

SUPPORTING INFORMATION

Machine learning-driven antiviral libraries targeting respiratory viruses

Gabriela Valle-Núñez, Raziel Cedillo-González, Juan F. Avellaneda-Tamayo, Fernanda I. Saldívar-González, Diana L. Prado-Romero, José L. Medina-Franco*

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico

*Correspondence author: medinajl@unam.mx; Tel.: +52-55-5622-3899

Contents

	Page
Table S1	Priority Diseases Identified by the World Health Organization in 2024 with a Focus on Medium and High PHEIC Risk and Respiratory Diseases. S2
Table S2	Molecular targets according to modelability index criteria. S4
Table S3	Machine learning algorithms used in this work, generated by default with PyCaret. S5
Table S4	Average statistics for Matthews Correlation Coefficient (MCC) for each target. S6
Figure S1	Distribution of physicochemical and constitutional properties computed for antivirals by target compared with approved antivirals from DrugBank. S11
Table S5	Descriptive statistics of physicochemical and constitutional descriptors computed for antivirals by target compared with approved antivirals from DrugBank. S13
Figure S2	Most frequent scaffolds presented in the data set utilized for constructing the ML models S19
Figure S3	Most frequent ring systems presented in the data set utilized for constructing the ML models. S20
Figure S4	Most frequent scaffolds identified in the newly designed antiviral-focused VS data set library. S20
Figure S5	Most frequent ring systems identified in the newly designed antiviral-focused VS data set library. S21
Figure S6	Chemical multiverse visualization of seven antivirals data sets focused on different targets as compared with approved antivirals from DrugBank. S21
Figure S7	Constellation plot of scaffolds from the data set utilized for constructing the ML models, categorized by target. S22
Table S6	MCC values calculated for different validation phases. S23
Table S7	Descriptive statistics of Jaccard and Euclidean distances computed for training sets by target. S25
Figure S8	Chemical space visualization of active compounds generated with t-SNE based on the Morgan Chiral of radius 2 (2048-bit) fingerprint for each newly designed library. S26
References	S27

Table S1. Priority Diseases Identified by the World Health Organization in 2024 with a Focus on Medium and High PHEIC Risk and Respiratory Diseases.

Family	PHEIC risk	Priority Pathogens	Prototype Pathogens	Pathogen X	Geographic Distribution*						Respiratory syndromes**	
					AFR	AMR	EMR	EUR	SEAR	WPR	Common	Less Common
Adenoviridae	Low-Medium	–	Recombinant Mastadenovirus	Mastadenovirus blackbeardi 21, Mastadenovirus blackbeardi 55, Mastadenovirus blackbeardi 7, Mastadenovirus exoticum	X	X	X	X	X	X	Bronchiolitis, Pneumonia.	Influenza-like illness, Croup, Common cold.
			Mastadenovirus blackbeardi serotype 14			X				X		
Arenaviridae	High	Mammarenavirus lassaense	Mammarenavirus lassaense	Mammarenavirus chapareense, Mammarenavirus choriomeningitis, Mammarenavirus lujoense, Mammarenavirus lujoense	X						–	–
		–	Mammarenavirus juninense			X						
		–	Mammarenavirus lujoense		X							
Coronaviridae	High	Subgenus Merbecovirus	Subgenus Merbecovirus	Alphacoronavirus suis (CCoV-HuPn-2018), Alphacoronavirus porci			X				Common cold, Pneumonia. Include SARS, SARS-CoV-2 and MERS.	Influenza-like illness.
		Subgenus Sarbecovirus	Subgenus Sarbecovirus		X	X	X	X	X	X		
Filoviridae	High	Orthoebolavirus zaireense	Orthoebolavirus zaireense	Orthoebolavirus bombaliense, Orthoebolavirus X, Orthoebolavirus restonense	X						–	–
		Orthomarburgvirus marburgense	–		X						–	–
		Orthoebolavirus sudanense	–		X		X				–	–
Flaviviridae	High	Orthoflavivirus zikaense	Orthoflavivirus zikaense	Orthoflavivirus japonicum, Orthoflavivirus encephalitis, Orthoflavivirus nilense	X	X			X	X	–	–
		Orthoflavivirus denguei	Orthoflavivirus denguei		X	X	X	X	X	X		
		Orthoflavivirus flavi	–		X	X						
		–	Orthoflavivirus encephalitis					X		X	–	–
		–	Orthoflavivirus nilense		X	X	X	X	X	X		
Hantaviridae	High	Orthohantavirus sinnombreense	Orthohantavirus sinnombreense	–		X					Hantavirus	–

