AI-driven drug design (AIDD):
Coupling high-throughput pharmacokinetic simulation (HTPK) to multi-objective molecular evolution of triazolopyrimidine antimalarial leads

Robert D. Clark,
Indiana University, Bloomington

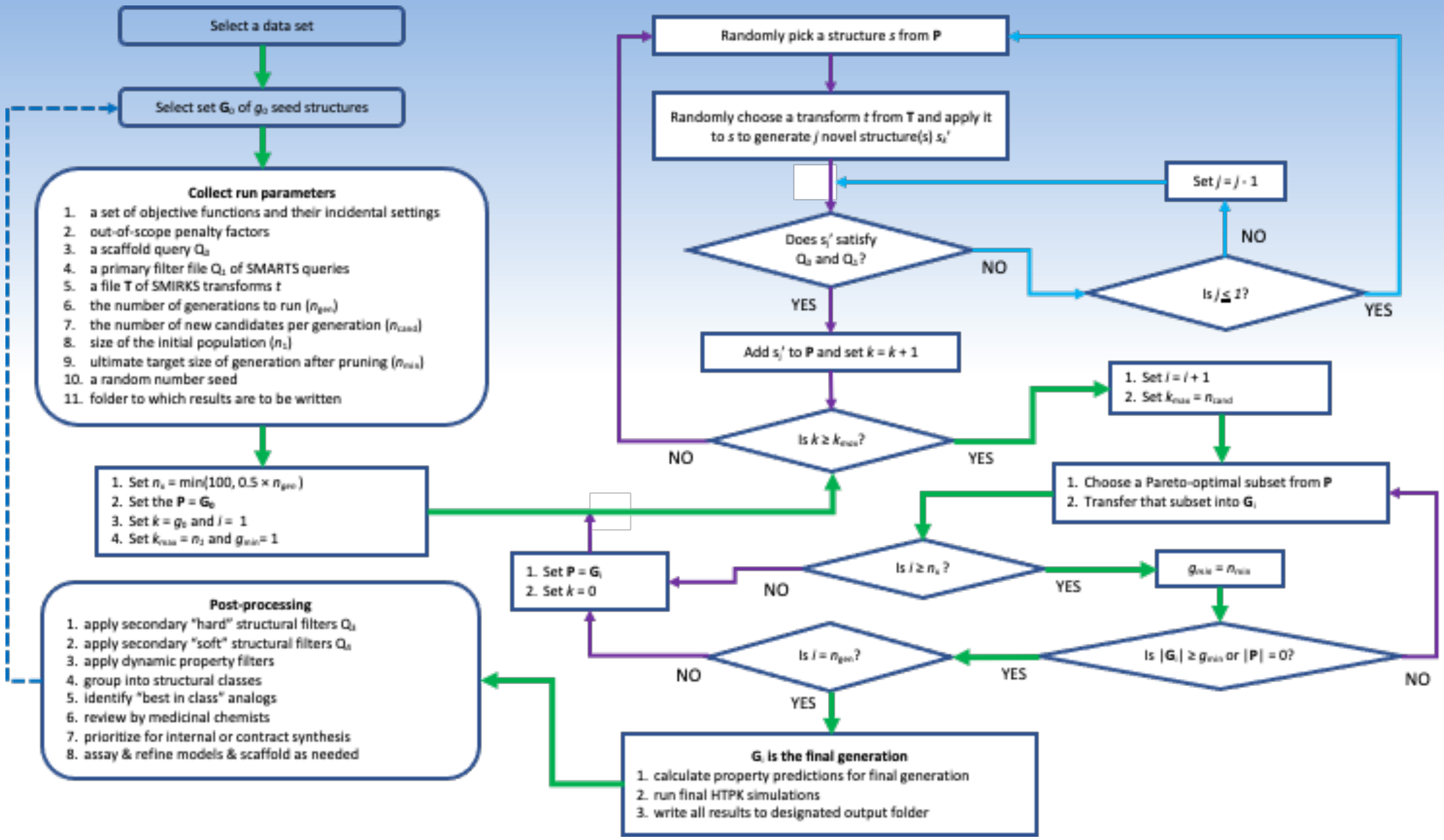
Michael S. Lawless, David W. Miller and Marvin Waldman
Simulations Plus, Inc.

Chicago ACS 2022

SimulationsPlus

INDIANA UNIVERSITY

What is AIDD and how does it work?



AI-driven Drug Discovery:

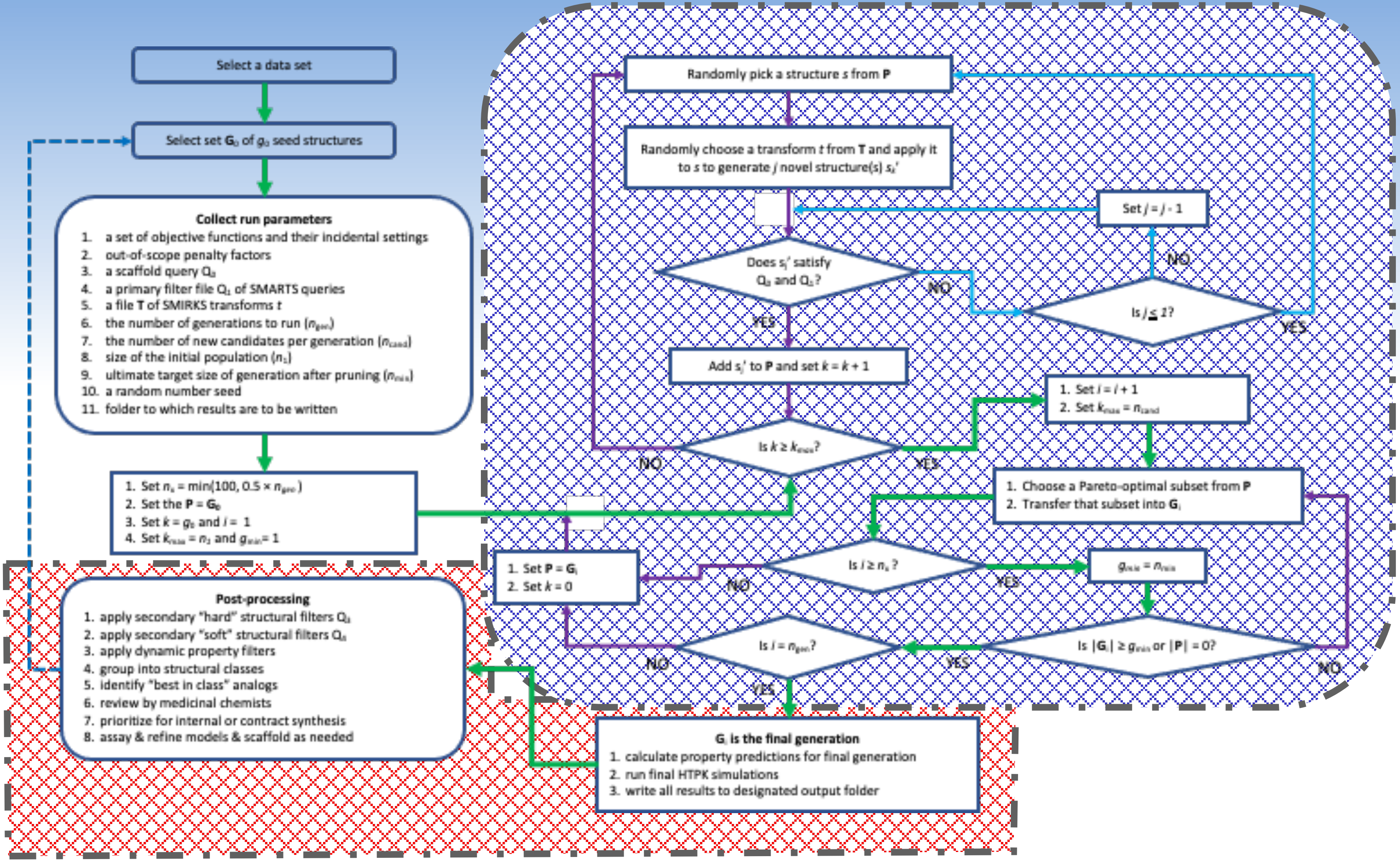
- Is complicated (see left).
- Uses a set of **SMIRKS transforms** to generate new molecules from ones that are already present in a population.
- Relies on an **evolutionary algorithm** to select for high-quality & diverse molecules.
- Periodically prunes the population based on **Pareto rank**; survivors make up the next generation.
- **Adjusts the chances** of a survivor being selected in the next round of molecule generation based on its **fitness** & how many children it produced.
- Can use a **wide range of objective functions** (including ones external to the program) and **filters** to steer selection.
- Is designed to **provide ideas for med chemists to work from** as well as opportunities for them to reshape output molecules on the fly.

AIDD does not:

- Use **deep neural networks** to generate or evaluate candidate molecules.

INDIANA UNIVERSITY

What is AIDD* and how does it work?



AI-driven Drug Discovery:

- Is complicated (see left).
- Uses a set of **SMIRKS transforms** to generate new molecules from ones that are already present in a population.
- Relies on an **evolutionary algorithm** to select for high-quality & diverse molecules.
- Periodically prunes the population based on **Pareto rank**; survivors make up the next generation.
- **Adjusts the chances** of a survivor being selected in the next round of molecule generation based on its **fitness** & how many children it produced.
- Can use a **wide range of objective functions** (including ones external to the program) and **filters** to steer selection.
- Is designed to **provide ideas for med chemists to work from** as well as opportunities for them to reshape output molecules on the fly.

AIDD does not:

- Use **deep neural networks** to generate or evaluate candidate molecules.

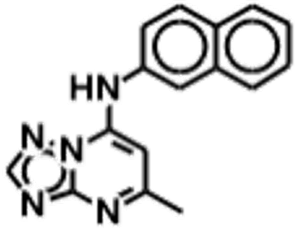
*AIDD is a module of ADMET Predictor®, which is distributed by Simulations Plus, Inc.

INDIANA UNIVERSITY

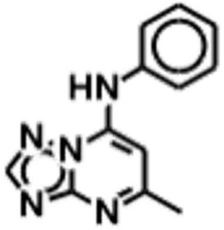
Example: the antimalarial triazolopyrimidine (TzP) data set

- MA Phillips et al. *J Med Chem* **2008**, *51*, 3649-3653.
- R Gujjar et al. *J. Med. Chem* **2009**, *52*, 1864-1872.
- R Gujjar et al. *J Med Chem* **2011**, *54*, 3935-3949.
- JM Coteron et al. *J Med Chem* **2011**, *54*, 5540-5561.
- A Marwaha et al. *J Med Chem* **2012**, *55*, 7425-7436.
- X Deng et al. *J Med Chem* **2014**, *57*, 5381-5394.
- MA Phillips et al. *Science Translat. Med.* **2015**, *7*, 296ra111-296ra111.
- S Kokkonda et al. *J Med Chem* **2016**, *59*, 5416-5431.

