Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting information

*i*SIM: Instant Similarity

Kenneth López-Pérez,[1] Taewon D. Kim,[1] Ramón Alain Miranda-Quintana[1*]

1. Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, USA.

[*] Email: quintana@chem.ufl.edu

Table 1: Library codes and number of molecules in the used CHEMBL libraries obtained from van Tilborg et al.[1]

Libraries	n
CHEMBL2034_Ki	750
CHEMBL218_EC50	1031
CHEMBL4616_EC50	682
CHEMBL231_Ki	973
CHEMBL2971_Ki	976
CHEMBL234_Ki	3657
CHEMBL237_Ki	2602
CHEMBL2047_EC50	631
CHEMBL233_Ki	3142
CHEMBL204_Ki	2754
CHEMBL244_Ki	3097
CHEMBL236_Ki	2598
CHEMBL287_Ki	1328
CHEMBL4203_Ki	731
CHEMBL1871_Ki	659
CHEMBL1862_Ki	794
CHEMBL235_EC50	2349

CHEMBL264_Ki	2862
CHEMBL4005_Ki	960
CHEMBL2147_Ki	1456
CHEMBL238_Ki	1052
CHEMBL237_EC50	955
CHEMBL4792_Ki	1471
CHEMBL262_Ki	856
CHEMBL214_Ki	3317
CHEMBL228_Ki	1704
CHEMBL239_EC50	1721
CHEMBL2835_Ki	615
CHEMBL3979_EC50	1125
CHEMBL219_Ki	1859

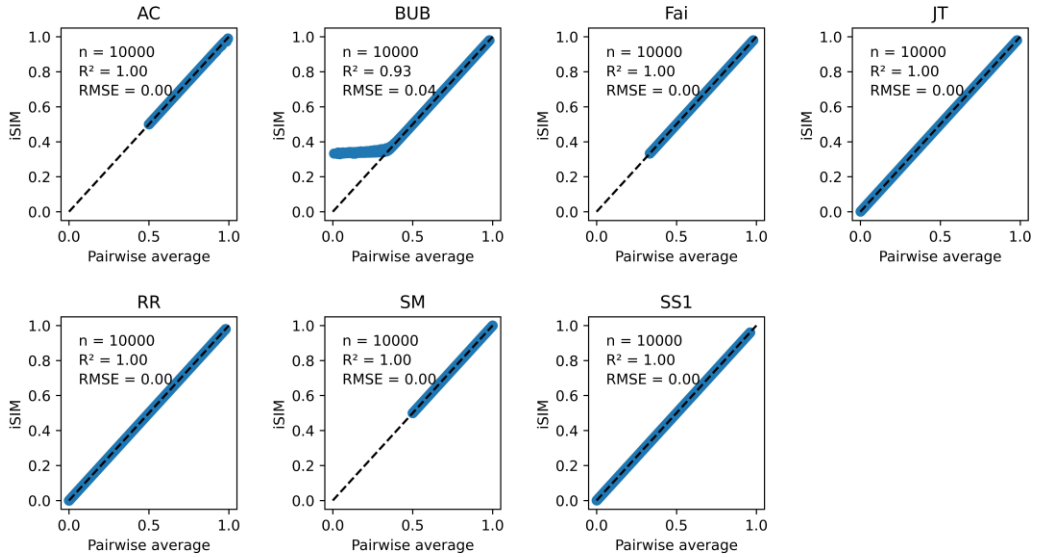

Figure S1: *i*SIM vs pairwise results for 10,000 randomly generated libraries. Molecules represented with random generated binary fingerprints. Plots separated by similarity index: AC (Austin-Colwell)[2], BUB (Baroni-Urbani-Buser)[3], Fai (Faith)[4], JT (Jaccard-Tanimoto)[5,6], RR (Russel-Rao)[7], SM (Sokal-Michener)[8] and SS1 (Sokal-Sneath 1)[9].

