution
68675643	-1.34
253103978	-3.56
485463469	0.49
882399112	0.82
---	--

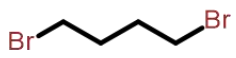
fragment contribution database

Modification 1: Change of Database

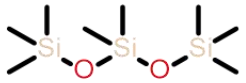
PubChem (1M cpds)

⬇

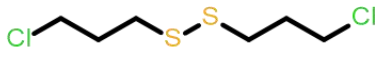
ZINC InStock (10M cpds)



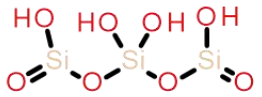
1. CHEMBL3185714
(0.95/-1.80)



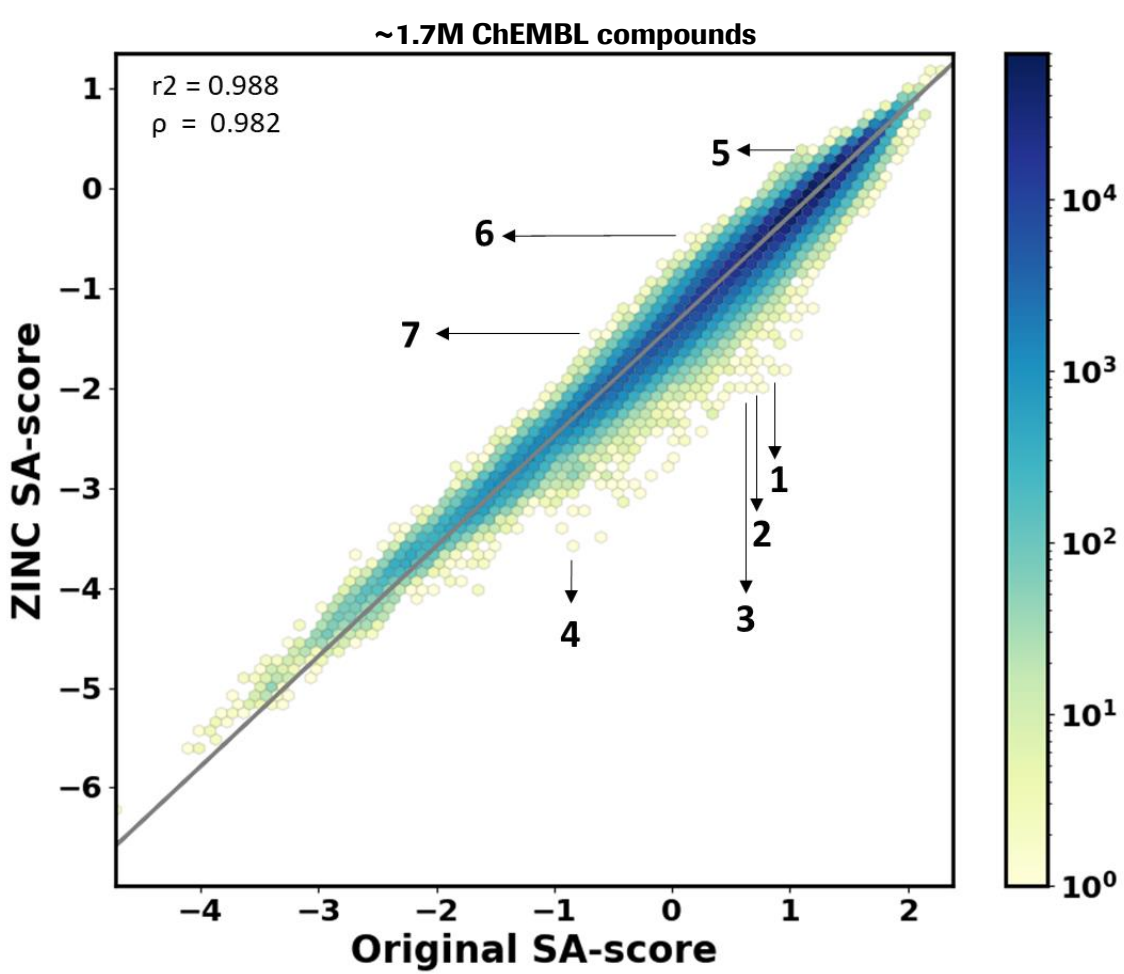
2. CHEMBL2142985
(0.79/-2)

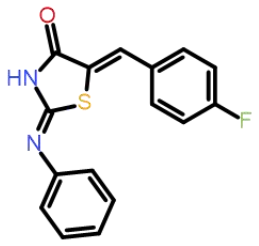


3. CHEMBL1966651
(0.68/-2.03)

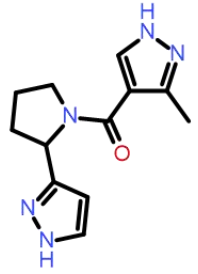


4. CHEMBL2109882
(-0.84/-3.58)

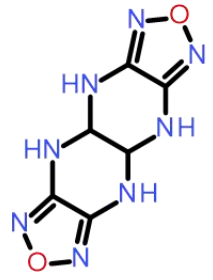




5. CHEMBL3969996
(1.10/0.38)

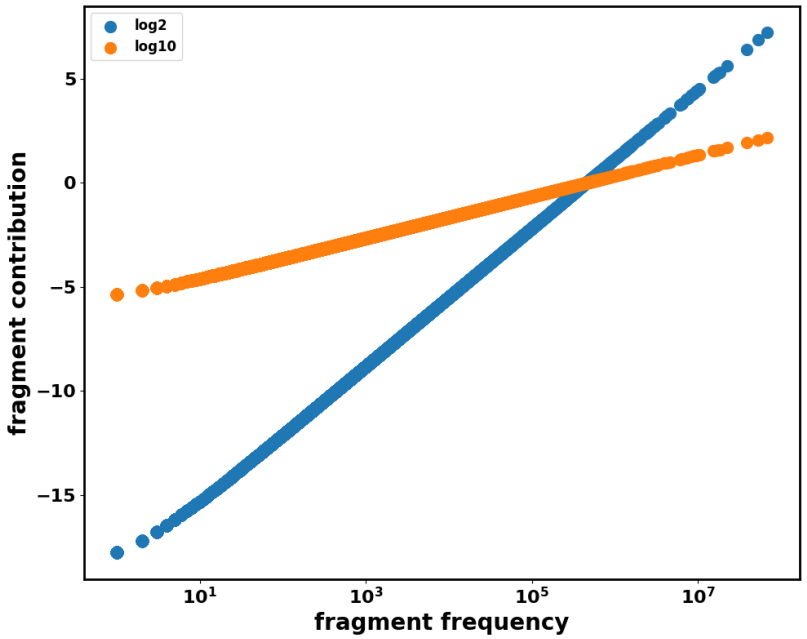


6. CHEMBL3448799
(0.14/-0.49)



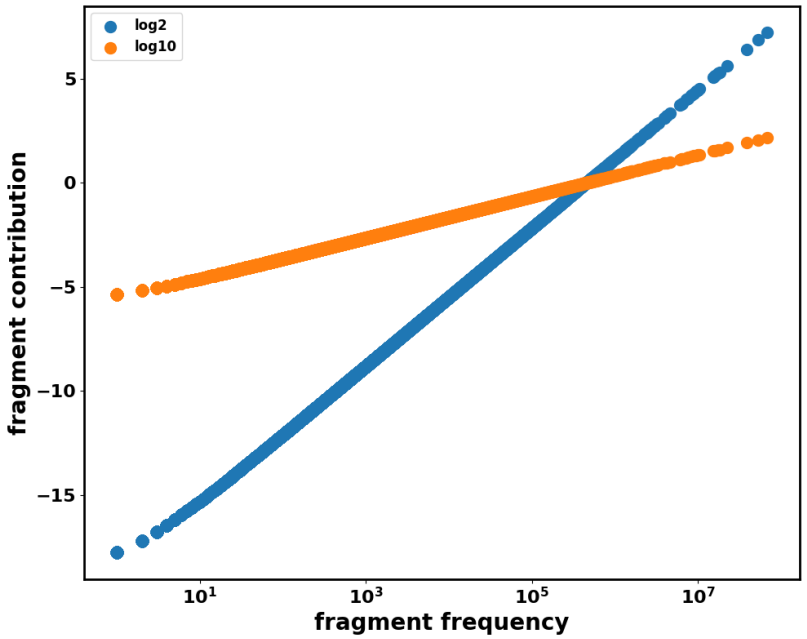
7. CHEMBL1984992
(-0.46/-1.20)