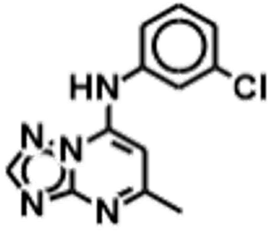
Used in building the activity model



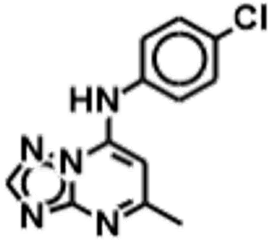
DSM1



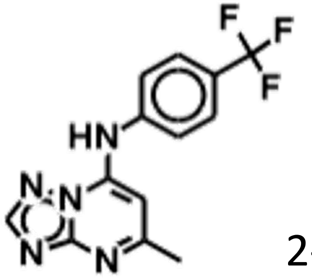
DSM12



DSM75

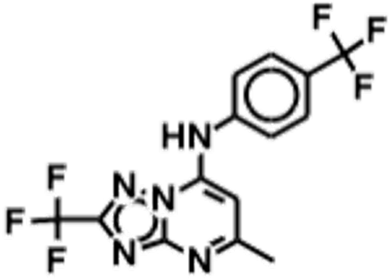


DSM89

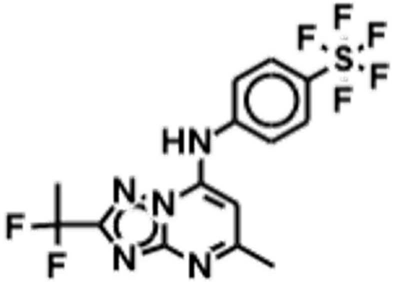


DSM74

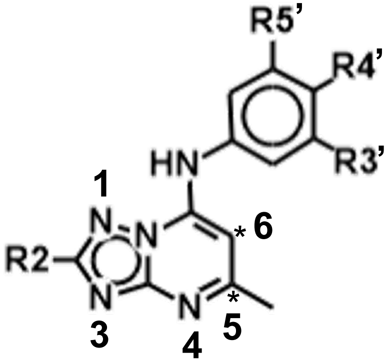
+ 37 other 2-unsubstituted analogs



DSM195



DSM265

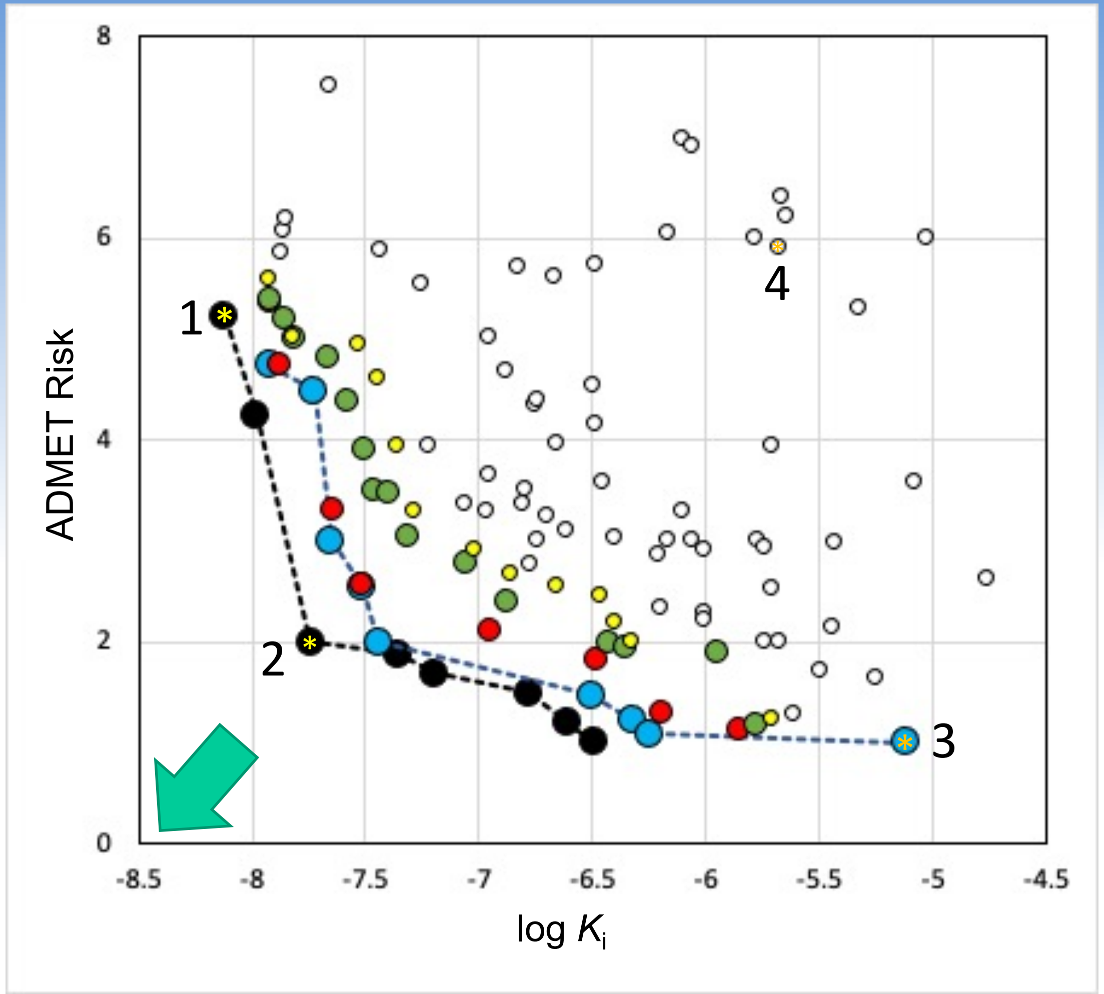


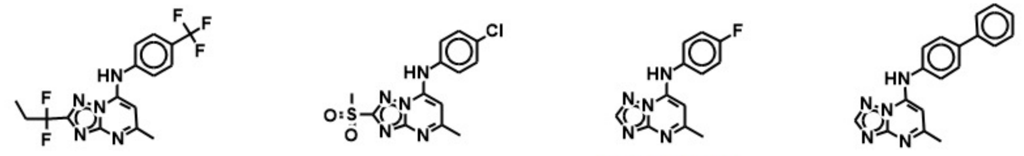
Scaffold

INDIANA UNIVERSITY

Pareto ranking TzPs (2 objectives)

- A member x_i of a set is *dominated* by another member x_j of that set unless x_i is superior to x_j with respect to some Pareto objective attribute.
- A (sub)set is *Pareto optimal* when no member is dominated by any other member.
- The *Pareto rank r* of x_i is 1 plus the number of Pareto optimal subsets that must be removed from a set before x_i is Pareto optimal in the residual set.[a]
- The plot at right shows the first five Pareto ranks for the set of literature TzPs that are "hit" by the consensus "active" scaffold.
- The two attributes considered here were:
  - experimental log K_i with respect to malarial dihydroorotate dehydrogenase (*Pf*DHODH)
  - an *ADMET Risk* score[b] based on 22 fuzzy-logic rules calibrated against a reference set of oral drugs, 10% of which "break" > 7



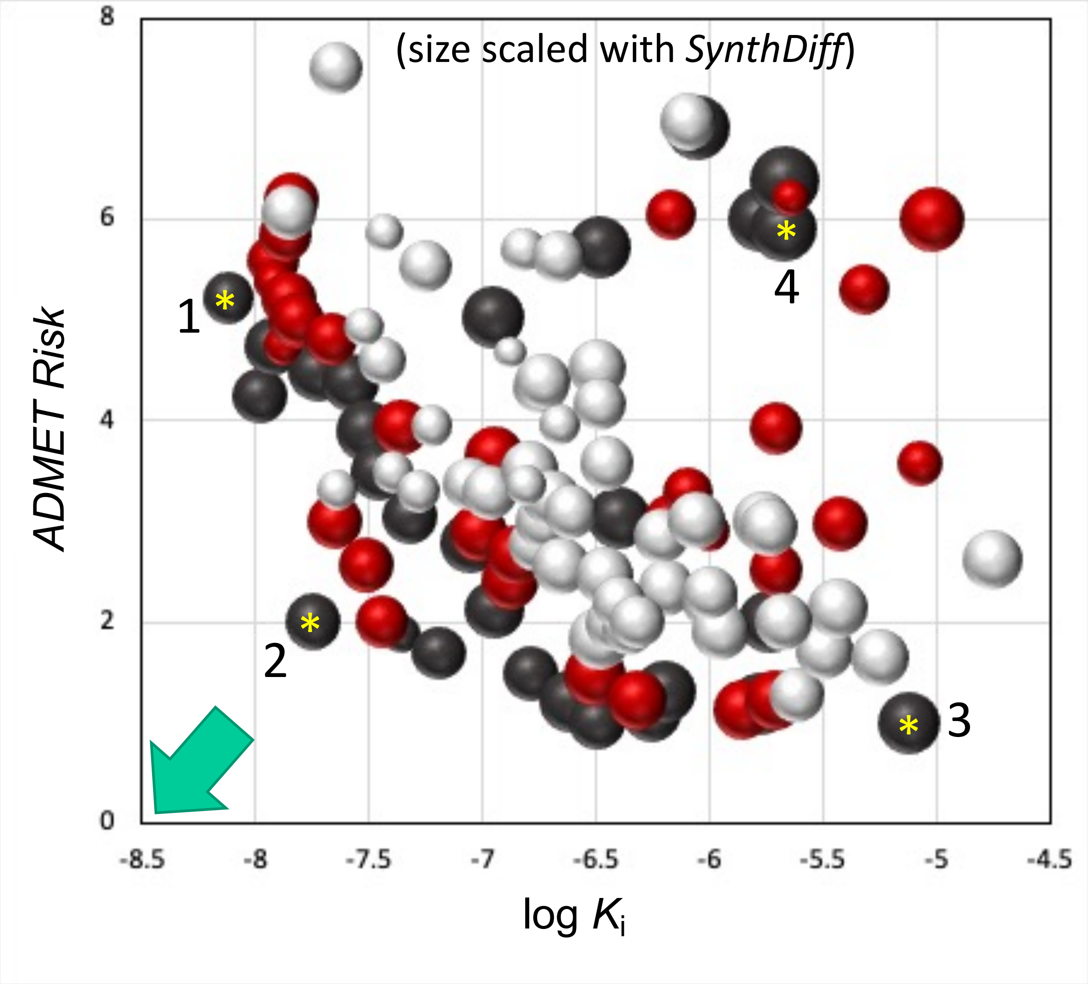


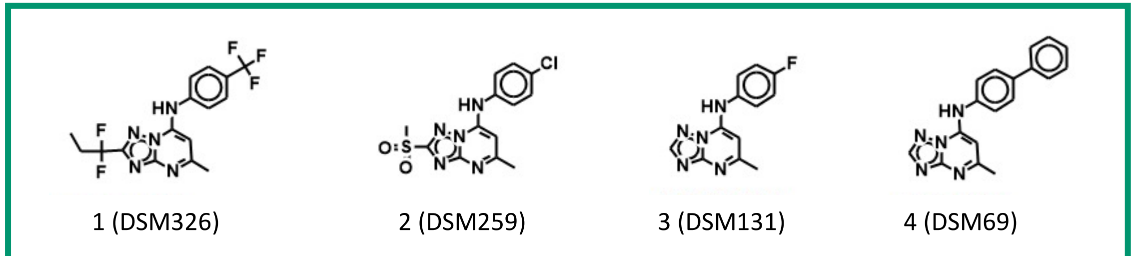
[a]See, for example: Abdou *et al.*, 12th Euro Conf Evolutionary Computation in Combinatorial Optimization (EvoCOP) **2012**, Spain. 194–205 (hal-00940119)
[b]M Lawless et al., Handb Exp Pharmacol **2016**, 232, 139-168 (doi: 10.1007/164_2015_23)