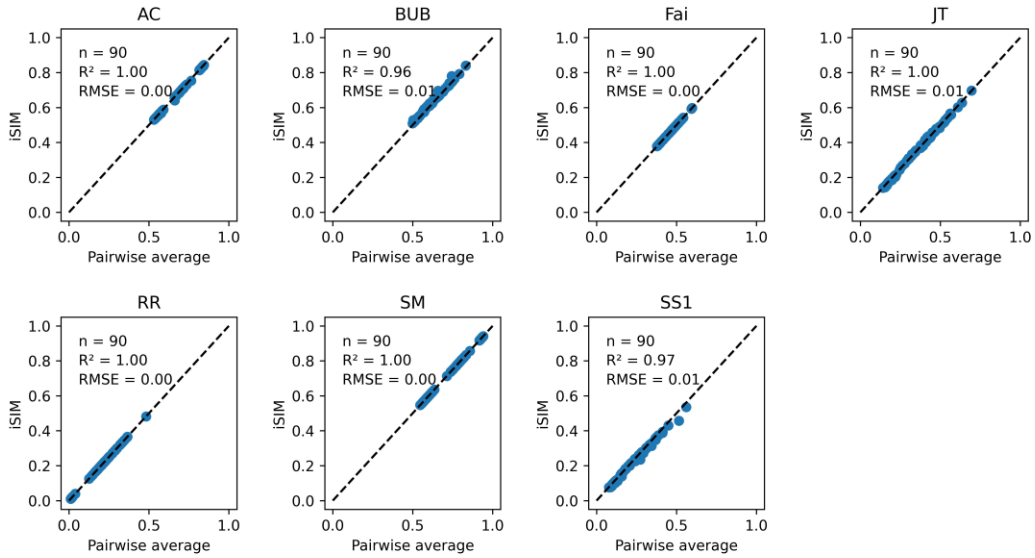

Figure S2: *i*SIM vs pairwise results for 30 CHEMBL libraries. Molecules represented with three different types of binary fingerprints: RDKIT (m = 2048), MACCS (m = 167) and ECFP4 (m = 2014). Plots separated by similarity index: AC (Austin-Colwell)[2], BUB (Baroni-Urbani-Buser)[3], Fai (Faith)[4], JT (Jaccard-Tanimoto)[5,6], RR (Russel-Rao)[7], SM (Sokal-Michener)[8] and SS1 (Sokal-Sneath 1)[9].

Table 2: List of real number descriptors for libraries computed from RDKit. [10]

MaxEStateIndex	BertzCT	PEOE_VSA3	SlogP_VSA7	HeavyAtomCount	fr_Ar_OH	fr_barbitur	fr_nitro_arom
MinEStateIndex	Chi0	PEOE_VSA4	SlogP_VSA8	NHOHCount	fr_COO	fr_benzene	fr_nitro_arom_nonortho
MaxAbsEStateIndex	Chi0n	PEOE_VSA5	SlogP_VSA9	NOCount	fr_COO2	fr_benzodiazepine	fr_nitroso
MinAbsEStateIndex	Chi0v	PEOE_VSA6	TPSA	NumAliphaticCarbocycles	fr_C_O	fr_bicyclic	fr_oxazole
qed	Chi1	PEOE_VSA7	EState_VSA1	NumAliphaticHeterocycles	fr_C_O_noCOO	fr_diazo	fr_oxime
MolWt	Chi1n	PEOE_VSA8	EState_VSA10	NumAliphaticRings	fr_C_S	fr_dihydropyridine	fr_para_hydroxylation
HeavyAtomMolWt	Chi1v	PEOE_VSA9	EState_VSA11	NumAromaticCarbocycles	fr_HOCCN	fr_epoxide	fr_phenol
ExactMolWt	Chi2n	SMR_VSA1	EState_VSA2	NumAromaticHeterocycles	fr_Imine	fr_ester	fr_phenol_noOrthoHbond
NumValenceElectrons	Chi2v	SMR_VSA10	EState_VSA3	NumAromaticRings	fr_NH0	fr_ether	fr_phos_acid
NumRadicalElectrons	Chi3n	SMR_VSA2	EState_VSA4	NumHAcceptors	fr_NH1	fr_furan	fr_phos_ester
MaxPartialCharge	Chi3v	SMR_VSA3	EState_VSA5	NumHDonors	fr_NH2	fr_guanido	fr_piperdine
MinPartialCharge	Chi4n	SMR_VSA4	EState_VSA6	NumHeteroatoms	fr_N_O	fr_halogen	fr_piperzine
MaxAbsPartialCharge	Chi4v	SMR_VSA5	EState_VSA7	NumRotatableBonds	fr_Ndealkylation1	fr_hdrzine	fr_priamide
MinAbsPartialCharge	HallKierAlpha	SMR_VSA6	EState_VSA8	NumSaturatedCarbocycles	fr_Ndealkylation2	fr_hdrzone	fr_prisulfonamd
FpDensityMorgan1	Ipc	SMR_VSA7	EState_VSA9	NumSaturatedHeterocycles	fr_Nhpyrrole	fr_imidazole	fr_pyridine
FpDensityMorgan2	Kappa1	SMR_VSA8	VSA_EState1	NumSaturatedRings	fr_SH	fr_imide	fr_quatN
FpDensityMorgan3	Kappa2	SMR_VSA9	VSA_EState10	RingCount	fr_aldehyde	fr_isocyan	fr_sulfide
BCUT2D_MWHI	Kappa3	SlogP_VSA1	VSA_EState2	MolLogP	fr_alkyl_carbamate	fr_isothiocyan	fr_sulfonamd
BCUT2D_MWLOW	LabuteASA	SlogP_VSA10	VSA_EState3	MolMR	fr_alkyl_halide	fr_ketone	fr_sulfone
BCUT2D_CHGHI	PEOE_VSA1	SlogP_VSA11	VSA_EState4	fr_Al_COO	fr_allylic_oxid	fr_ketone_Topliss	fr_term_acetylene
BCUT2D_CHGLO	PEOE_VSA10	SlogP_VSA12	VSA_EState5	fr_Al_OH	fr_amide	fr_lactam	fr_tetrazole
BCUT2D_LOGPHI	PEOE_VSA11	SlogP_VSA2	VSA_EState6	fr_Al_OH_noTert	fr_amidine	fr_lactone	fr_thiazole
BCUT2D_LOGPLOW	PEOE_VSA12	SlogP_VSA3	VSA_EState7	fr_ArN	fr_aniline	fr_methoxy	fr_thiocyan
BCUT2D_MRHI	PEOE_VSA13	SlogP_VSA4	VSA_EState8	fr_Ar_COO	fr_aryl_methyl	fr_morpholine	fr_thiophene
BCUT2D_MRLOW	PEOE_VSA14	SlogP_VSA5	VSA_EState9	fr_Ar_N	fr_azide	fr_nitrile	fr_unbrch_alkane
BalabanJ	PEOE_VSA2	SlogP_VSA6	FractionCSP3	fr_Ar_NH	fr_azo	fr_nitro	fr_urea