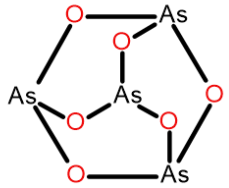
Modification 2: Change of log function



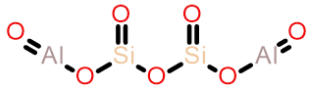
$$f_i = \log\left(\frac{frequency\ of\ fragment\ i}{no.\ of\ unique\ fragments\ in\ database\ *\ 0.8}\right)$$

Modification 2: Change of log function



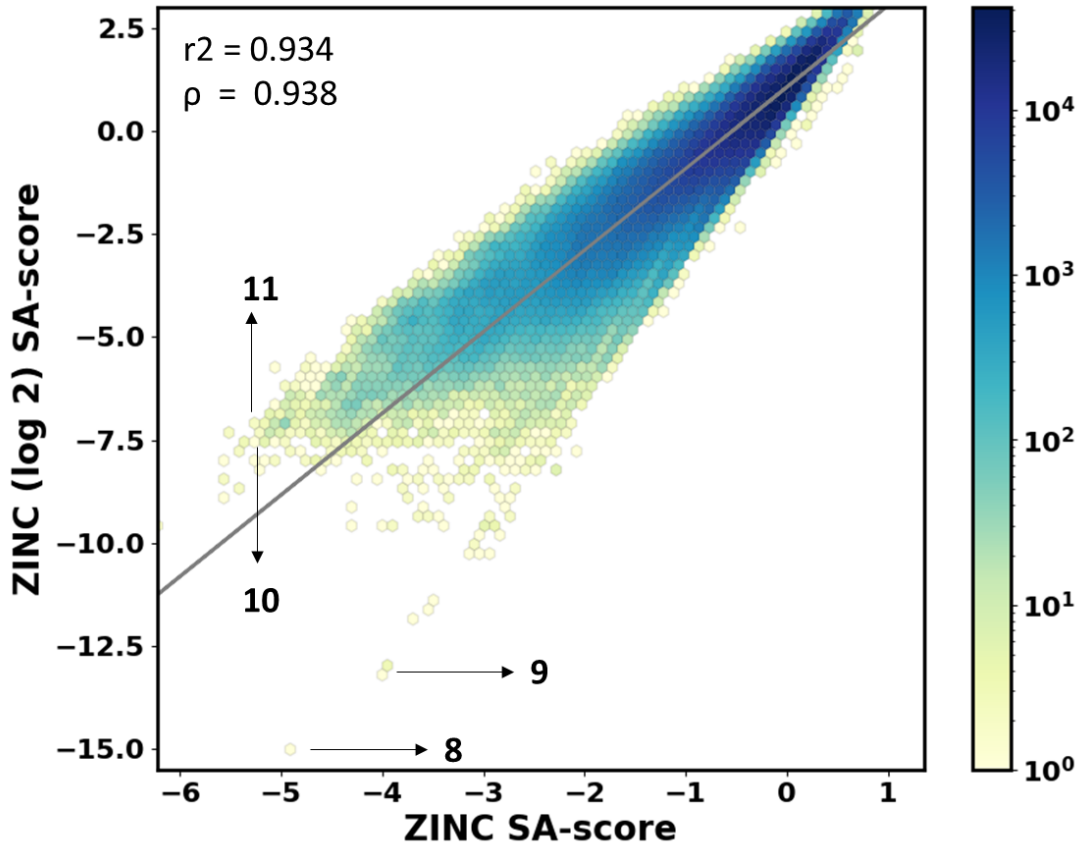


8. CHEMBL2362016
(-4.91/-15.01)

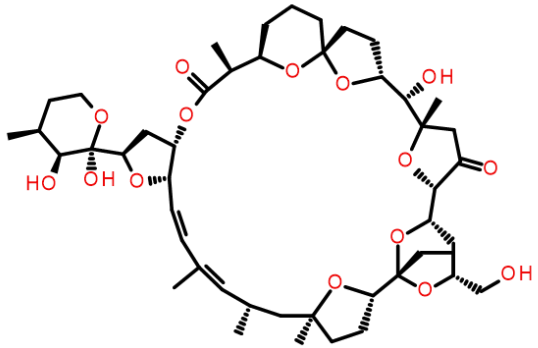


9. CHEMBL3833410
(-3.99/-12.95)

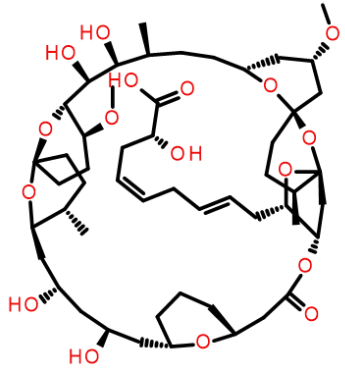
~1.7M ChEMBL compounds



Roche



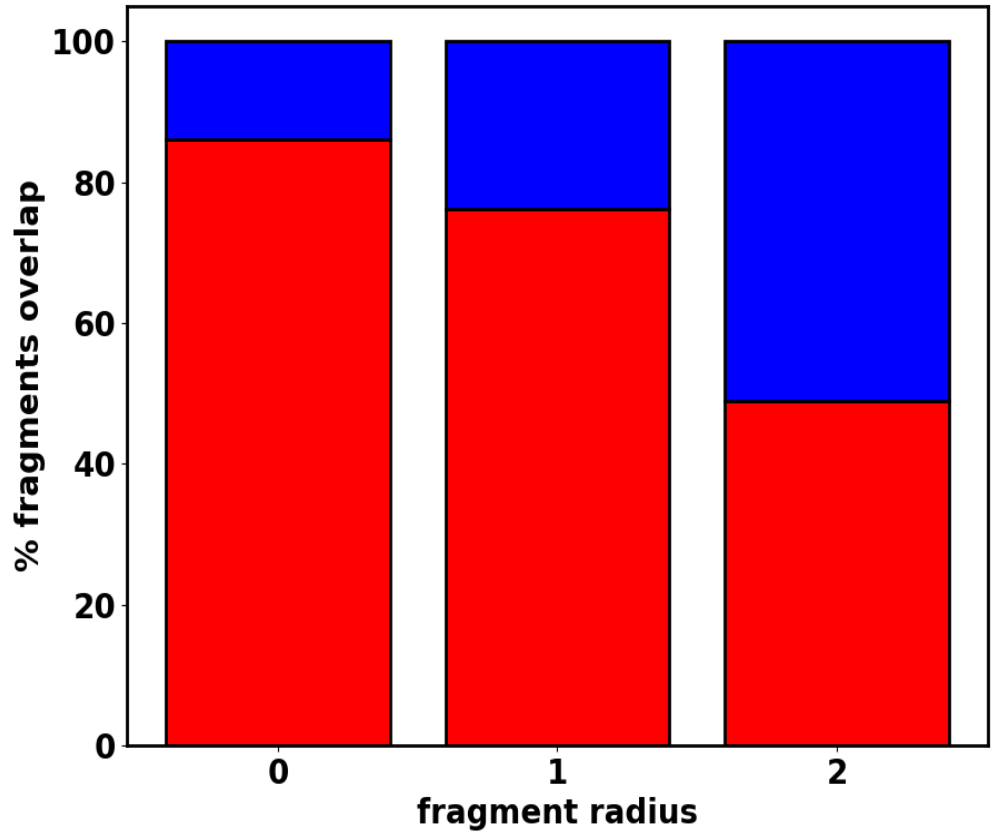
10. CHEMBL2448517
(-5.40/-7.68)