		Orthohantavirus hantanense	–	–				X		X	pulmonary syndrome.	
Nairoviridae	High	Orthonairovirus haemorrhagiae	Orthonairovirus haemorrhagiae	–	X		X	X		X	–	–
Orthomyxoviridae	High	Alphainfluenzavirus Influenzae H1	Alphainfluenzavirus Influenzae H1	Alphainfluenzavirus influenzae (H9N2), Betainfluenzavirus influenzae	X	X	X	X	X	X	Bronchiolitis, Influenza-like illness, Pneumonia.	Croup, Common cold.
		Alphainfluenzavirus Influenzae H2	–		X	X	X	X	X	X		
		Alphainfluenzavirus Influenzae H3	–		X	X	X	X	X	X		
		Alphainfluenzavirus Influenzae H5	Alphainfluenzavirus Influenzae H5		X	X	X	X	X	X		
		Alphainfluenzavirus Influenzae H6	–		X	X	X	X	X	X		
		Alphainfluenzavirus Influenzae H7	–		X	X	X	X	X	X		
		Alphainfluenzavirus Influenzae H10	–		X	X	X	X	X	X		
Paramyxoviridae	High	Henipavirus nipahense	Henipavirus nipahense	–					X	X	Bronchiolitis, Croup, Pneumonia.	Influenza-like illness, Common cold.
Parvoviridae	Low	–	Protoparvovirus carnivoran	Protoparvovirus carnivoran	X	X	X	X	X	X	–	–
Phenuiviridae	High	Bandavirus dabieense	Bandavirus dabieense	Phlebovirus riftense					X	X	–	–
		–	Phlebovirus riftense		X						–	–
Picornaviridae	Medium	Enterovirus coxsackiepol	–	Enterovirus deconjecti 68, Enterovirus alphacoxsackie 71	X		X		X		Bronchiolitis, Common cold.	Pneumonia, Common cold.
		–	Enterovirus alphacoxsackie 71		X	X	X	X	X	X		
		–	Enterovirus deconjecti 68		X	X	X	X	X	X		
Pneumoviridae	Low-Medium	–	Metapneumovirus hominis	–	X	X	X	X	X	X	Respiratory syncytial virus infections, bronchiolitis, and pneumonia.	–
Poxviridae	High	Orthopoxvirus variola	–	Orthopoxvirus cowpox							–	–
		–	Orthopoxvirus vaccinia			X	X		X		–	–

		Orthopoxvirus monkeypox	Orthopoxvirus monkeypox		X	X	X	X	X	X	–	–
Retroviridae	Medium	Lentivirus humimdefl	Lentivirus humimdefl	Gammaretrovirus gibleu-like viruses in koalas, bats and rodents, Lentivirus simimdef	X	X	X	X	X	X	–	–
Togaviridae	High	Alphavirus chikungunya	Alphavirus chikungunya	Alphavirus eastern, Alphavirus madariaga, Alphavirus mayaro, Alphavirus rossriver	X	X			X	X	–	–
		Alphavirus venezuelan	Alphavirus venezuelan			X					–	–
Pathogen X	Pathogen X	Pathogen X	Pathogen X	–							–	–

Information compiled from: World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), and the MSD Manual.^{1,2} The families highlighted in bold represent the viral families associated with respiratory diseases, which are the primary focus of this project.

Prototype Pathogens: Representative pathogens within a family, selected to serve as models for fundamental and translational research, enabling the development of MCMMs (Medical Countermeasures) that can be applied to other family members.

*AFR represents the African region, AMR the American region, EMR the Eastern Mediterranean region, EUR the European region, SEAR the South-East Asia region, and WPR the Western Pacific region. The cell color in the "Geographic Distribution" column indicates the classification of the pathogen: blue for priority pathogens, red for prototype pathogens, and green for those classified as both.

Pathogen X: A pathogen not currently posing a PHEIC threat but with evidence suggesting it could become one in the future due to significant changes in its biology, transmission patterns, or virulence.

Table S2. Molecular targets according to modelability index criteria. Selected targets are highlighted in yellow.

Target	Organism	Count	pIC ₅₀ Median	Active	Inactive	MODI MACCS keys	MODI Morgan Chiral 2
Capsid protein	HEV-71	22	6.26	20	2	0.95	0.95
	HRV	18	6.27	15	3	0.83	0.89
Fusion glycoprotein F0	HRSV	249	8.00	240	9	0.96	0.96
Helicase (NSP13)	SARS-CoV	9	4.30	0	9	1.00	1.00
	SARS-CoV-2	3	4.52	1	2	0.67	0.67
Hemagglutinin	IAV	25	5.58	19	6	0.80	0.72
Hemagglutinin- neuraminidase	HPIV-1	36	3.82	2	34	1.00	1.00
M2 proton channel	HRSV	1	8.70	1	0	-	-
	IAV	92	5.45	68	24	0.82	0.83
MTase (NSP14)	SARS-CoV-2	39	6.36	38	1	0.97	0.97
Mpro	FCoV	12	5.74	12	0	1.00	1.00
	HCoV-229E	11	6.54	7	4	0.45	0.45
	HEV-71	26	6.70	26	0	1.00	1.00
	HRV	21	7.10	21	0	1.00	1.00
	MERS-CoV	12	6.15	12	0	1.00	1.00
	SARS-CoV	197	4.52	77	120	0.66	0.77
	SARS-CoV-2	815	6.35	651	164	0.88	0.91
Neuraminidase	IAV	1123	5.72	733	390	0.88	0.91
	IBV	202	5.47	132	70	0.72	0.71
PLP	HCoV-NL63	5	4.36	0	5	1.00	1.00
	MERS-CoV	10	4.63	5	5	0.80	0.50
	SARS-CoV-2	107	5.98	94	13	0.87	0.87
Polymerase (PA)	IAV	256	5.40	151	105	0.84	0.88
Polymerase (PB2)	IAV	79	7.62	79	0	1.00	1.00
Protease	HRV	389	5.96	298	91	0.83	0.85
Protein P	HRSV	1	8.70	1	0	-	-
RdRp	HRSV	1	8.70	1	0	-	-
	IAV	143	3.93	13	130	0.90	0.93
	SARS-CoV-2	46	5.16	41	5	0.87	0.74
Spike glycoprotein	SARS-CoV	15	5.20	13	2	0.67	0.87
	SARS-CoV-2	44	5.40	28	16	0.86	0.82
gpG	NiV	7	5.40	6	1	0.86	0.86
TOTAL		4016				0.86	0.86