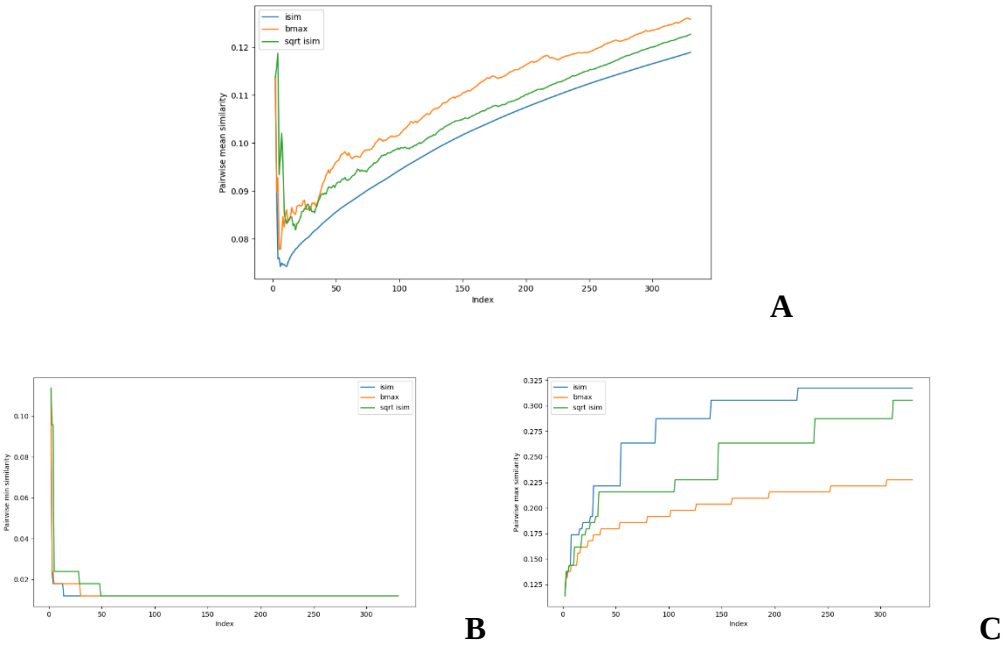

Figure S3: MaxMin[11] (bmax, yellow), *i*RR (isim, blue), and sqrt_*i*RR (sqrt_isim, green) results for the diversity sampling of the CHEMBL214 dataset represented with MACCS binary fingerprints: A) pairwise similarity of the Selected set, B) minimum similarity between elements of the Selected set, C) maximum similarity between elements of the selected set.