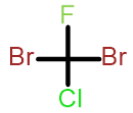
11. CHEMBL2062531
(-5.48/-7.28)

Modification 3:
remove radius 2 fragments from consideration

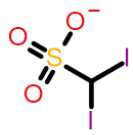
Overlap of ZINC InStock fragments with fragments from Original Publication



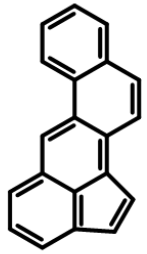
Modification 3: remove radius 2 fragments from consideration



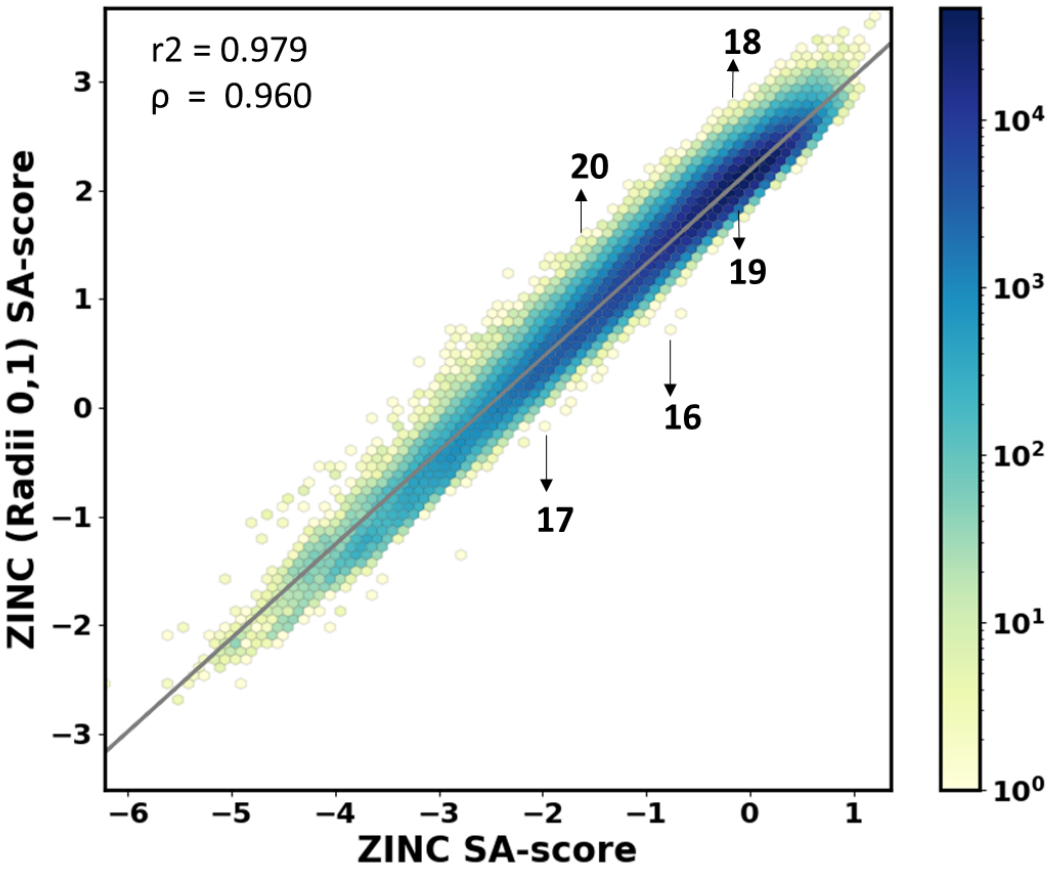
16. CHEMBL345249
(-0.80/0.69)
Total fragments: 10
Unique fragments: 8
R2 fragments: 0

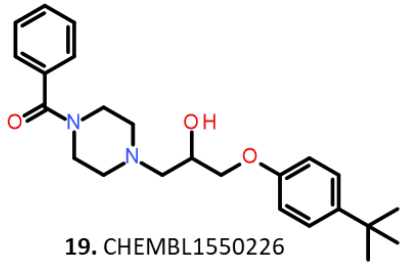


17. CHEMBL2105030
(-1.98/-0.22)
Total fragments: 15
Unique fragments: 11
R2 fragments: 1

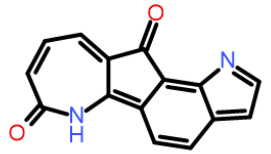


18. CHEMBL153437
(0.02/2.87)
Total fragments: 60
Unique fragments: 20
R2 fragments: 29



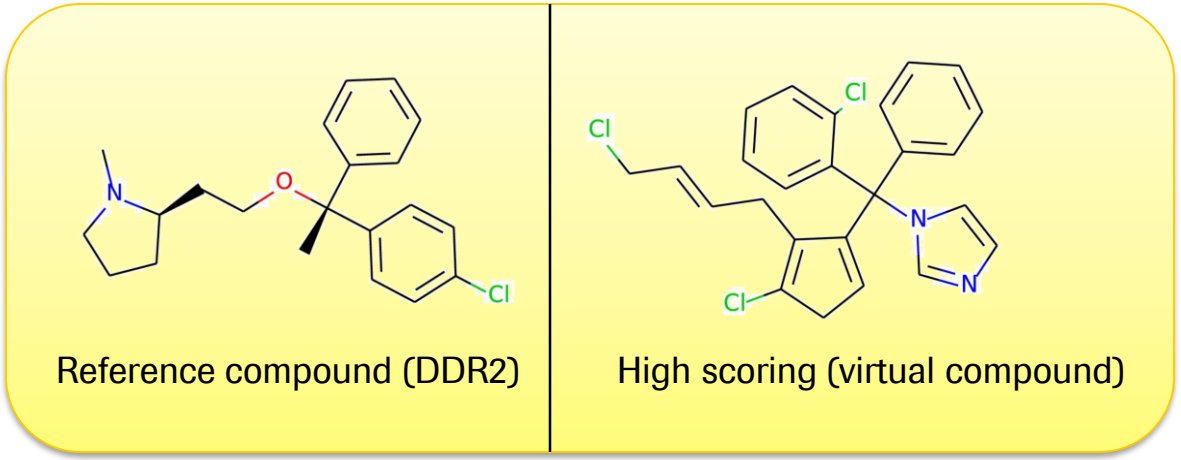


19. CHEMBL1550226
(0.02/1.93)
Total fragments: 82
Unique fragments: 47
R2 fragments: 24



20. CHEMBL488748
(-1.59/1.57)
Total fragments: 55
Unique fragments: 38
R2 fragments: 17

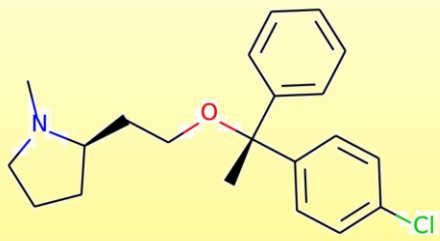
Modification 4: taking fragment frequency into account