* Seven selected targets are highlighted in yellow.

Table S3. Machine learning algorithms used in this work, generated by default with PyCaret.

Acronym	Model
lr	Logistic Regression
ridge	Ridge Classifier
lda	Linear Discriminant Analysis
rf	Random Forest Classifier
nb	Naive Bayes
gbc	Gradient Boosting Classifier
ada	Ada Boost Classifier
et	Extra Trees Classifier
qda	Quadratic Discriminant Analysis
lightbm	Light Gradient Boosting Machine
knn	K Neighbors Classifier
dt	Decision Tree Classifier
xgboost	Extreme Gradient Boosting
dummy	Dummy Classifier
svm	SVM - Linear Kernel

Table S4. Average statistics for Matthews Correlation Coefficient (MCC) for each target.

Target: IAV_Polymerase (PA)			
Model	MCC_cross_validation	MCC_external_validation	average_MCC
AdaBoost Classifier	0.666	0.587	0.627
Decision Tree Classifier	0.655	0.382	0.518
Dummy Classifier	0.075	-0.007	0.034
Extra Trees Classifier	0.712	0.786	0.749
Extreme Gradient Boosting	0.687	0.246	0.467
Gradient Boosting Classifier	0.715	0.679	0.697
K Neighbors Classifier	0.663	0.080	0.371
Light Gradient Boosting Machine	0.634	0.642	0.638
Linear Discriminant Analysis	0.771	0.495	0.633
Logistic Regression	0.771	0.515	0.643
Naive Bayes	0.572	0.086	0.329
Quadratic Discriminant Analysis	0.289	0.000	0.144
Random Forest Classifier	0.716	0.551	0.634
Ridge Classifier	0.771	0.495	0.633
SVM - Linear Kernel	0.807	0.733	0.770
Target: HRV_Protease			
AdaBoost Classifier	0.643	0.398	0.520
Decision Tree Classifier	0.555	0.315	0.435
Dummy Classifier	0.035	-0.080	-0.022
Extra Trees Classifier	0.619	0.177	0.398
Extreme Gradient Boosting	0.680	0.356	0.518
Gradient Boosting Classifier	0.744	0.500	0.622
K Neighbors Classifier	0.423	0.059	0.241
Light Gradient Boosting Machine	0.635	0.262	0.449
Linear Discriminant Analysis	0.530	0.271	0.400
Logistic Regression	0.520	0.301	0.410
Naive Bayes	0.346	0.181	0.264
Quadratic Discriminant Analysis	0.199	-0.295	-0.048
Random Forest Classifier	0.595	0.340	0.468

Table S4 (cont). Average statistics for Matthews Correlation Coefficient (MCC) for each target.

Target: HRV_Protease			
Model	MCC_cross_validation	MCC_external_validation	average_MCC
Ridge Classifier	0.489	0.358	0.423
SVM - Linear Kernel	0.566	0.000	0.283
Target: IAV_M2 proton channel			
AdaBoost Classifier	0.248	0.000	0.124
Decision Tree Classifier	0.368	-0.149	0.110
Dummy Classifier	0.338	0.077	0.207
Extra Trees Classifier	0.322	0.000	0.161
Extreme Gradient Boosting	0.152	0.000	0.076
Gradient Boosting Classifier	0.352	0.000	0.176
K Neighbors Classifier	0.204	0.167	0.186
Light Gradient Boosting Machine	0.387	-0.102	0.143
Linear Discriminant Analysis	0.388	-0.102	0.143
Logistic Regression	0.347	0.000	0.173
Naive Bayes	0.239	-0.016	0.111
Quadratic Discriminant Analysis	0.441	0.000	0.221
Random Forest Classifier	0.090	0.000	0.045
Ridge Classifier	0.302	0.000	0.151
SVM - Linear Kernel	0.318	0.000	0.159
Target: SARS-CoV_Mpro			
AdaBoost Classifier	0.511	0.118	0.314
Decision Tree Classifier	0.440	0.260	0.350
Dummy Classifier	0.001	0.135	0.068
Extra Trees Classifier	0.501	0.183	0.342
Extreme Gradient Boosting	0.416	0.332	0.374
Gradient Boosting Classifier	0.496	0.330	0.413
K Neighbors Classifier	0.535	0.308	0.421
Light Gradient Boosting Machine	0.416	0.373	0.395
Linear Discriminant Analysis	0.541	0.445	0.493
Logistic Regression	0.453	0.280	0.367