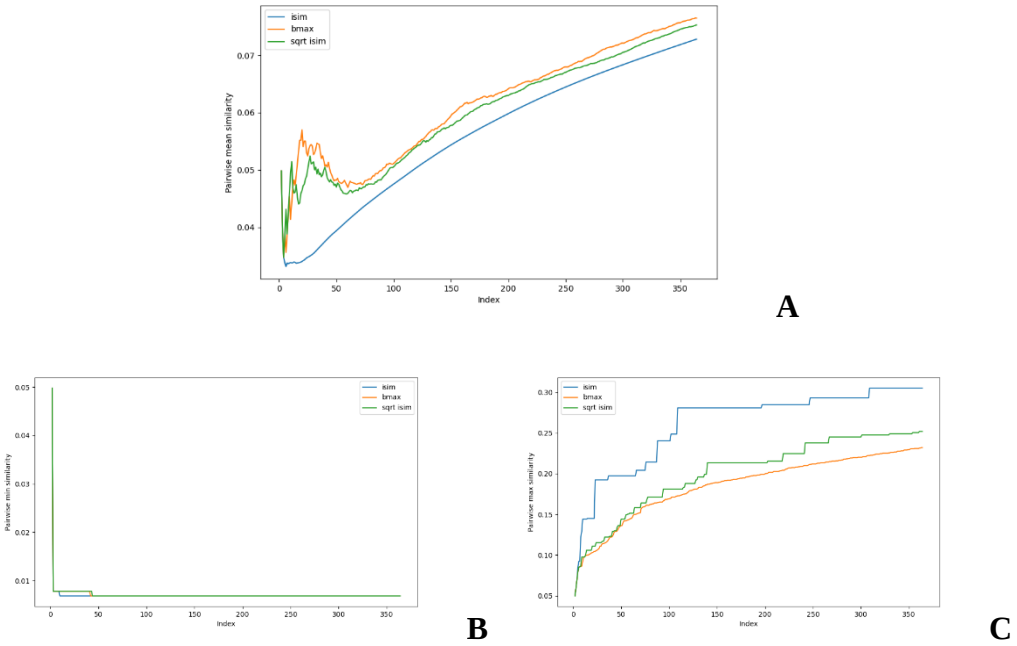

Figure S4: MaxMin[11] (bmax, yellow), *i*RR (isim, blue), and sqrt_*i*RR (sqrt_isim, green) results for the diversity sampling of the CHEMBL234 dataset represented with RDKIT binary fingerprints: A) pairwise similarity of the Selected set, B) minimum similarity between elements of the Selected set, C) maximum similarity between elements of the selected set.