Pareto ranking TzPs (3 objectives)

- A member x_i of a set is *dominated* by another member x_j of that set unless x_i is superior to x_j with respect to some Pareto objective attribute.
- A (sub)set is *Pareto optimal* when no member is dominated by any other member.
- The *Pareto rank r* of x_i is 1 plus the number of Pareto optimal subsets that must be removed from a set before x_i is Pareto optimal in the residual set.[a]
- The plot at right shows the first **two** Pareto ranks for the set of literature TzPs that are "hit" by the consensus "active" scaffold.
- The **three** attributes considered here were:
  - experimental log K_i with respect to malarial dihydroorotate dehydrogenase (*Pf*DHODH)
  - ADMET Risk
  - estimated synthetic difficulty (*SynthDiff*)[a]

[a] à la Ertl & Schuffenhauer, J Cheminformatics **2009**, *1*, 8 (doi: 10.1186/1758-2946-1-8)





Models & settings used for illustrative TzP AIDD runs

- Primary filters to check scaffold and weed out problematic ("undruglike") substructures
- *log* K_i^{gen} model from Clark et al. (*JCAMD* **2020**, *34*, 1117-1132; doi: 10.1007/s10822-020-00333-x)
    - ANNE model based on 89 diverse DHODH inhibitors, 42 of which were 2-unsubstituted TzPs
    - SEP $\pm$0.5 log units; capped at -7.4 minimum
- Bioavailability from ADMET Predictor's HTPK module: *%Fb*
    - estimated based on 1 mg oral dose for 70 kg human; capped at 90% max
- Synthetic difficulty score augmented with "toxicophoric" penalties: *SynthDiff+*
    - Capped at a minimum of 2
- *AIDD Risk*: a reweighted version of *ADMET Risk* with broadened thresholds

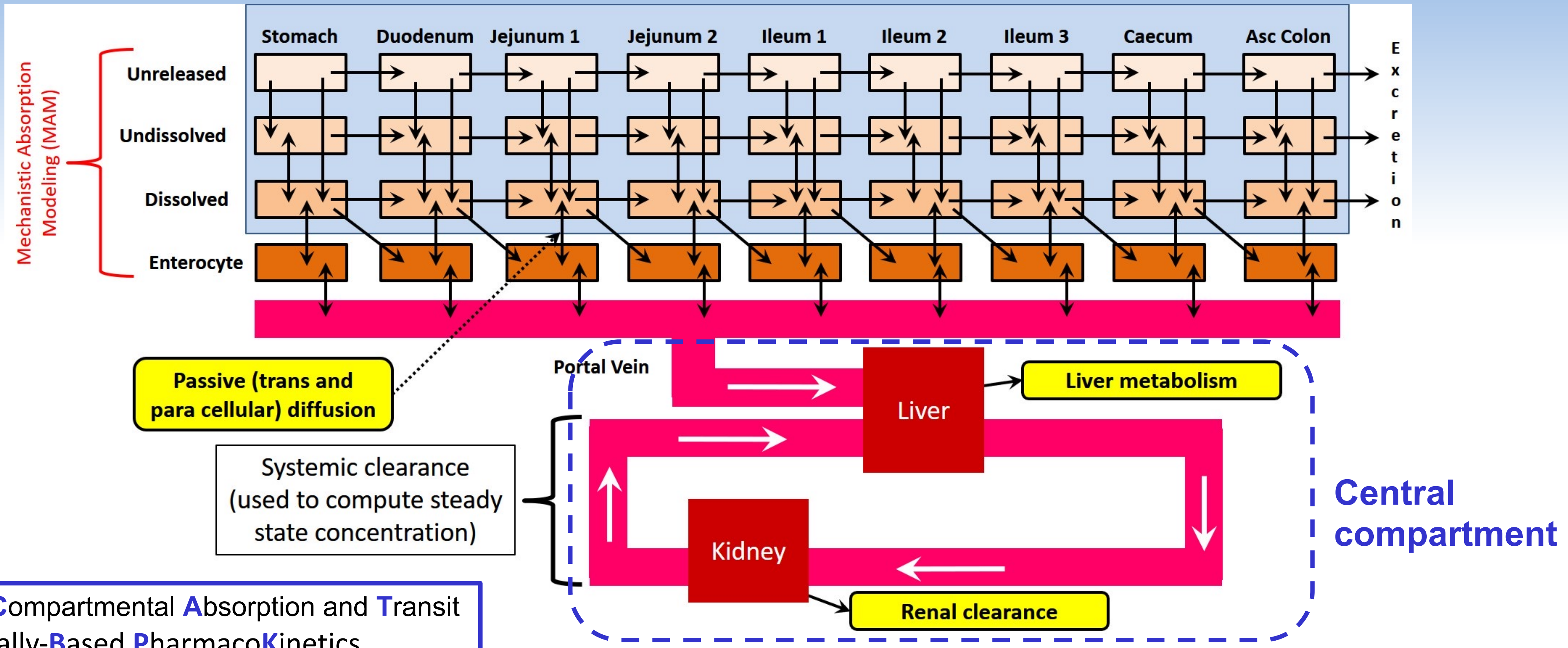
Objective functions used for Pareto ranking within the evolutionary cycle

- Create an initial population of 500 molecules; create 500 new ones per generation; and keep at least 500 per generation after the 100th (or half-way through the run)
- Run for 500 or 50 generations
- *%Fb*, ADMET Risk, *log* K_i and "simple" *SynthDiff* were used for post-processing
    - minimum of 70% and maxima of 6, -7.2, and 5, respectively, yielded ~300 products per run
    - "post" out-of-scope penalties are less harsh than those that were used during molecular evolution

INDIANA UNIVERSITY

Mechanistic High-Throughput Pharmacokinetic Simulation (HTPK)

GastroPlus® ACAT™ Model* + Compartmental (Minimal PBPK) Model*

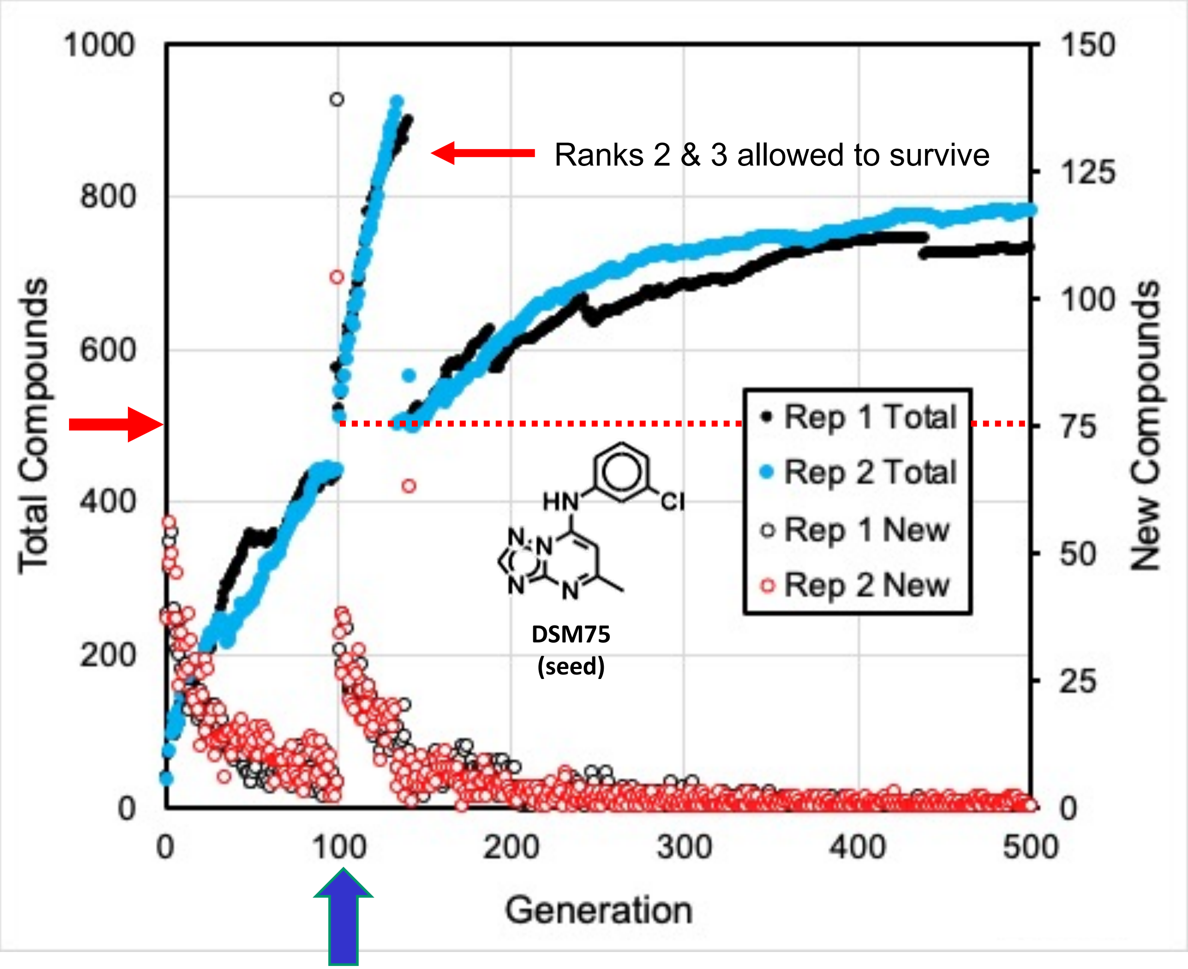


*Advanced Compartmental Absorption and Transit
**Physiologically-Based PharmacoKinetics