Target: SARS-CoV_Mpro			
Model	MCC_cross_validation	MCC_external_validation	average_MCC
Naive Bayes	0.547	0.097	0.322
Quadratic Discriminant Analysis	0.250	0.112	0.181
Random Forest Classifier	0.553	0.150	0.351
Ridge Classifier	0.379	0.332	0.355
SVM - Linear Kernel	0.483	0.445	0.464
Target: SARS-CoV-2_Mpro			
AdaBoost Classifier	0.672	-0.035	0.319
Decision Tree Classifier	0.654	0.403	0.529
Dummy Classifier	0.053	0.028	0.041
Extra Trees Classifier	0.705	0.156	0.431
Extreme Gradient Boosting	0.729	0.297	0.513
Gradient Boosting Classifier	0.741	0.082	0.411
K Neighbors Classifier	0.646	0.113	0.379
Light Gradient Boosting Machine	0.648	0.457	0.552
Linear Discriminant Analysis	0.751	-0.008	0.371
Logistic Regression	0.730	-0.010	0.360
Naive Bayes	0.501	-0.037	0.232
Quadratic Discriminant Analysis	0.131	0.037	0.084
Random Forest Classifier	0.708	0.240	0.474
Ridge Classifier	0.626	0.052	0.339
SVM - Linear Kernel	0.713	0.078	0.395
Target: IAV_Neuraminidase			
AdaBoost Classifier	0.628	0.084	0.356
Decision Tree Classifier	0.694	-0.118	0.288
Dummy Classifier	0.060	-0.013	0.024
Extra Trees Classifier	0.758	0.320	0.539
Extreme Gradient Boosting	0.698	0.103	0.400
Gradient Boosting Classifier	0.753	-0.054	0.350
K Neighbors Classifier	0.677	-0.091	0.293

Table S4 (cont). Average statistics for Matthews Correlation Coefficient (MCC) for each target.

Target: IAV_Neuraminidase			
Model	MCC_cross_validation	MCC_external_validation	average_MCC
Light Gradient Boosting Machine	0.742	-0.114	0.314
Linear Discriminant Analysis	0.752	0.245	0.499
Logistic Regression	0.788	0.286	0.537
Naive Bayes	0.545	0.119	0.332
Quadratic Discriminant Analysis	0.288	-0.074	0.107
Random Forest Classifier	0.762	0.397	0.579
Ridge Classifier	0.714	0.121	0.417
SVM - Linear Kernel	0.760	0.218	0.489
Target: IBV_Neuraminidase			
AdaBoost Classifier	0.641	0.310	0.476
Decision Tree Classifier	0.525	0.124	0.324
Dummy Classifier	0.000	0.000	0.000
Extra Trees Classifier	0.642	0.107	0.374
Extreme Gradient Boosting	0.538	0.124	0.331
Gradient Boosting Classifier	0.609	0.050	0.330
K Neighbors Classifier	0.569	0.210	0.390
Light Gradient Boosting Machine	0.557	0.050	0.304
Linear Discriminant Analysis	0.583	-0.107	0.238
Logistic Regression	0.603	0.097	0.350
Naive Bayes	0.533	-0.072	0.231
Quadratic Discriminant Analysis	0.446	-0.032	0.207
Random Forest Classifier	0.483	0.050	0.266
Ridge Classifier	0.573	0.087	0.330
SVM - Linear Kernel	0.546	-0.156	0.195

Table S4 (cont). Average statistics for Matthews Correlation Coefficient (MCC) for each target.