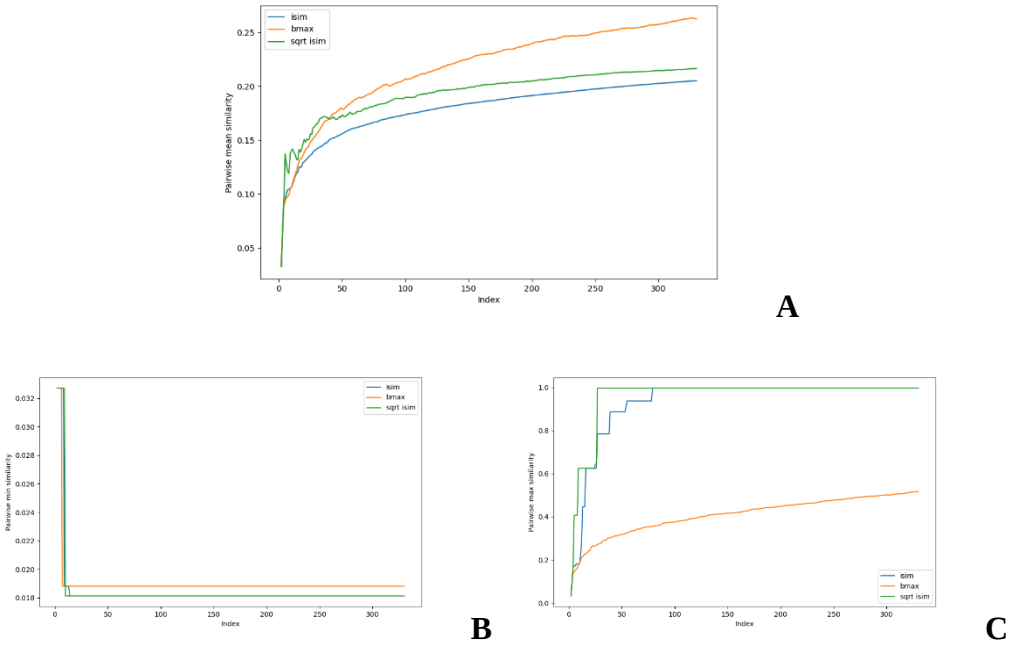

Figure S5: MaxMin[11] (bmax, yellow), *i*JT (isim, blue), and sqrt_*i*JT (sqrt_isim, green) results for the diversity sampling of the CHEMBL214 dataset represented with RDKIT binary fingerprints: A) pairwise similarity of the Selected set, B) minimum similarity between elements of the Selected set, C) maximum similarity between elements of the selected set.

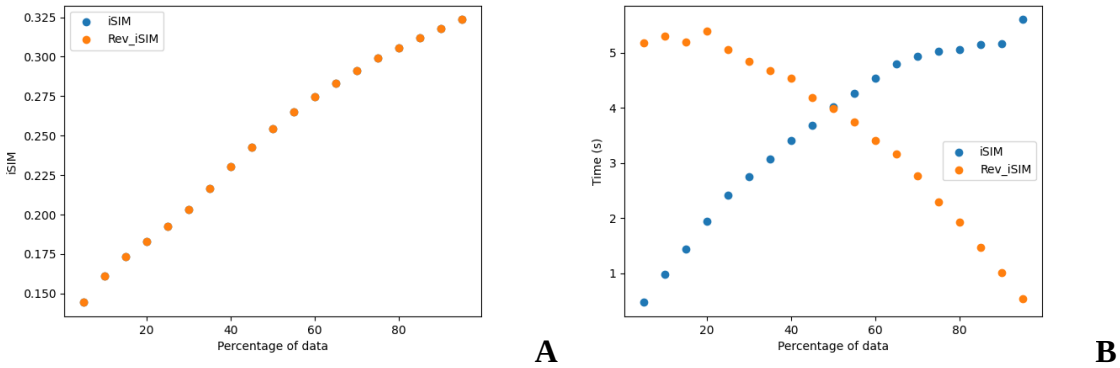

Figure S6: A) *i*SIMDiv and *i*SIMRevDiv selections for different data percentages (5-95%, in 5% steps) for a n = 1,000 randomly generated dataset of binary fingerprints of length m = 166. B) Computing time variation of the diversity selection methods with the data percentage selected.

References

(1) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J Chem Inf Model* **2022**, *62* (23), 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.

(2) Austin, B.; Colwell, R. R. Evaluation of Some Coefficients for Use in Numerical Taxonomy of Microorganisms. *Int J Syst Bacteriol* **1977**, *27* (3), 204–210. https://doi.org/10.1099/00207713-27-3-204.

(3) Baroni-Urbani, C.; Buser, M. W. Similarity of Binary Data. *Syst Zool* **1976**, *25* (3), 251. https://doi.org/10.2307/2412493.

(4) Faith, D. P.; Minchin, P. R.; Belbin, L. Compositional Dissimilarity as a Robust Measure of Ecological Distance. *Vegetatio* **1987**, *69* (1–3), 57–68. https://doi.org/10.1007/BF00038687.

(5) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science (1979)* **1960**, *132* (3434), 1115–1118. https://doi.org/10.1126/science.132.3434.1115.

(6) Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE. *New Phytologist* **1912**, *11* (2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

(7) Russell, P. F.; Rao, T. R.; others. On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras. *J Malar Inst India* **1940**, *3* (1).

(8) Sokal, R. R.; Michener, C. D. University of Kansas. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas science bulletin. University of Kansas* **1958**.

(9) Sokal, R. R.; Sneath, P. H. A.; others. Principles of Numerical Taxonomy. *Principles of numerical taxonomy.* **1963**.

(10) RDKit: Open-Source Cheminformatics. Https://Www.Rdkit.Org.

(11) Kuo, C.; Glover, F.; Dhir, K. S. Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming*. *Decision Sciences* **1993**, *24* (6), 1171–1185. https://doi.org/10.1111/j.1540-5915.1993.tb00509.x.