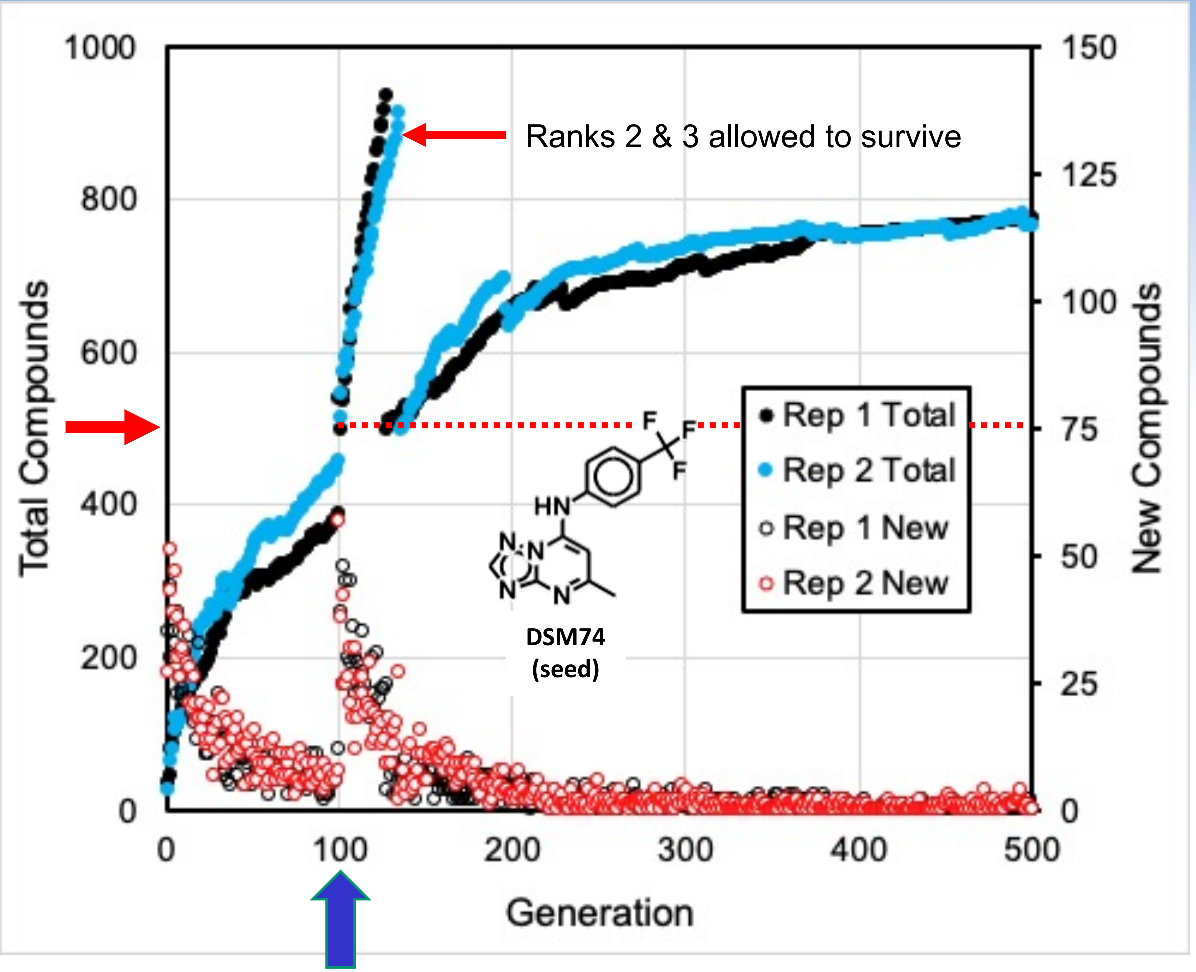
P. Daga *et al.* Physiologically Based Pharmacokinetic Modeling in Lead Optimization. 1 & 2, *Mol Pharmaceutics* **2018**, *15*, 821-830 & 831-839.

INDIANA UNIVERSITY

Population growth across generations



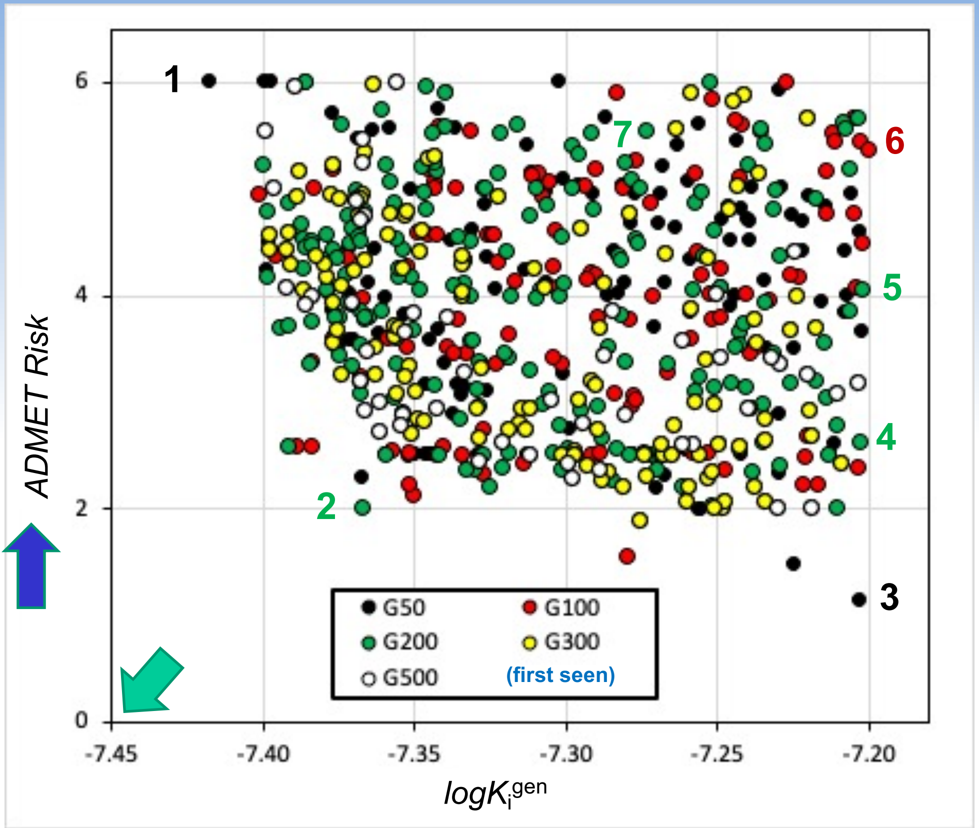
Minimum population size (500) takes effect

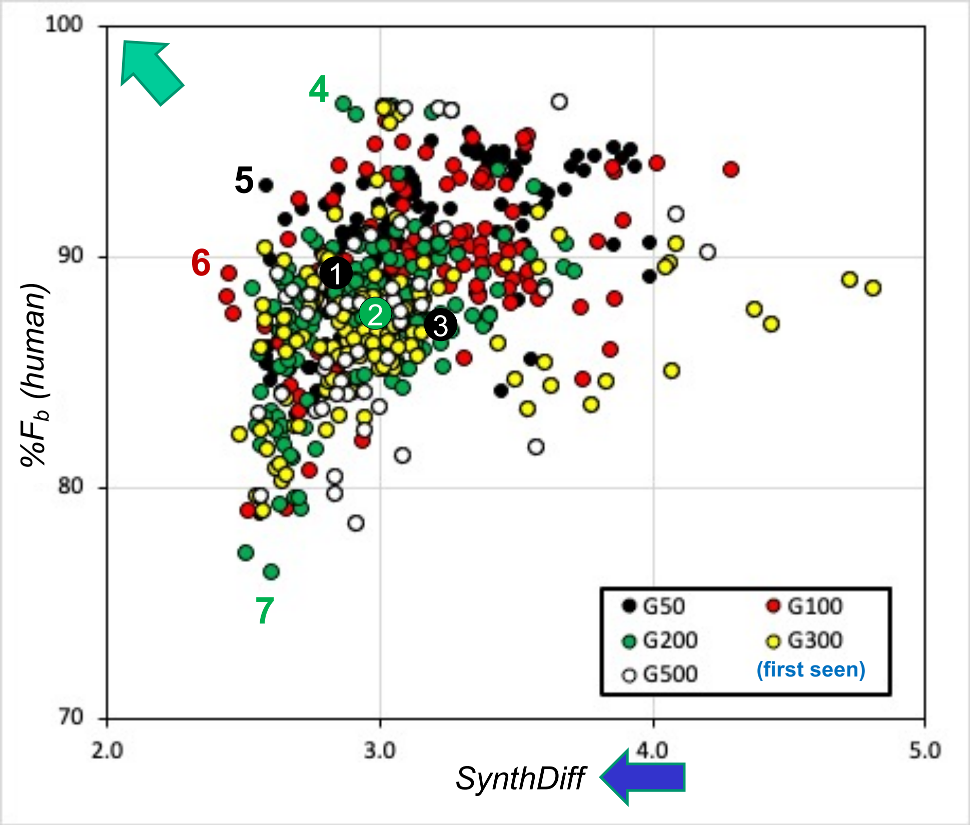


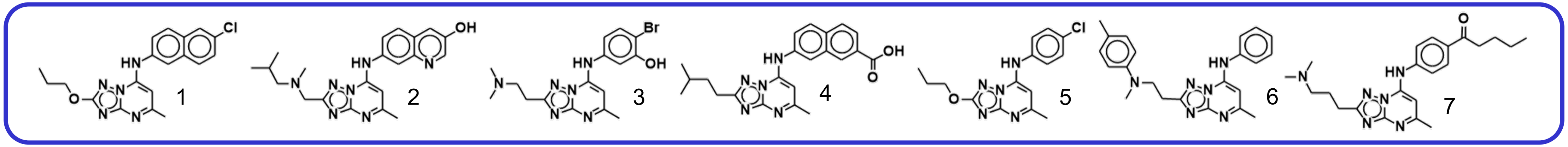
Minimum population size (500) takes effect