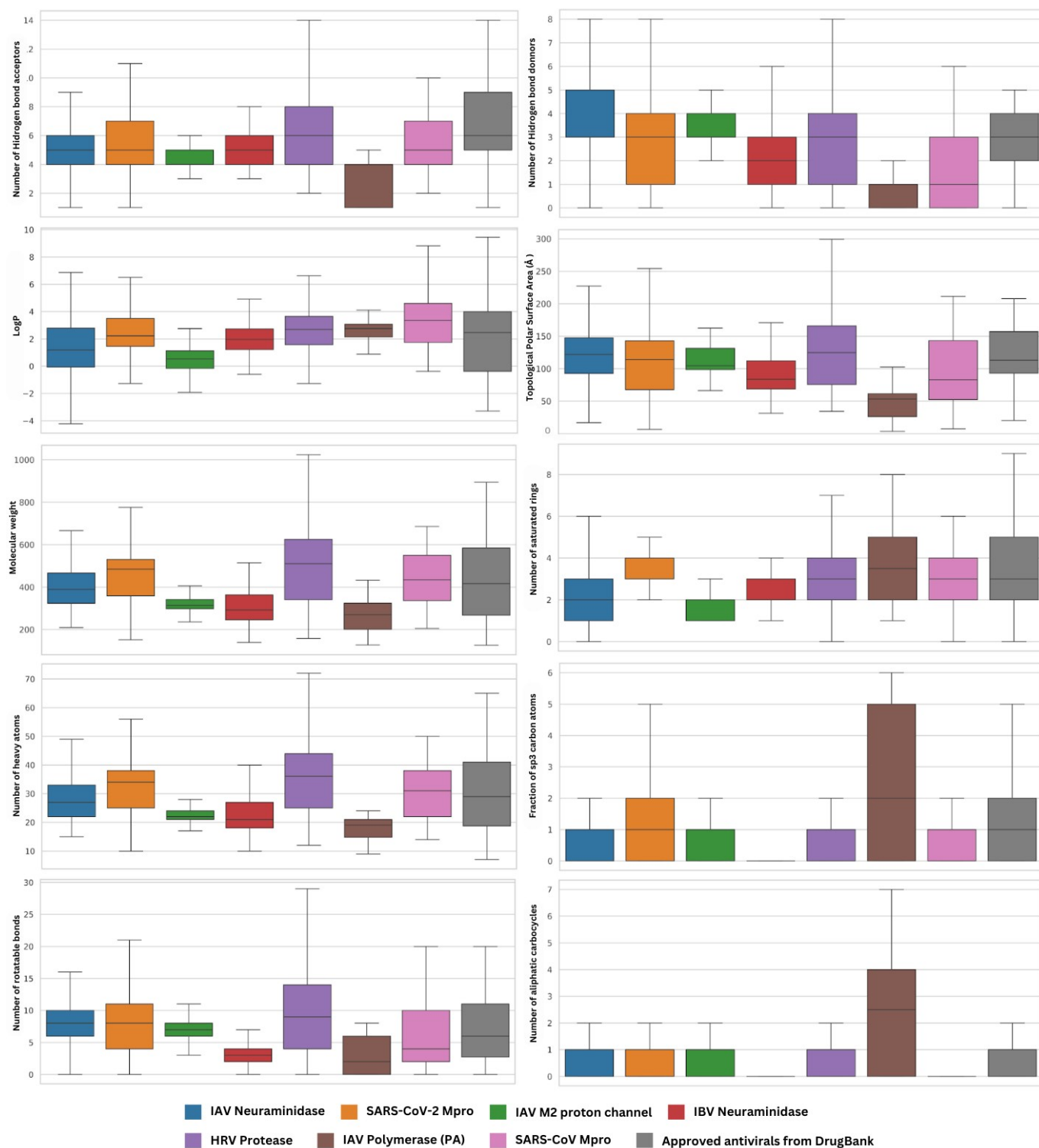


Figure S1. Distribution of physicochemical and constitutional properties computed for antivirals by target compared with approved antivirals from DrugBank.