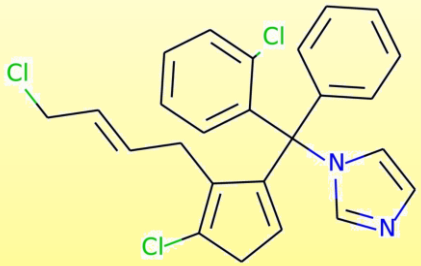
Reference compound (DDR2) | High scoring (virtual compound)

On failure modes in molecule generation and optimization, Günter Klambauer *etal*, *Drug Discovery Today*, **2019**

Modification 4: taking fragment frequency into account

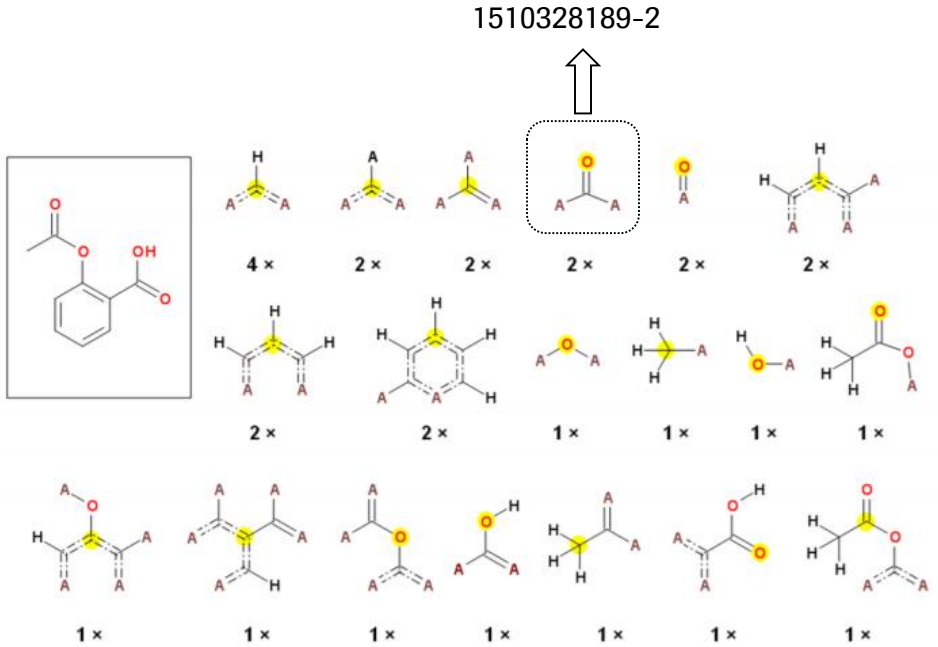


Reference compound (DDR2)



High scoring (virtual compound)

On failure modes in molecule generation and optimization, Günter Klambauer *etal*, *Drug Discovery Today*, **2019**

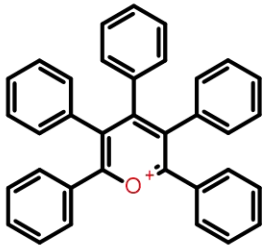


Original SA score: Fragments are defined as morgan fingerprint bitids

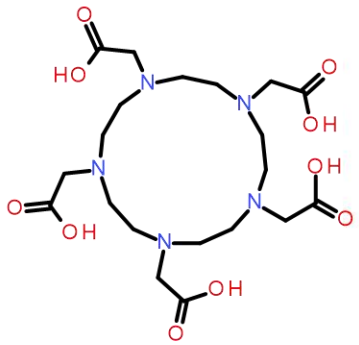
Modified SA score:

Fragment = string(bitid) + "-" + string(frequency within a molecule)

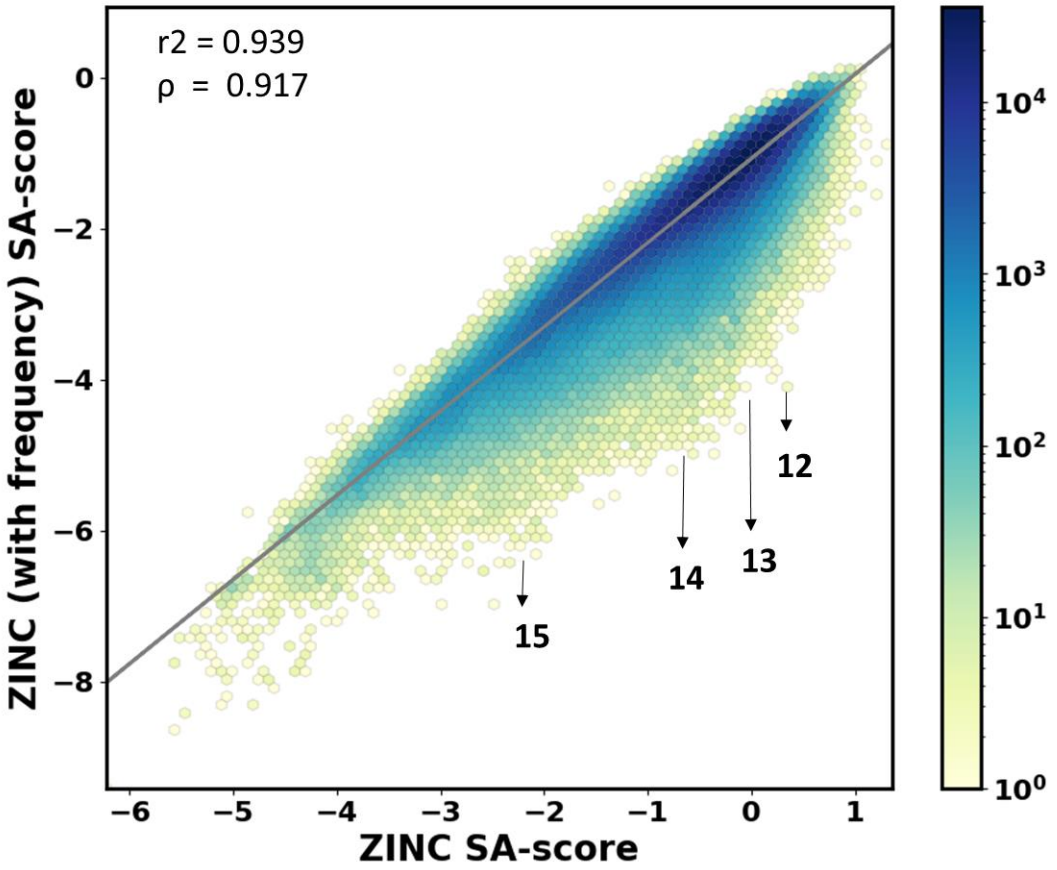
Modification 4: taking fragment frequency into account

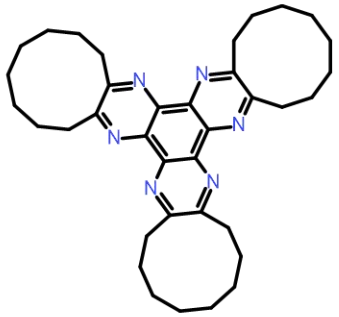


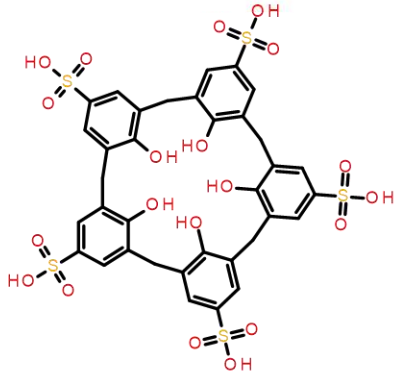
12. CHEMBL2010315
(-0.37/-4.07)