INDIANA UNIVERSITY

Pairwise progress on Pareto objectives (by origin)*



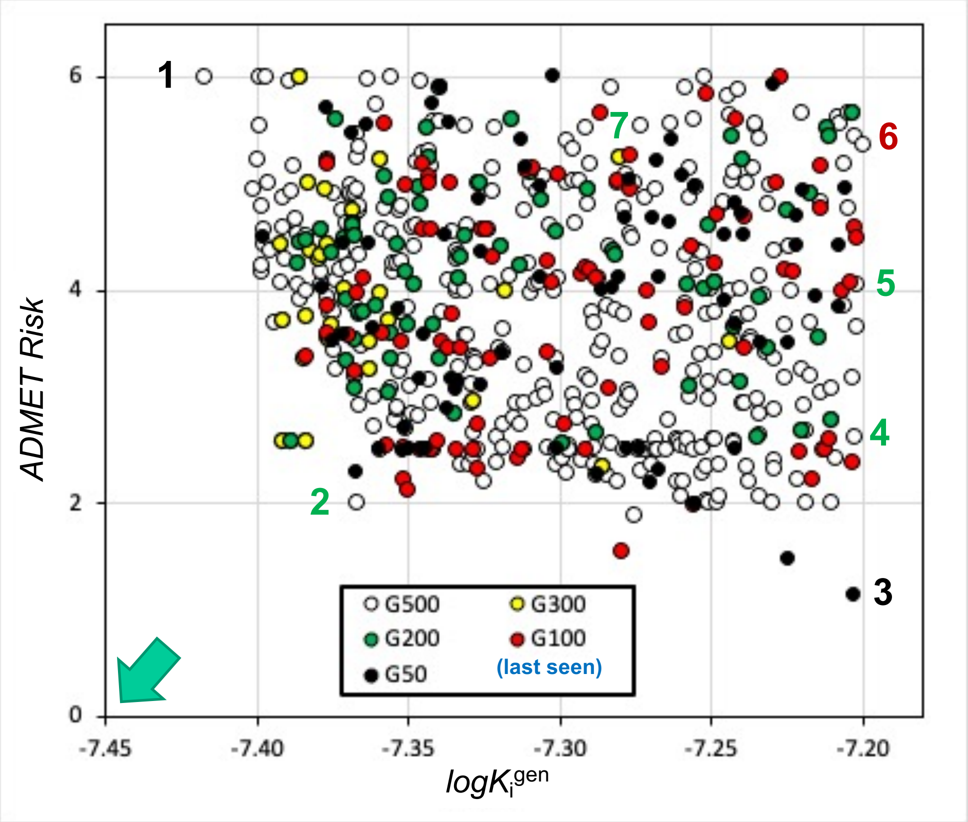


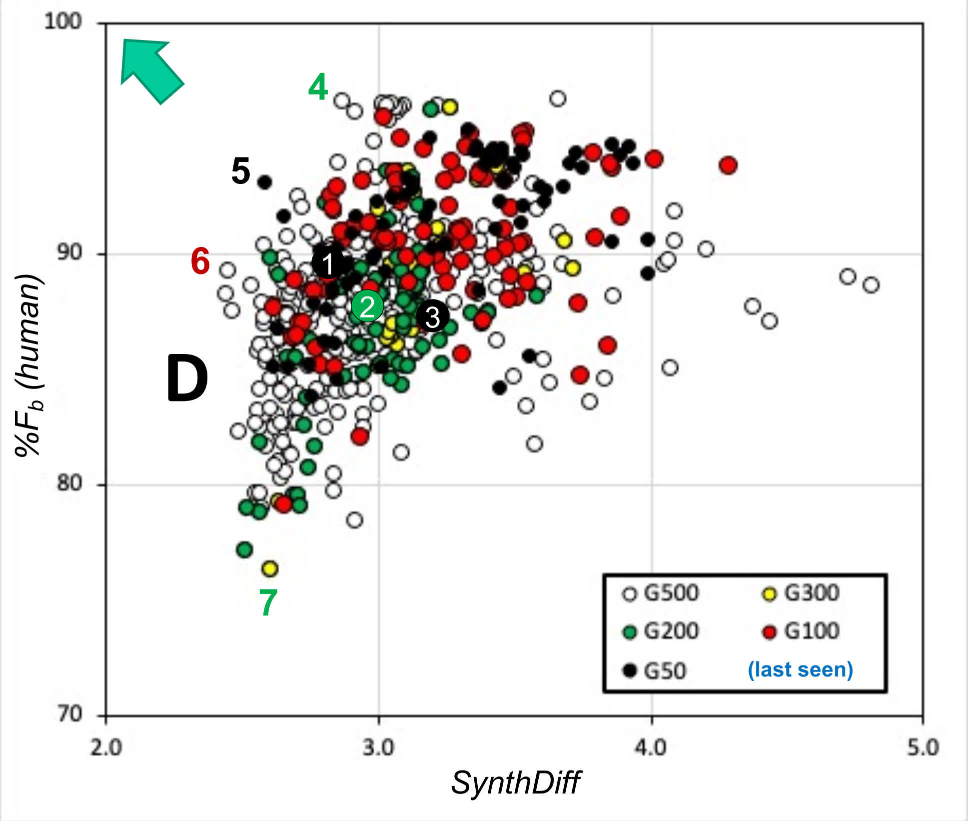


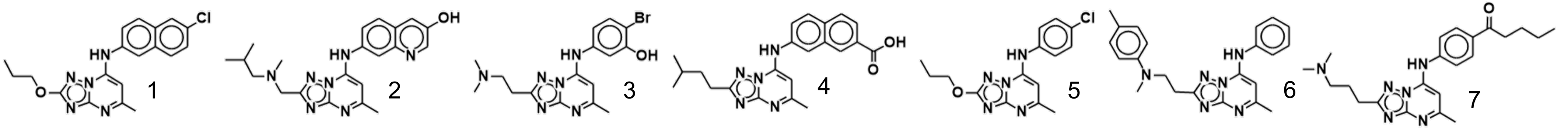
*Seed structure DSM75 (3’-Cl)

INDIANA UNIVERSITY

Pairwise progress on Pareto objectives (by extinction)



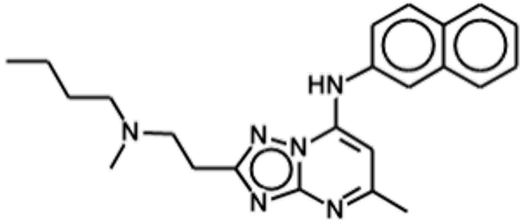




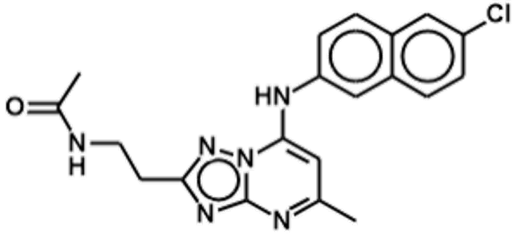
*Seed structure DSM75 (3'-Cl)

INDIANA UNIVERSITY

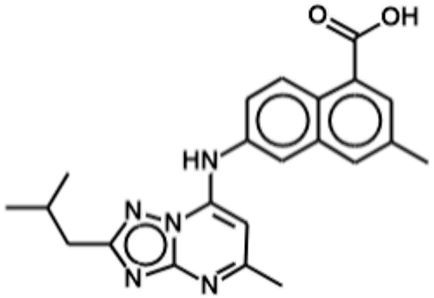
Examples from different product classes



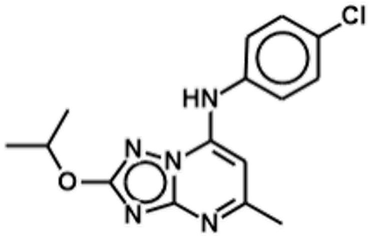
9 (A1)



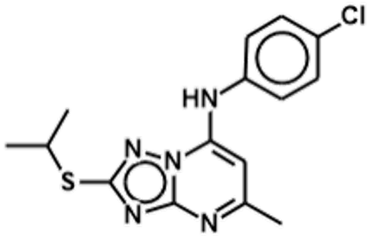
10 (A2)



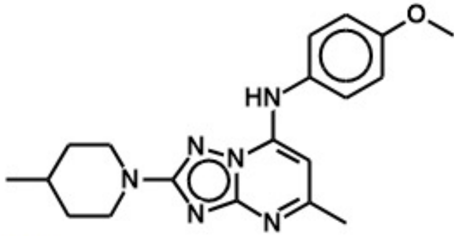
11 (A2)



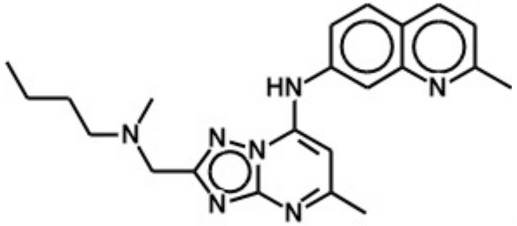
B3 (12)



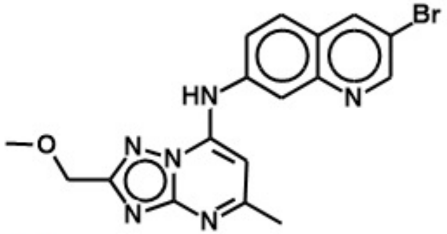
13 (B4)



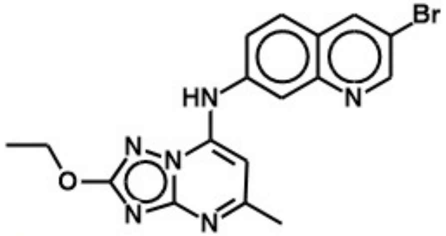
14 (B6)



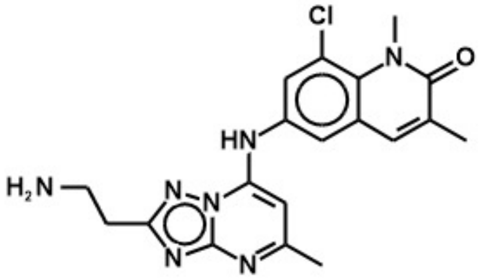
15 (D1)



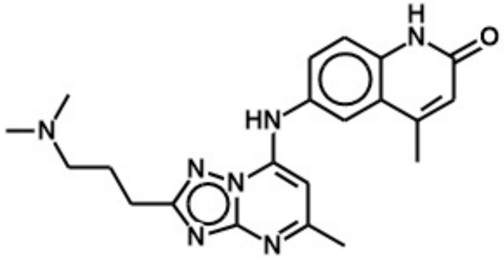
16 (D2)



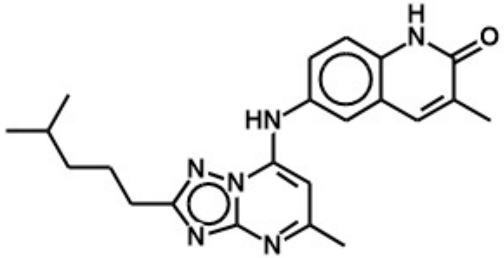
17 (D3)



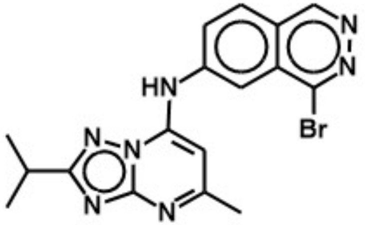
18 (E1)



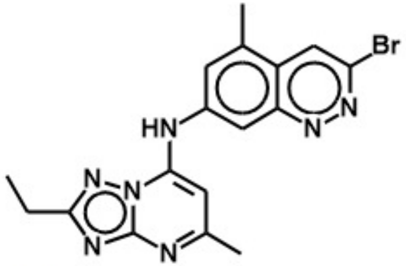
19 (E1)



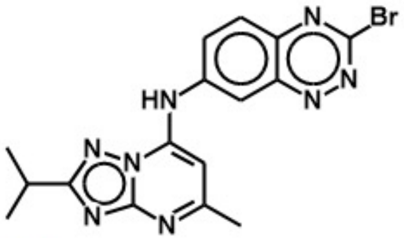
20 (E2)



21 (U2)



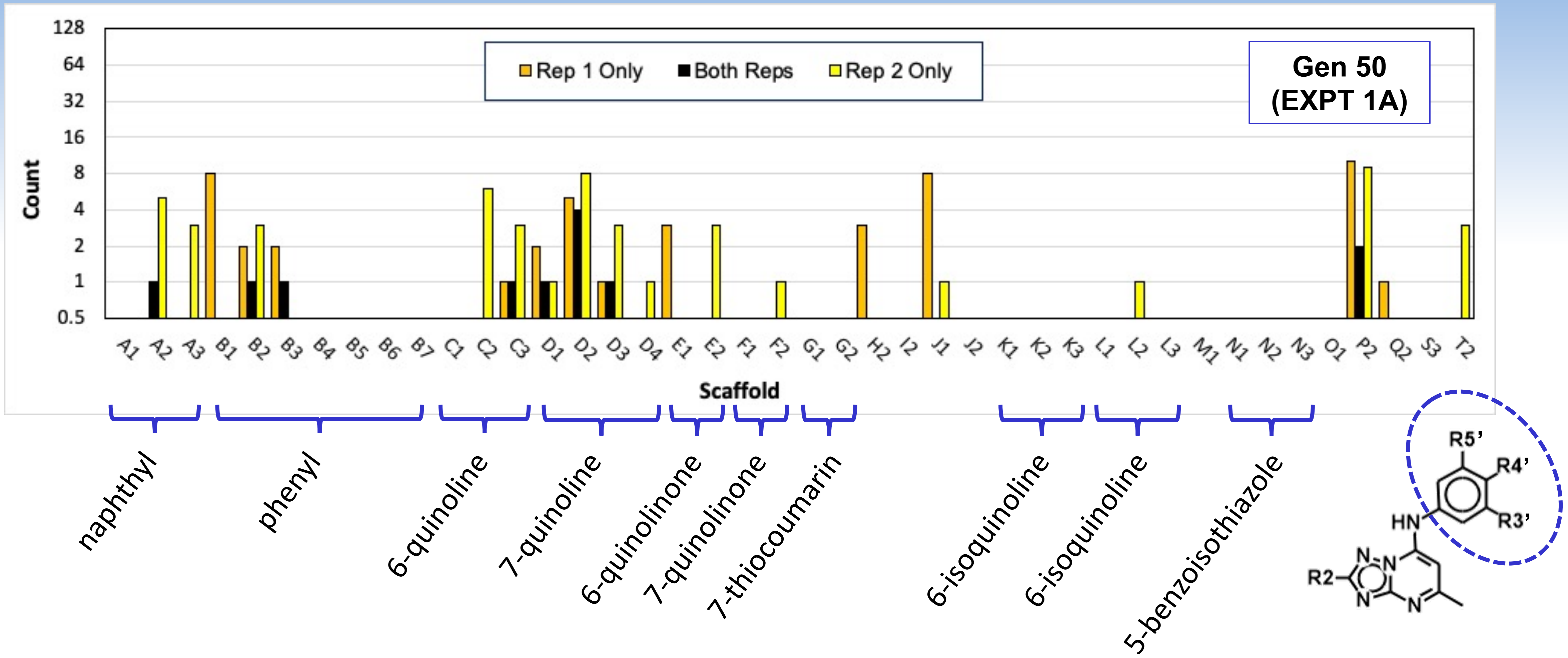
22 (U2)