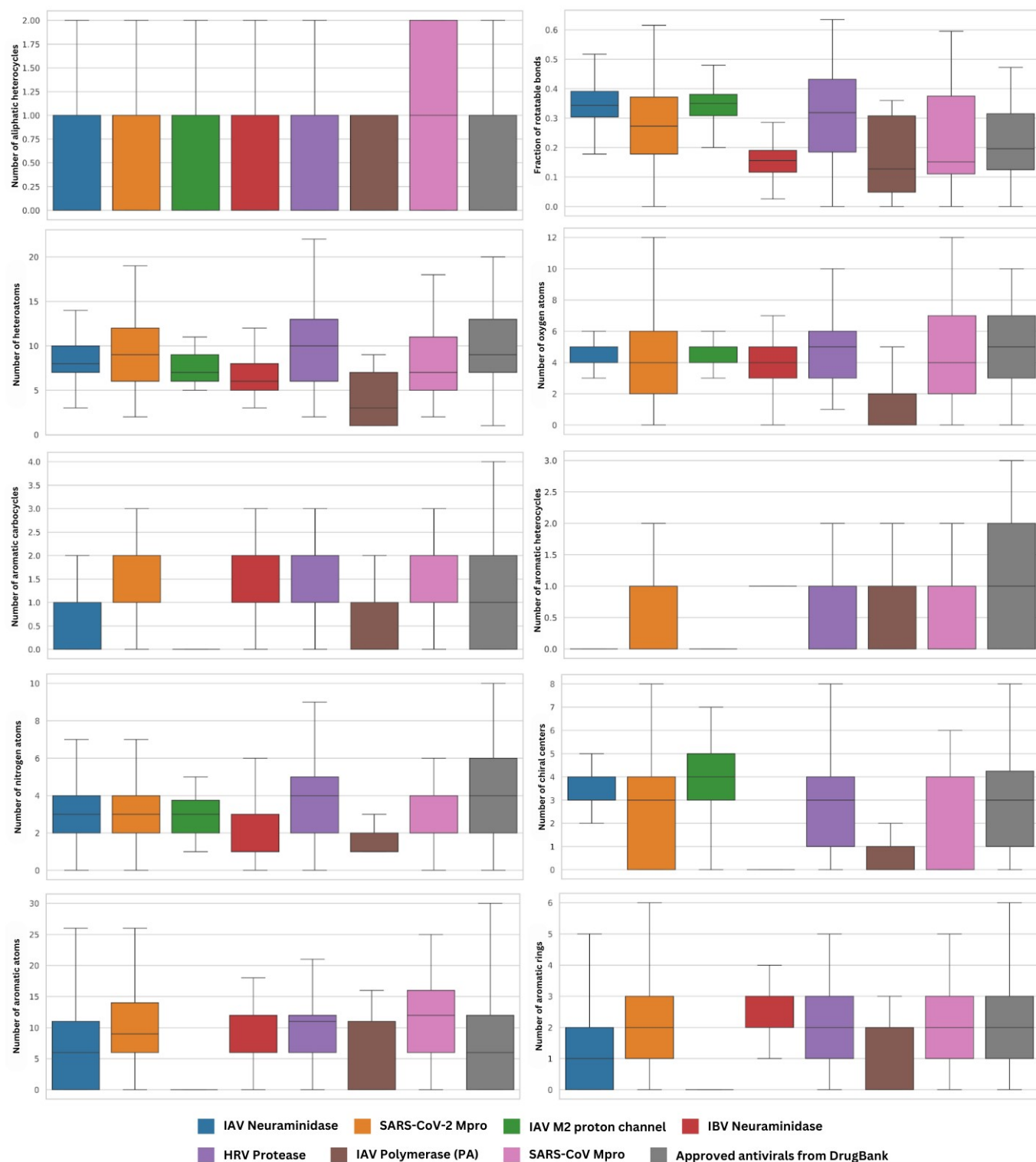


Figure S1 (cont). Distribution of physicochemical and constitutional properties computed for antivirals by target compared with approved antivirals from DrugBank.

Table S5. Descriptive statistics of physicochemical and constitutional descriptors computed for antivirals by target compared with approved antivirals from DrugBank.

Descriptor	Data set	Size	mean	std	min	Q1	50%	Q3	max
Number of acidic atoms	DB	92	0.16	0.68	0.00	0.00	0.00	0.00	3.00
	HRV_Protease	298	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IAV_M2 proton channel	68	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IAV_Neuraminidase	733	0.04	0.26	0.00	0.00	0.00	0.00	2.00
	IAV_Polymerase (PA)	151	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IBV_Neuraminidase	132	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SARS-CoV-2_Mpro	651	0.18	0.71	0.00	0.00	0.00	0.00	3.00
	SARS-CoV_Mpro	77	0.23	0.81	0.00	0.00	0.00	0.00	3.00
Number of aromatic rings	DB	92	1.98	1.35	0.00	1.00	2.00	3.00	6.00
	HRV_Protease	298	2.08	1.21	0.00	1.00	2.00	3.00	10.00
	IAV_M2 proton channel	68	1.00	1.09	0.00	0.00	0.00	2.00	3.00
	IAV_Neuraminidase	733	1.05	1.27	0.00	0.00	1.00	2.00	11.00
	IAV_Polymerase (PA)	151	2.18	0.80	0.00	2.00	2.00	3.00	4.00
	IBV_Neuraminidase	132	0.23	0.48	0.00	0.00	0.00	0.00	2.00
	SARS-CoV-2_Mpro	651	2.06	1.47	0.00	1.00	2.00	3.00	6.00
	SARS-CoV_Mpro	77	2.13	1.44	0.00	1.00	2.00	3.00	7.00
Number of aromatic atoms	DB	92	8.21	7.94	0.00	0.00	6.00	12.00	31.00
	HRV_Protease	298	10.39	6.55	0.00	6.00	11.00	12.00	48.00
	IAV_M2 proton channel	68	5.44	5.90	0.00	0.00	0.00	11.00	16.00
	IAV_Neuraminidase	733	5.81	6.76	0.00	0.00	6.00	11.00	54.00
	IAV_Polymerase (PA)	151	8.49	4.77	0.00	6.00	6.00	12.00	18.00
	IBV_Neuraminidase	132	1.36	2.79	0.00	0.00	0.00	0.00	12.00
	SARS-CoV-2_Mpro	651	10.22	7.55	0.00	6.00	9.00	14.00	35.00
	SARS-CoV_Mpro	77	11.61	7.44	0.00	6.00	12.00	16.00	33.00
Number of basic atoms	DB	92	0.04	0.42	0.00	0.00	0.00	0.00	4.00
	HRV_Protease	298	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IAV_M2 proton channel	68	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IAV_Neuraminidase	733	0.03	0.17	0.00	0.00	0.00	0.00	1.00