13. CHEMBL103884
(-0.09/-4.10)







Other modifications tested

1) Morgan fingerprint ➡ Daylight type path based fingerprint

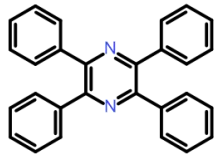
2) Increasing penalty for a fragment not present in fragment DB

3) Increasing penalty for a fragment not present in fragment DB: consider fragment size

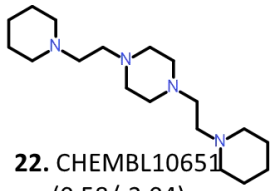
4) Consider worst *n* number of fragments only

5) Inverse document frequency (from NLP approach)

Combining all modifications: new “SA-score”

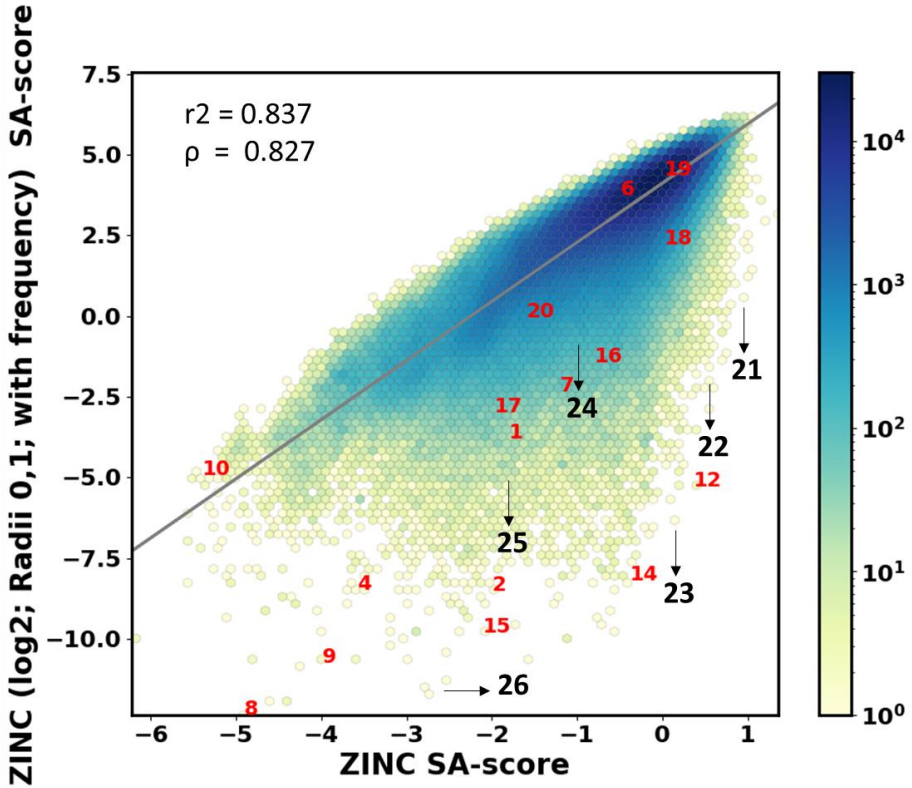


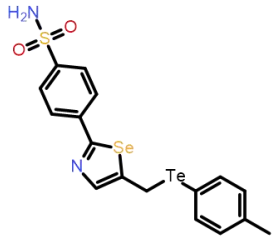
21. CHEMBL11842
(0.91/0.6)



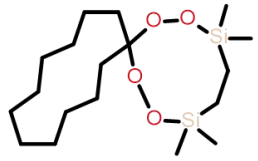
22. CHEMBL10651
(0.58/-2.04)

23. CHEMBL3184154
(0.14/-6.39)

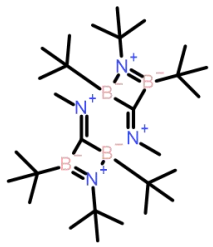




24. CHEMBL4285683
(-1.11/-1.23)



25. CHEMBL3597039
(-2/-5.75)



26. CHEMBL1989831
(-2.74/-11.84)

New SA-score evaluation: Test set 1

https://www.benevolent.com/guacamol

Task	Best of Dataset	SMILES GA
Osimertinib_MPO	100	100
Ranolazine_MPO	100	100
Perindopril_MPO	100	100
Amlodipine_MPO	100	100
Sitagliptin_MPO	100	100

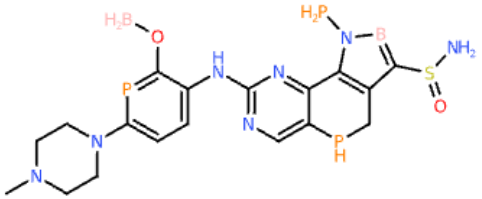
Merged two sets

Best of Dataset = Flag as good

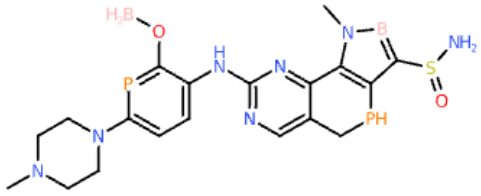
SMILES GA = Flag as Bad

Task	Evaluation set
Osimertinib_MPO	200
Ranolazine_MPO	200
Perindopril_MPO	200
Amlodipine_MPO	200
Sitagliptin_MPO	200

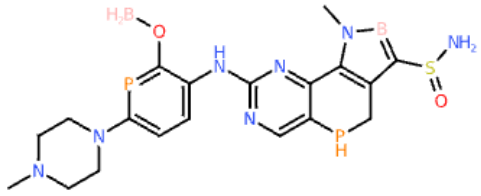
Osimertinib_MPO compounds from SMILES GA