23 (U2)

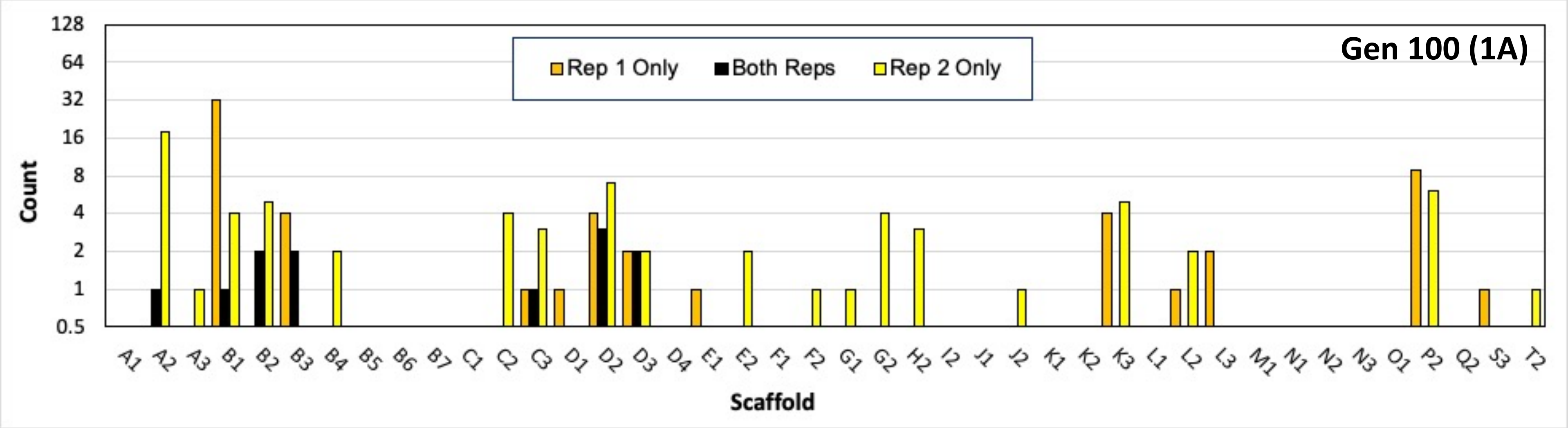
INDIANA UNIVERSITY

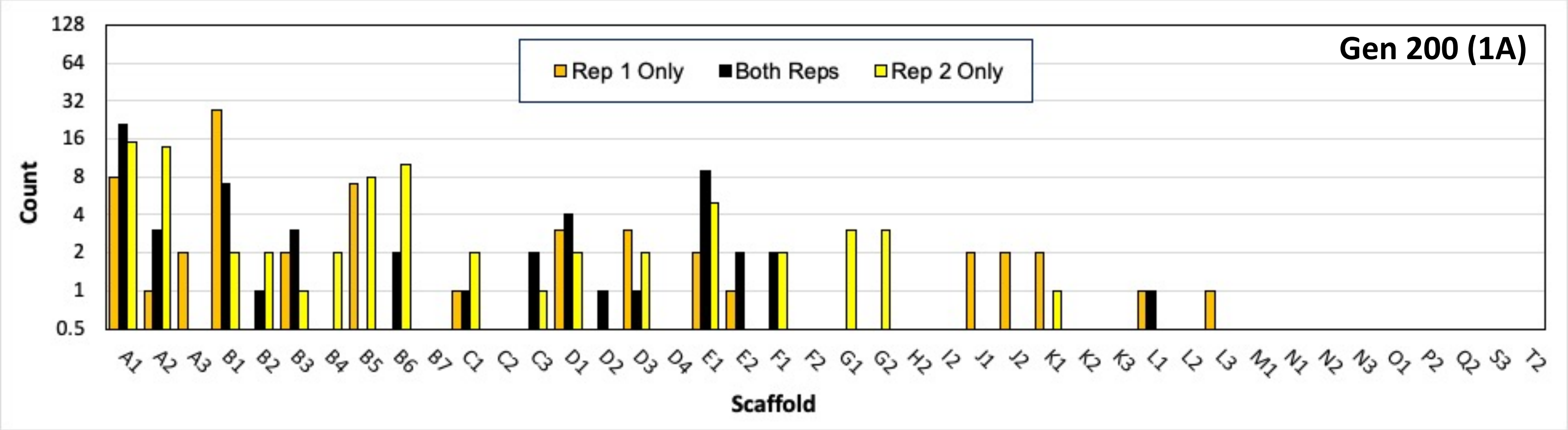
Products are structurally diverse, even early on



INDIANA UNIVERSITY

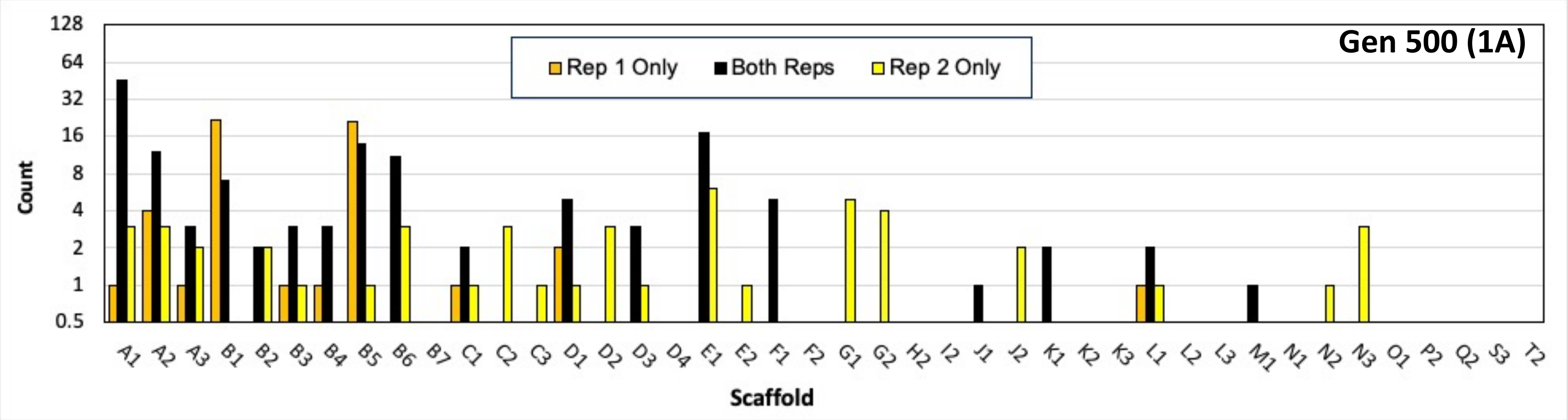
Distribution of AIDD products becomes more focused in mid-run

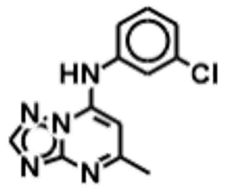




INDIANA UNIVERSITY

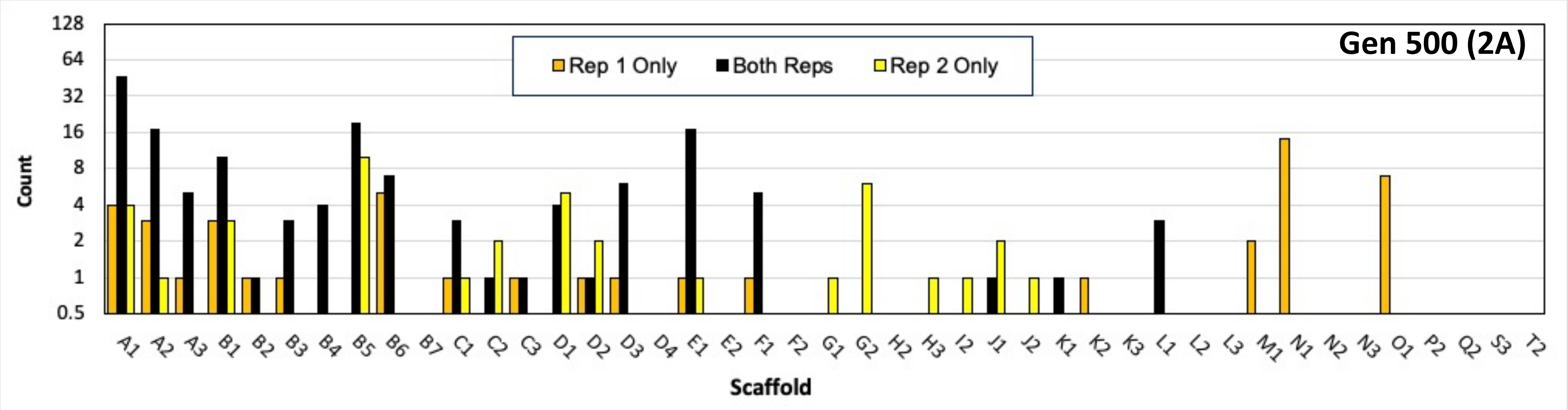
Different seeds yield similar final distributions of products

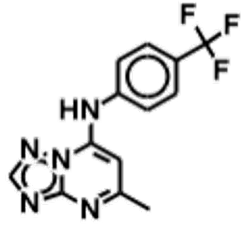




DSM75

(seeds)