	IAV_Polymerase (PA)	151	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IBV_Neuraminidase	132	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SARS-CoV-2_Mpro	651	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SARS-CoV_Mpro	77	0.09	0.29	0.00	0.00	0.00	0.00	1.00
Number of alicyclic rings of carbon	DB	92	0.62	1.16	0.00	0.00	0.00	1.00	5.00
	HRV_Protease	298	0.34	0.67	0.00	0.00	0.00	1.00	4.00
	IAV_M2 proton channel	68	2.13	2.14	0.00	0.00	2.50	4.00	7.00
	IAV_Neuraminidase	733	0.56	0.53	0.00	0.00	1.00	1.00	4.00
	IAV_Polymerase (PA)	151	0.05	0.21	0.00	0.00	0.00	0.00	1.00
	IBV_Neuraminidase	132	0.42	0.55	0.00	0.00	0.00	1.00	2.00
	SARS-CoV-2_Mpro	651	0.47	0.78	0.00	0.00	0.00	1.00	4.00
	SARS-CoV_Mpro	77	0.22	0.98	0.00	0.00	0.00	0.00	5.00
Number of aromatic rings of carbon	DB	92	0.90	1.04	0.00	0.00	1.00	2.00	4.00
	HRV_Protease	298	1.49	1.05	0.00	1.00	1.00	2.00	6.00
	IAV_M2 proton channel	68	0.53	0.61	0.00	0.00	0.00	1.00	2.00
	IAV_Neuraminidase	733	0.74	0.88	0.00	0.00	1.00	1.00	4.00
	IAV_Polymerase (PA)	151	1.23	0.79	0.00	1.00	1.00	2.00	3.00
	IBV_Neuraminidase	132	0.19	0.45	0.00	0.00	0.00	0.00	2.00
	SARS-CoV-2_Mpro	651	1.25	0.92	0.00	1.00	1.00	2.00	4.00
	SARS-CoV_Mpro	77	1.42	1.16	0.00	1.00	1.00	2.00	4.00
Number of chiral centers	DB	92	3.00	2.82	0.00	1.00	3.00	4.25	19.00
	HRV_Protease	298	2.96	2.48	0.00	1.00	3.00	4.00	12.00
	IAV_M2 proton channel	68	1.18	2.43	0.00	0.00	0.00	1.00	12.00
	IAV_Neuraminidase	733	3.24	2.18	0.00	3.00	3.00	4.00	20.00
	IAV_Polymerase (PA)	151	0.28	0.58	0.00	0.00	0.00	0.00	2.00
	IBV_Neuraminidase	132	3.84	1.04	0.00	3.00	4.00	5.00	7.00
	SARS-CoV-2_Mpro	651	2.41	2.14	0.00	0.00	3.00	4.00	11.00
	SARS-CoV_Mpro	77	1.49	2.06	0.00	0.00	0.00	4.00	6.00
CSP3	DB	92	1.28	1.39	0.00	0.00	1.00	2.00	6.00
	HRV_Protease	298	0.61	0.89	0.00	0.00	0.00	1.00	4.00