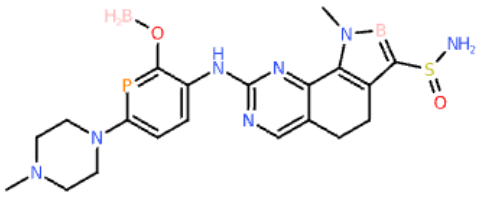
0.88914



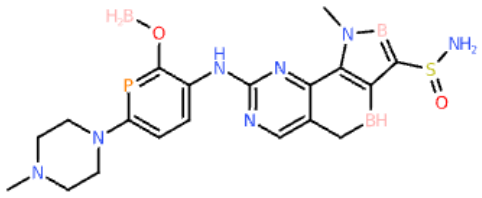
0.88914



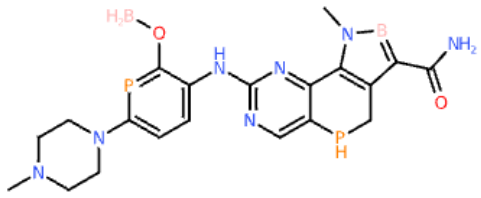
0.88914



0.88914



0.88914



0.88914

New SA-score evaluation: Test set 1

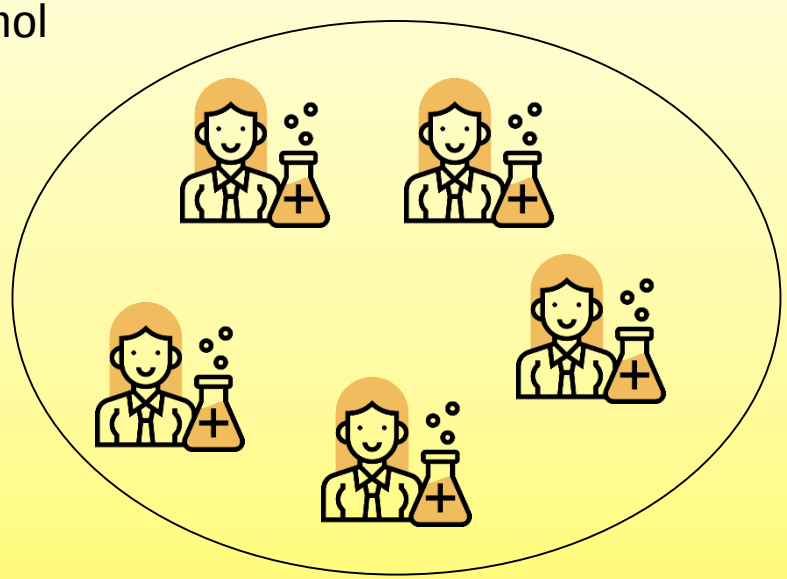
Area under Receiver Operating Curve

task	Original SA-score	ZINC SA-score	QED	ZINC (log2; Radii 0,1; with frequency) SA-score
Osimertinib_MPO	**0.99**	**0.99**	0.86	**0.99**
Ranolazine_MPO	**0.99**	**0.99**	0.62	**0.99**
Perindopril_MPO	**0.97**	**0.97**	0.39	0.96
Amlodipine_MPO	**0.98**	**0.98**	0.95	**0.98**
Sitagliptin_MPO	**0.99**	**0.99**	0.97	**0.99**
average	**0.98**	**0.98**	0.76	**0.98**

New SA-score evaluation: Test set 2

https://www.benevolent.com/guacamol

Task	Graph GA	LSTM
Osimertinib_MPO	100	100
Fexofenadine_MPO	100	100
Ranolazine_MPO	100	100
Perindopril_MPO	100	100
Amlodipine_MPO	100	100
Sitagliptin_MPO	100	100
Zaleplon_MPO	100	100
Osimertinib_MPO	100	100



We ask medicinal chemists

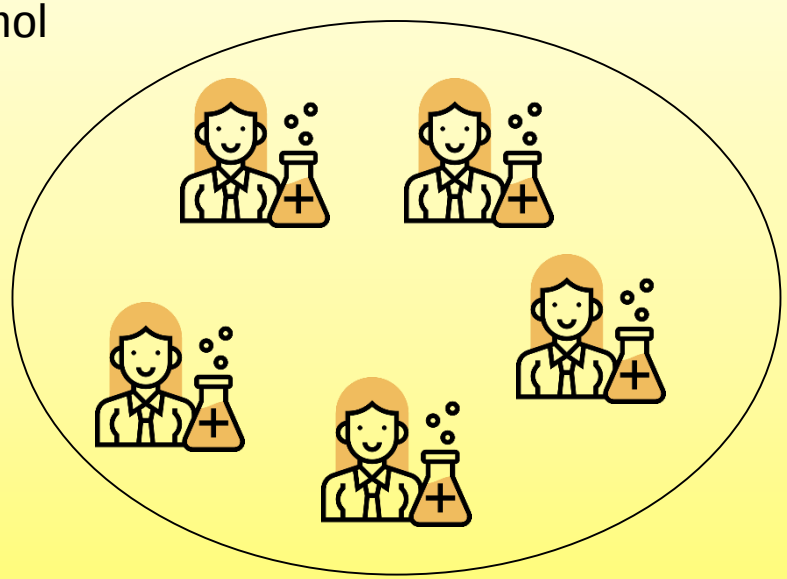
to flag bad compounds:

- Which compounds are undesirable from their structure
- Will never be interesting for drug discovery projects
- Ease of synthesis is not an important parameter
- Only spend few seconds/compound

New SA-score evaluation: Test set 2

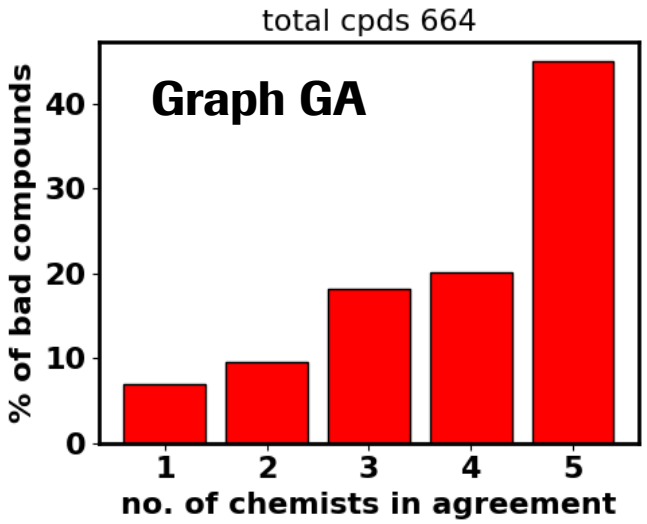
https://www.benevolent.com/guacamol

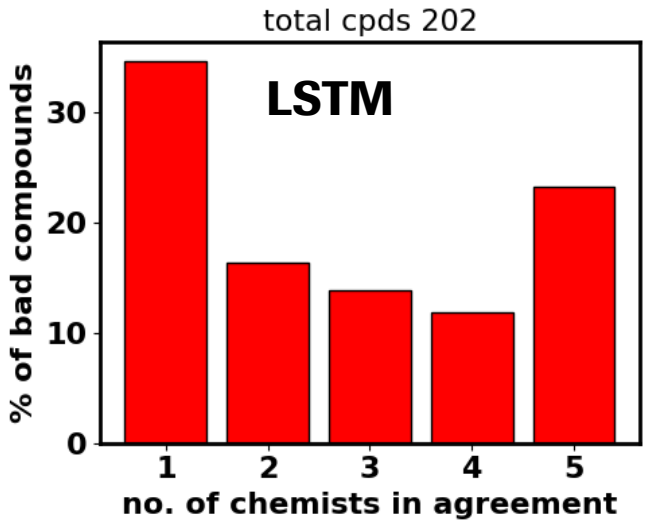
Task	Graph GA	LSTM
Osimertinib_MPO	100	100
Fexofenadine_MPO	100	100
Ranolazine_MPO	100	100
Perindopril_MPO	100	100
Amlodipine_MPO	100	100
Sitagliptin_MPO	100	100
Zaleplon_MPO	100	100
Osimertinib_MPO	100	100



We ask medicinal chemists to flag bad compounds:

- Which compounds are undesirable from their structure
- Will never be interesting for drug discovery projects
- Ease of synthesis is not an important parameter
- Only spend few seconds/compound