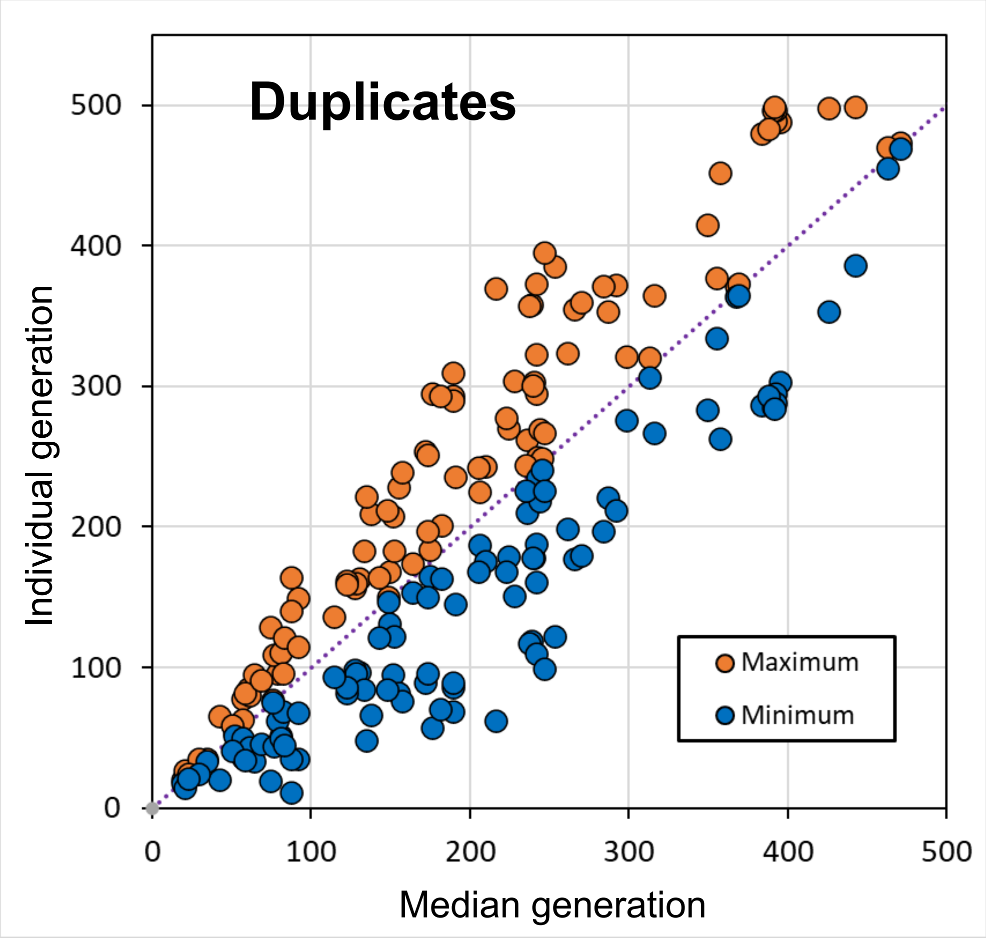


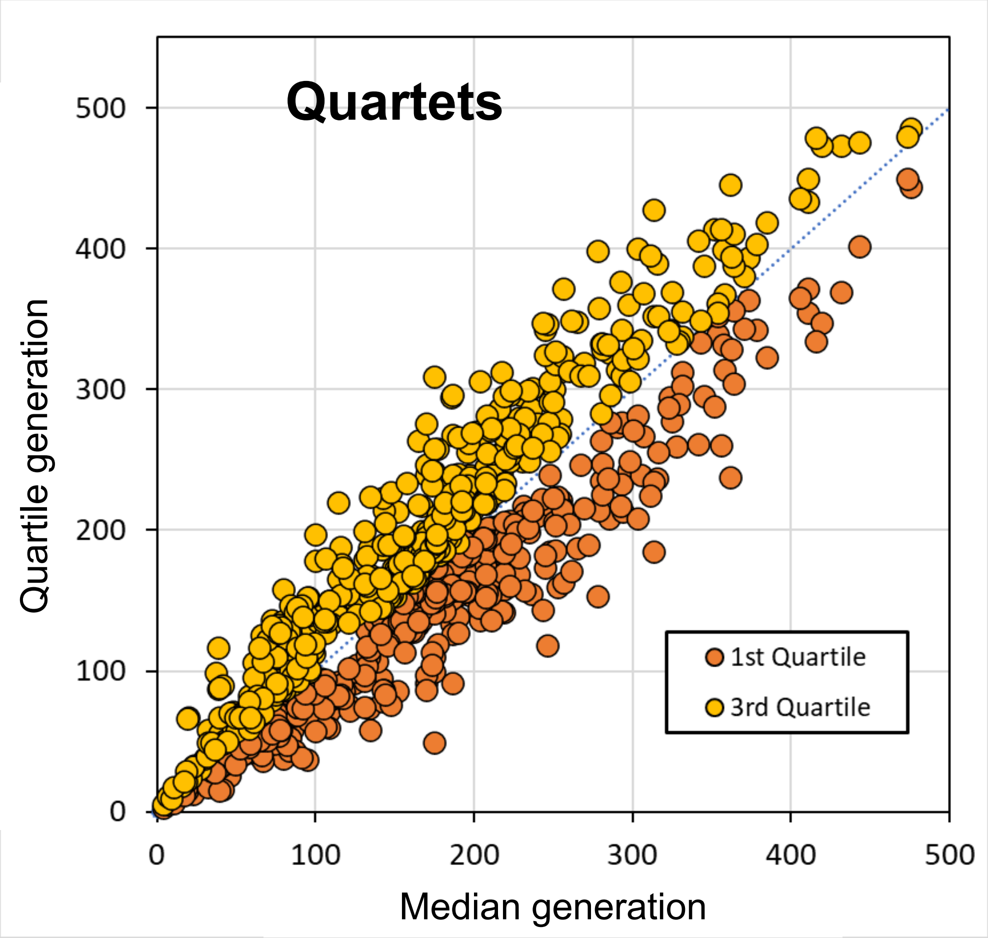


DSM74

INDIANA UNIVERSITY

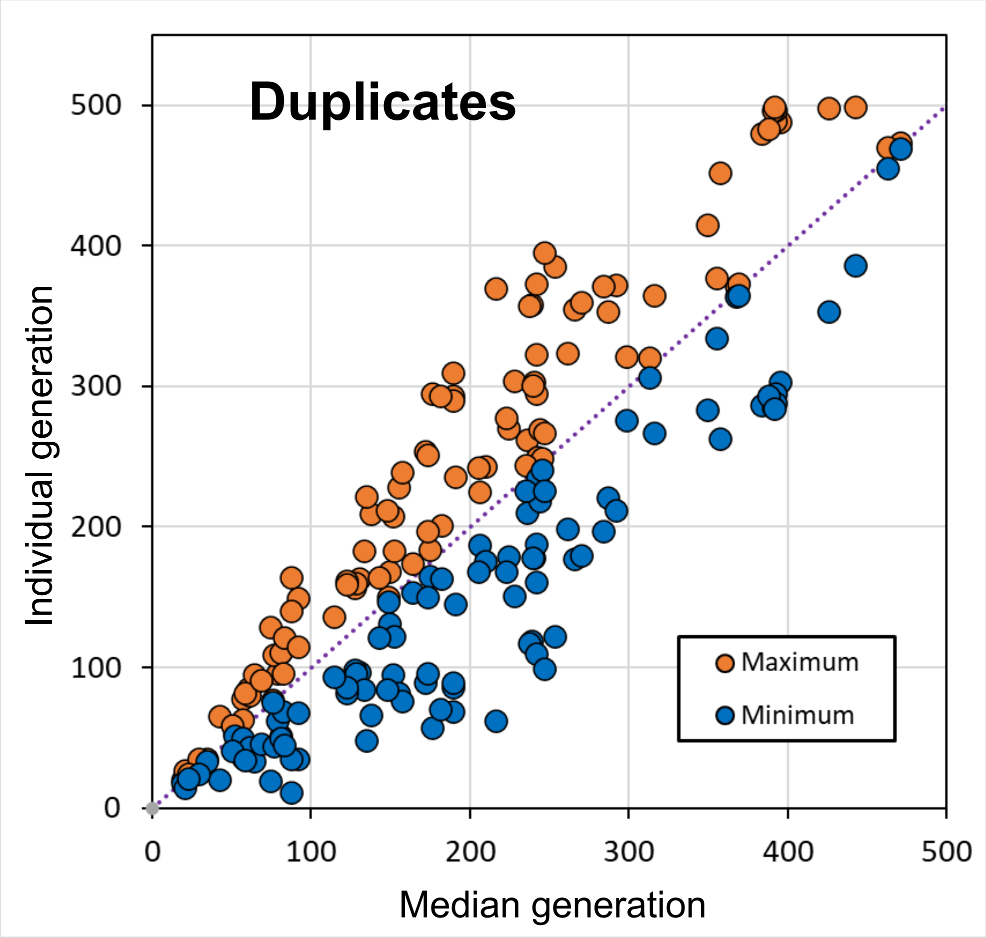
A molecule can be “born” at different times in different runs

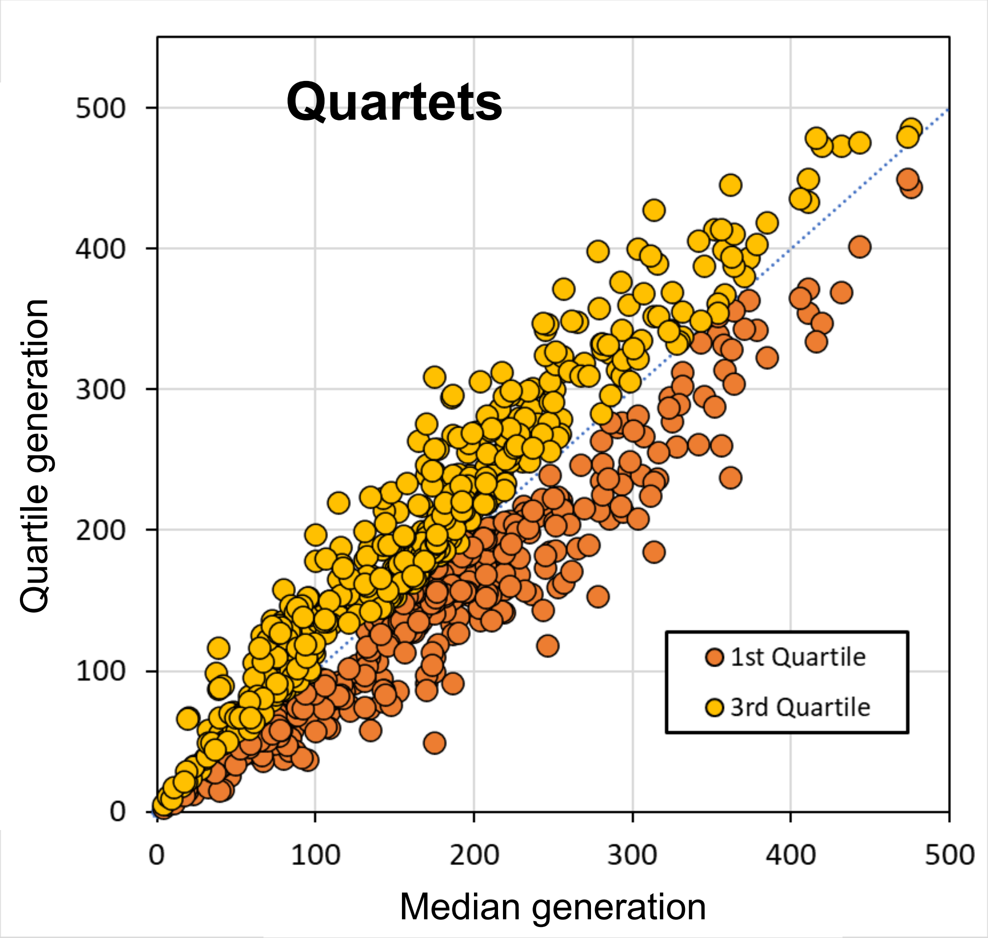




INDIANA UNIVERSITY

A *good* molecule can be "born" at different times in different runs

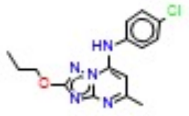
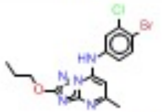
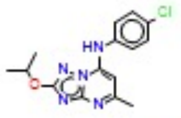
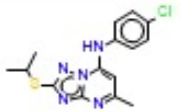
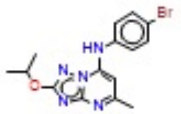
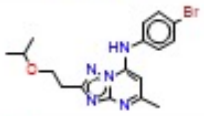
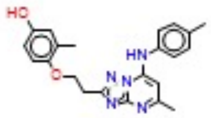




Source: Gen 500 compounds from 1A and 2A replicate experiments after removal of compounds with out-of-scope activity predictions but before application of any secondary filters.

INDIANA UNIVERSITY

Molecular evolution can be complicated – very complicated…

	Structure	Identifier	Generati...	GenZ	RxnCou...	Rxns
361		6470	13	50	10	88,69,90,138,78,86,54,78,80,54
362		10239	21	100	12	88,69,90,138,78,86,54,78,80,54,90,68
363		6694	14	50	10	88,69,90,138,78,86,54,78,80,85
364		18764	38	50	11	88,69,90,138,78,86,54,78,80,85,81
365		13610	28	500	11	88,69,90,138,78,86,54,78,80,85,84
366		35159	71	500	13	88,69,90,138,78,86,54,78,80,85,84,54,54
367		103538	208	500	18	88,69,90,138,78,86,91,92,138,68,85,18,83,80,5

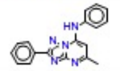
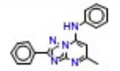
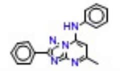
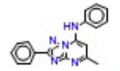
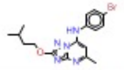
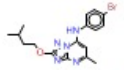
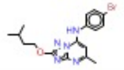
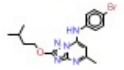
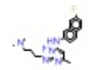
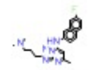
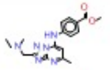
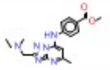
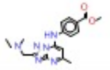
original generation

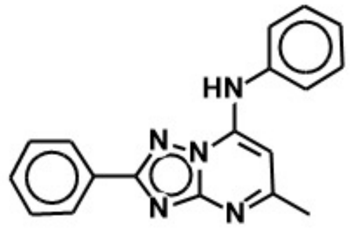
terminal snapshot (50, 100, 200 or 500)

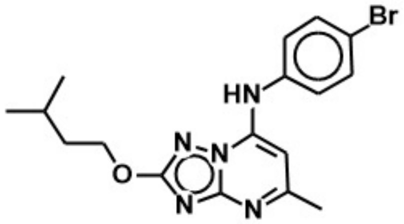
59	Arom_6_ring_to_5(1)
60	Arom_6_ring_to_5(2)
61	Arom_5_ring_to_6
62	Increase_ring_size
63	Decrease_ring_size
64	Change_ring_topology(1)
65	Change_ring_topology(2)
66	Shift_ring_substituents(1)
67	Shift_ring_substituents(2)
68	Shift_ring_substituents(3)
69	Shift_ring_substituents(4)
70	Single_to_double_bond
71	Double_or_triple_to_single_bond
72	Aromatic_to_single_bond
73	Triple_to_double_bond
74	Aromatize_6-membered_ring
75	Aromatize_5-membered_ring
76	De-aromatize_6-membered_ring
77	De-aromatize_5-membered_ring
78	Non-carbon_to_carbon
79	Non-nitrogen_to_nitrogen
80	Non-oxygen_to_oxygen
81	Non-sulfur_to_sulfur
82	Non-fluorine_to_fluorine
83	Non-chlorine_to_chlorine
84	Non-bromine_to_bromine
85	Add_methyl
86	Add_hydroxyl
87	Add_amine
88	Add_fluro

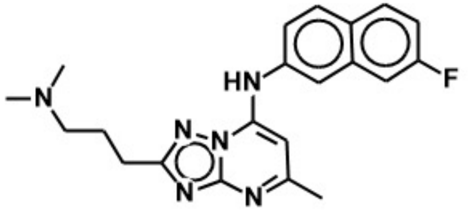
INDIANA UNIVERSITY

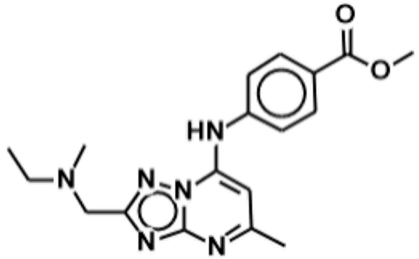
Different evolutionary paths lead to the same molecule

Structure	Identifier	Duplicates	SEED STRUCT...	Generation	RxnCount	Rxns
	542		DSM74 4'-CF3	2	2	91,142
	2422	Duplicate of 542	DSM74 4'-CF3	5	6	97,47,71,139,91,138
	1089	Duplicate of 542	DSM75 3'-Cl	3	4	138,107,5,61
	3337	Duplicate of 542	DSM75 3'-Cl	7	8	92,97,138,139,110,61,78,141
	25412		DSM74 4'-CF3	51	10	86,90,54,85,68,142,35,85,80,54
	148002	Duplicate of 25412	DSM74 4'-CF3	297	17	86,47,54,89,78,81,83,138,54,78,80,54,84,85,55,54,54
	40938	Duplicate of 25412	DSM75 3'-Cl	82	18	138,95,16,89,15,84,138,89,85,78,68,54,68,138,85,35,36,54
	46847	Duplicate of 25412	DSM75 3'-Cl	94	22	86,83,79,88,85,54,68,85,53,54,69,142,90,68,54,68,138,80,78,55,68,54
	48552		DSM74 4'-CF3	98	25	86,90,54,85,68,142,81,83,78,138,122,85,92,64,54,69,138,139,79,54,85,55,54,64,82
	78037	Duplicate of 48552	DSM75 3'-Cl	157	43	138,100,80,130,54,91,86,78,68,76,85,54,82,79,55,88,78,138,58,138,82,125,130,84,55,86,78,80,138,122,78,138,86,69,...
	82679		DSM75 3'-Cl	166	26	92,138,67,138,139,92,78,89,78,85,81,85,54,85,80,54,55,54,139,85,79,54,54,85,55,55
	54341	Duplicate of 82679	DSM75 3'-Cl	109	28	138,115,24,91,85,76,69,58,70,138,85,82,140,78,54,55,54,85,54,36,54,79,35,36,85,54,139,55
	87078	Duplicate of 82679	DSM74 4'-CF3	175	28	86,90,54,85,68,142,35,85,80,54,55,35,79,69,54,68,55,78,138,54,80,42,1,85,85,54,36,55

542


25412


48552


82679


INDIANA UNIVERSITY

”Rediscovered” literature triazolopyrimidines

ID	Substituents			Experiment								
					1A[a] (500)		1B (50)		2A (500)		2B (50)	
	R2	R3’	R4’	Train	rep1	rep2	rep1	rep2	rep1	rep2	rep1	rep2
DSM75	H	Cl	H	+	0 - <50	0 - <50	0 - <5	0 - <5				
DSM74	H	H	CF3	+				5 – 5	0 - <50	0 - <50	0 - 5	0 - <5
DSM1	H	benzo (naphthyl)		+							29 - 30	
DSM89	H	H	Cl	+				3 - 15			2 - 5	3 - 20
DSM100	H	H	OMe	+		23 - 100						
DSM156	H	H	OCH2Ph	+	17 - 500	36 - 500				14 - 500		
DSM227	OMe	H	Cl	-			7 - 40					11 - 15
DSM245	OEt	H	Cl	-			8 - 50	23 - 45		18 - 50	14 - 50	15 - 50
DSM246	OEt	Cl	H	-								
DSM257	SMe	H	Cl	-		27 - 50	5 - 50	6 - 20			10 - 20	10 - 50
DSM268	CH2OH	H	Cl	-								4 - 10
DSM271	Et	H	Cl	-			4 - 50	5 - 5		11 - 50	6 - 20	5 - 50
DSM278	CH2NHMe	H	Cl	-			25 - 30					
DSM279	CH2NMe2	H	Cl	-	21 - 150		1 - 5					19 - 50
DSM282	CH2NMe2	Cl	H	-								18 - 50
DSM299	CH2OMe	H	Cl	-				5 - 5				25 - 35
DSM301	CH2CH2OMe	H	Cl	-		41 - 50	37 - 50					16 - 50
DSM303	CH2CH2OMe	H	CF3	-			38 - 50					
DSM305	CH2OMe	H	CF3	-				6 - 15				5 - 5
DSM307	iPr	H	CF3	-							5 - 5	
DSM309	iPr	H	Cl	-			18 - 50				8 - 20	13 - 50
DSM311	iBu	H	CF3	-			43 - 45					5 - 5
DSM317	CH2CH2OH	H	CF3	-				38-45				

KEY

- DSM75 was the seed structure for Experiments 1A and 1B.
- DSM74 was the seed structure for Experiments 2A and 2B.
- Experiments 1A and 2A were run for 500 generations.
- Experiments 1B and 2B were run for 50 generations.
- The first number in each cell is the generation where the molecule was originally generated.
- The second number in each cell is the last checkpoint generation in which the molecule was observed.
- A “+” in the “Train” column means that the compound was part of the training set for $\log K_i^{gen}$.

INDIANA UNIVERSITY

A natural metaphor for AIDD’s output: trees

INDIANA UNIVERSITY

Summary

- The heart of AIDD is an **evolutionary molecular design engine** that:
  - randomly selects molecules for mutation from a seeded population;
  - generates new analogs by applying randomly selected SMIRKS transforms to them;
  - periodically prunes back the population based on Pareto ranking to create each new generation;
  - revises roulette wheel weights for surviving molecules based on their fitness.
- **Primary structural filters** are used to require or avoid avoid particular substructures.
- **HTPK properties**, activity models, **Risk scores**, synthetic difficulty estimates and external functions can be used as Pareto ranking objectives.
- **Interactive post-processing** with secondary filters is a key part of the workflow.
- The output molecules **are reasonable** from a medicinal chemistry point of view.
- The output molecules are **structurally diverse but focused** into natural subgroups.
- Molecular evolution is **remarkably consistent** overall, shaped more by the Pareto objectives and constraints than by the seed structure(s) or random number seed used.
- Separate runs generally take **different paths** to produce recurrent molecules.

INDIANA UNIVERSITY

Thanks to:

coauthors
- Michael S. Lawless
- David Miller
- Marvin Waldman

Other SLP developers
- Pankaj R. Daga
- Robert Fraczkiewicz
- Dechuan Zhuang
- Jinhua Zhang

drbobclark@gmail.com
clarkrod@indiana.edu

Simulations Plus, Inc., makes ADMET Predictor freely available through their University+ academic licensing program and underwrote my ACS attendance.

Thank you for your kind attention!

SimulationsPlus

INDIANA UNIVERSITY