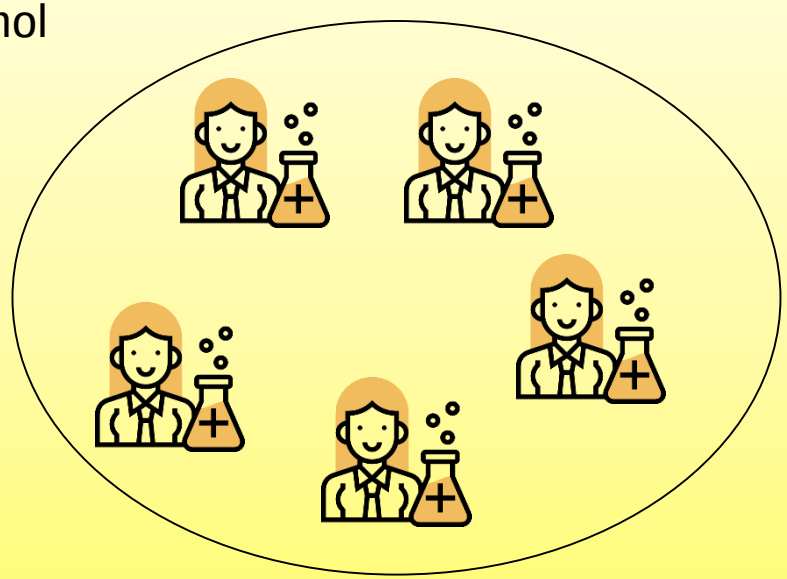




New SA-score evaluation: Test set 2

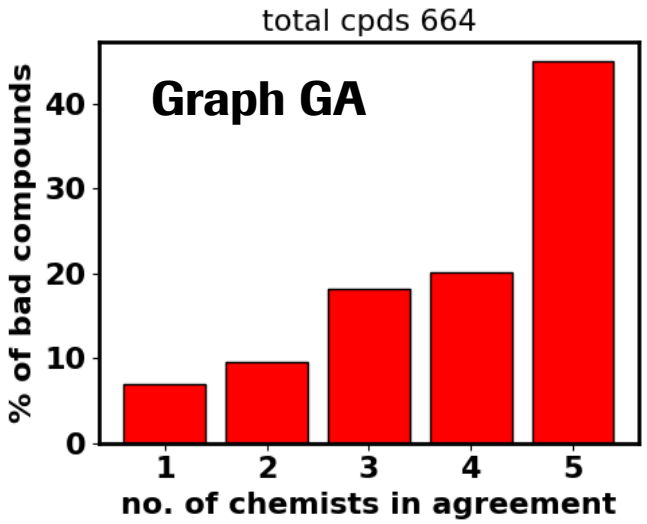
https://www.benevolent.com/guacamol

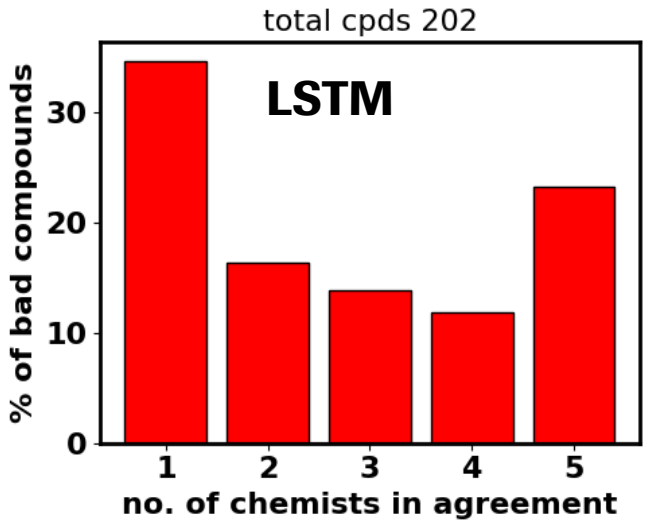
Task	Graph GA	LSTM
Osimertinib_MPO	100	100
Fexofenadine_MPO	100	100
Ranolazine_MPO	100	100
Perindopril_MPO	100	100
Amlodipine_MPO	100	100
Sitagliptin_MPO	100	100
Zaleplon_MPO	100	100
Osimertinib_MPO	100	100



We ask medicinal chemists to flag bad compounds:

- Which compounds are undesirable from their structure
- Will never be interesting for drug discovery projects
- Ease of synthesis is not an important parameter
- Only spend few seconds/compound





We used majority vote to set a flag

Task	Graph GA	LSTM
Osimertinib_MPO	84	1
Fexofenadine_MPO	95	8
Ranolazine_MPO	98	2
Perindopril_MPO	22	5
Amlodipine_MPO	65	38
Sitagliptin_MPO	100	9
Zaleplon_MPO	61	0
Osimertinib_MPO	92	13

New SA-score evaluation: Test set 2

Area under Receiver Operating Curve, Graph GA

task	Original SA-score	ZINC SA-score	QED	ZINC (log2; Radii 0,1; with frequency) SA-score
Osimertinib_MPO	0.61	0.65	**0.75**	0.74
Fexofenadine_MPO	**0.80**	**0.80**	0.79	**0.80**
Ranolazine_MPO	0.70	0.73	0.58	**0.74**
Perindopril_MPO	0.92	0.93	0.90	**0.95**
Amlodipine_MPO	0.43	0.46	**0.65**	0.57
Sitagliptin_MPO	NA	NA	NA	NA
Zaleplon_MPO	0.76	0.74	0.65	**0.79**
Valsartan_SMARTS	0.53	0.53	**0.83**	0.67
average	**0.68**	**0.69**	**0.73**	**0.75**

New SA-score evaluation: Test set 2

Area under Receiver Operating Curve, LSTM

task	Original SA-score	ZINC SA-score	QED	ZINC (log2; Radii 0,1; with frequency) SA-score
Osimertinib_MPO	0.65	0.69	0.61	**0.88**
Fexofenadine_MPO	0.45	0.43	**0.93**	0.51
Ranolazine_MPO	0.68	0.70	0.77	**0.79**
Perindopril_MPO	0.97	**0.98**	0.73	**0.98**
Amlodipine_MPO	0.85	0.87	0.65	**0.90**
Sitagliptin_MPO	0.48	0.56	**0.87**	0.60
Zaleplon_MPO	NA	NA	NA	NA
Valsartan_SMARTS	0.74	0.75	0.54	**0.81**
average	0.69	0.71	0.73	**0.78**

Take away

➢ **Original SA score ⟹ ZINC SA score with modifications**

- Improved differentiation between good and bad compounds
- Not all bad cases are captured
- Use it along with QED or some other metrics

➢ **Modifications**

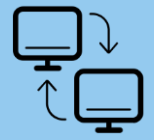
- Changed compound database for FCD
- Dropped radius 2 fragments
- Introduced log2 function
- Modified definition of fragment (fragment + frequency in a molecule)

➢ **More ideas for modifications?**

- Please reach out to us

➢ **Manuscript in progress**

➢ **Code and Data sharing**



⟹ Benchmark datasets

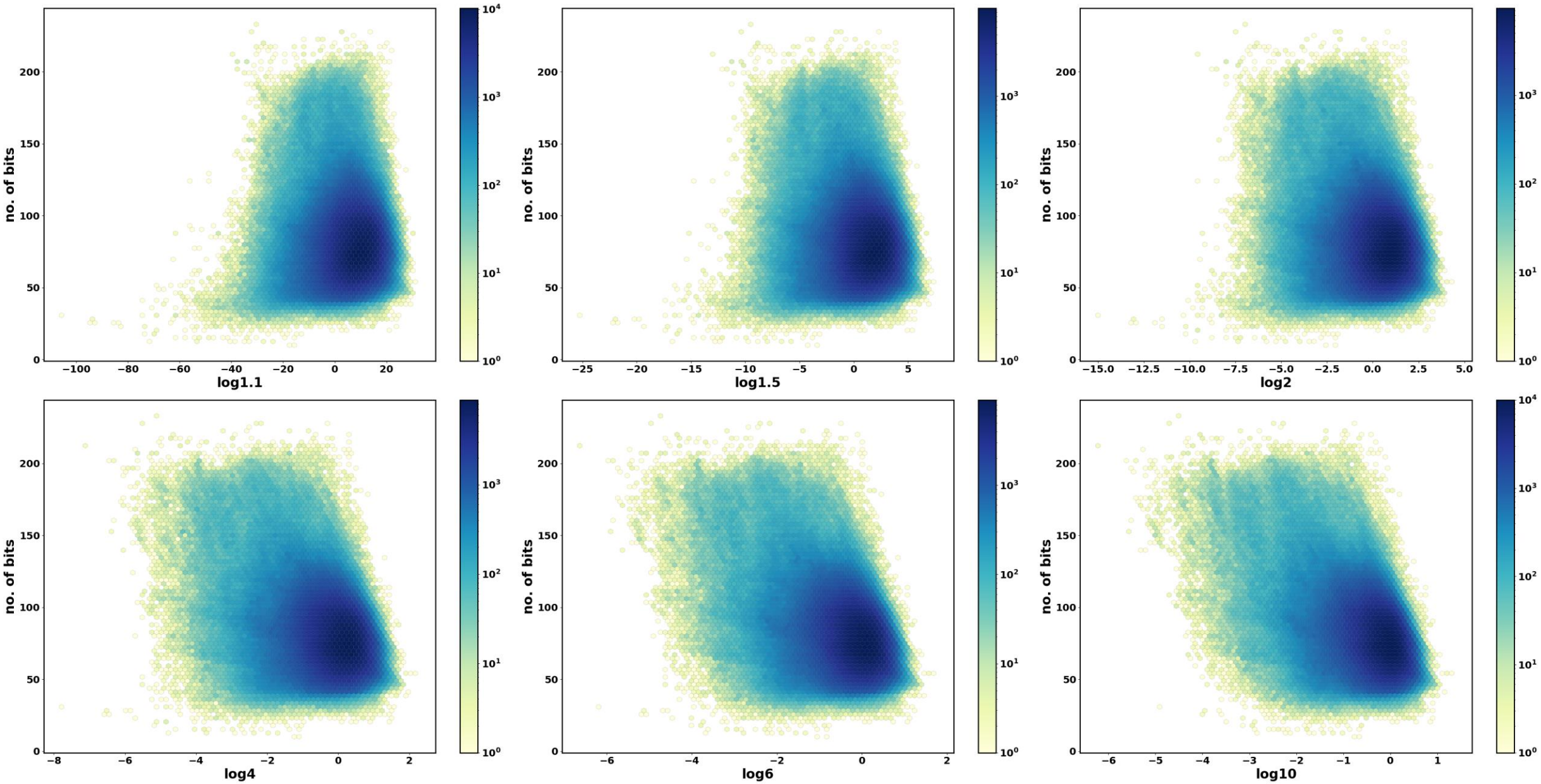
⟹ Code to train SA score with your own database of choice

⟹ Code to calculate new SA score

Acknowledgement

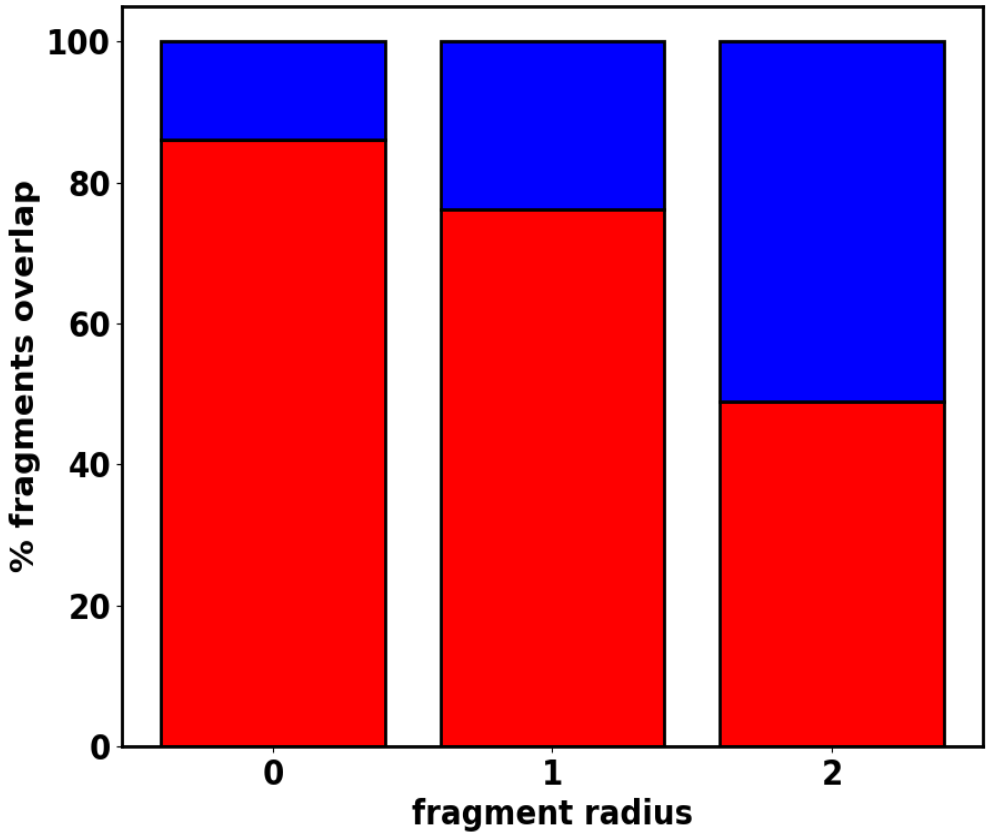
- Jerome Hert (CADD, Roche)
- Christian Kramer (CADD, Roche)
- Sereina Riniker (ETH)
- Greg Landrum (ETH)

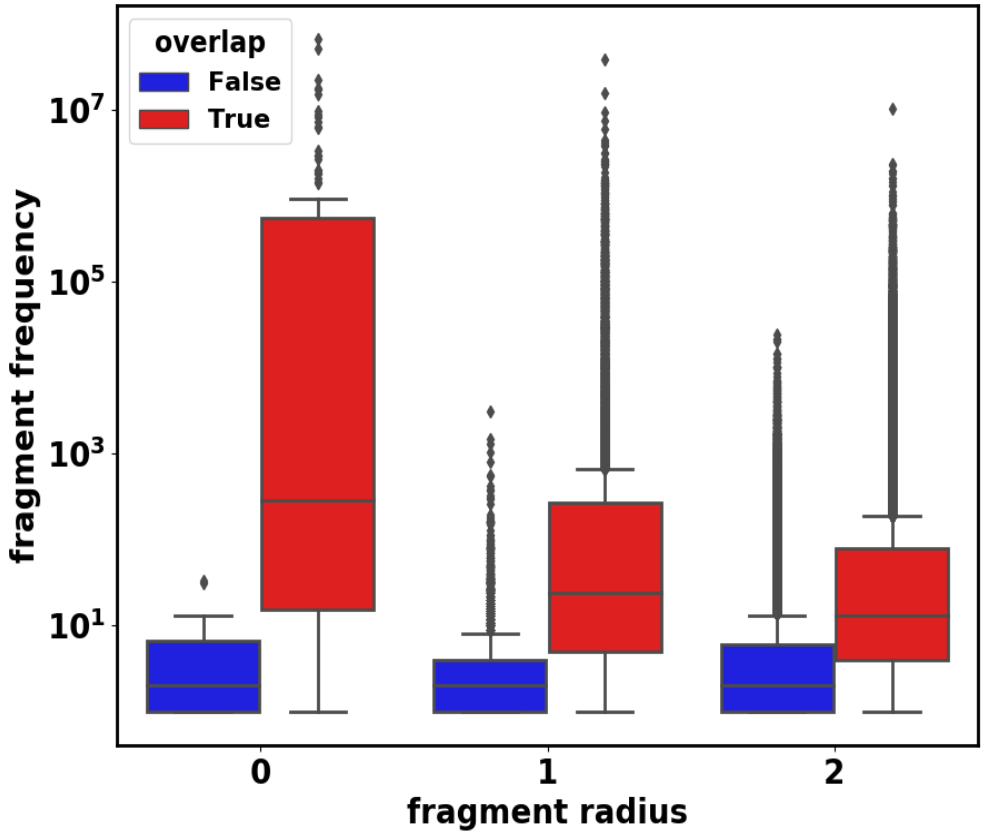
Log functions: dependency of score on number of fragments



Modification 3:
remove radius 2 fragments from consideration

Overlap of ZINC InStock fragments with fragments from Original Publication





Doing now what patients need next