	IAV_M2 proton channel	68	2.31	2.36	0.00	0.00	2.00	5.00	6.00
	IAV_Neuraminidase	733	0.33	0.52	0.00	0.00	0.00	1.00	2.00
	IAV_Polymerase (PA)	151	0.15	0.38	0.00	0.00	0.00	0.00	2.00
	IBV_Neuraminidase	132	0.61	0.57	0.00	0.00	1.00	1.00	2.00
	SARS-CoV-2_Mpro	651	1.15	1.29	0.00	0.00	1.00	2.00	5.00
	SARS-CoV_Mpro	77	0.62	0.84	0.00	0.00	0.00	1.00	3.00
Fraction of rotatable bonds	DB	92	0.23	0.14	0.00	0.13	0.20	0.31	0.91
	HRV_Protease	298	0.31	0.15	0.00	0.19	0.32	0.43	0.63
	IAV_M2 proton channel	68	0.15	0.13	0.00	0.05	0.13	0.31	0.36
	IAV_Neuraminidase	733	0.34	0.08	0.03	0.31	0.34	0.39	0.65
	IAV_Polymerase (PA)	151	0.16	0.05	0.00	0.12	0.16	0.19	0.29
	IBV_Neuraminidase	132	0.35	0.06	0.20	0.31	0.35	0.38	0.50
	SARS-CoV-2_Mpro	651	0.28	0.12	0.00	0.18	0.27	0.37	0.62
	SARS-CoV_Mpro	77	0.22	0.15	0.00	0.11	0.15	0.38	0.60
Number of halogen atoms	DB	92	0.77	1.32	0.00	0.00	0.00	1.00	5.00
	HRV_Protease	298	0.49	1.03	0.00	0.00	0.00	0.00	5.00
	IAV_M2 proton channel	68	0.32	0.56	0.00	0.00	0.00	1.00	3.00
	IAV_Neuraminidase	733	0.25	0.67	0.00	0.00	0.00	0.00	4.00
	IAV_Polymerase (PA)	151	0.28	0.62	0.00	0.00	0.00	0.00	3.00
	IBV_Neuraminidase	132	0.15	0.57	0.00	0.00	0.00	0.00	3.00
	SARS-CoV-2_Mpro	651	0.74	1.25	0.00	0.00	0.00	1.00	8.00
	SARS-CoV_Mpro	77	0.60	1.17	0.00	0.00	0.00	1.00	6.00
HBA	DB	92	6.73	2.87	1.00	5.00	6.00	9.00	14.00
	HRV_Protease	298	6.25	2.92	2.00	4.00	6.00	8.00	25.00
	IAV_M2 proton channel	68	2.43	1.61	1.00	1.00	1.00	4.00	5.00
	IAV_Neuraminidase	733	5.96	5.25	1.00	4.00	5.00	6.00	70.00
	IAV_Polymerase (PA)	151	4.89	1.43	3.00	4.00	5.00	6.00	11.00
	IBV_Neuraminidase	132	4.36	0.86	3.00	4.00	4.00	5.00	7.00
	SARS-CoV-2_Mpro	651	5.69	2.09	1.00	4.00	5.00	7.00	14.00
	SARS-CoV_Mpro	77	5.44	2.26	2.00	4.00	5.00	7.00	10.00

HBD	DB	92	2.97	1.52	0.00	2.00	3.00	4.00	8.00
	HRV_Protease	298	3.23	3.20	0.00	1.00	3.00	4.00	18.00
	IAV_M2 proton channel	68	0.87	0.69	0.00	0.00	1.00	1.00	3.00
	IAV_Neuraminidase	733	4.06	3.30	0.00	3.00	3.00	5.00	38.00
	IAV_Polymerase (PA)	151	2.18	1.30	0.00	1.00	2.00	3.00	8.00
	IBV_Neuraminidase	132	3.64	0.96	2.00	3.00	3.00	4.00	8.00
	SARS-CoV-2_Mpro	651	2.48	1.68	0.00	1.00	3.00	4.00	8.00
	SARS-CoV_Mpro	77	1.70	2.19	0.00	0.00	1.00	3.00	10.00
Number of heavy atoms	DB	92	31.62	15.77	7.00	18.75	29.00	41.00	80.00
	HRV_Protease	298	36.96	15.86	12.00	25.00	36.00	44.00	141.00
	IAV_M2 proton channel	68	17.53	3.94	9.00	14.75	19.00	21.00	24.00
	IAV_Neuraminidase	733	29.23	16.84	15.00	22.00	27.00	33.00	252.00
	IAV_Polymerase (PA)	151	22.54	6.39	10.00	18.00	21.00	27.00	49.00
	IBV_Neuraminidase	132	23.14	3.60	17.00	21.00	22.00	24.25	35.00
	SARS-CoV-2_Mpro	651	31.92	8.76	10.00	25.00	34.00	38.00	66.00
	SARS-CoV_Mpro	77	30.61	9.26	14.00	22.00	31.00	38.00	50.00
Number of heteroatoms	DB	92	9.84	4.30	1.00	7.00	9.00	13.00	23.00
	HRV_Protease	298	10.05	5.35	2.00	6.00	10.00	13.00	44.00
	IAV_M2 proton channel	68	4.09	2.78	1.00	1.00	3.00	7.00	9.00
	IAV_Neuraminidase	733	9.22	6.97	1.00	7.00	8.00	10.00	96.00
	IAV_Polymerase (PA)	151	6.35	2.03	3.00	5.00	6.00	8.00	13.00
	IBV_Neuraminidase	132	7.36	1.56	5.00	6.00	7.00	9.00	11.00
	SARS-CoV-2_Mpro	651	9.10	3.30	2.00	6.00	9.00	12.00	19.00
	SARS-CoV_Mpro	77	8.45	3.77	2.00	5.00	7.00	11.00	18.00
Number of alicyclic rings with heteroatoms	DB	92	0.96	0.96	0.00	0.00	1.00	1.00	4.00
	HRV_Protease	298	0.56	0.79	0.00	0.00	0.00	1.00	3.00
	IAV_M2 proton channel	68	0.35	0.48	0.00	0.00	0.00	1.00	1.00
	IAV_Neuraminidase	733	0.52	0.81	0.00	0.00	0.00	1.00	6.00
	IAV_Polymerase (PA)	151	0.42	0.63	0.00	0.00	0.00	1.00	3.00
	IBV_Neuraminidase	132	0.68	0.53	0.00	0.00	1.00	1.00	2.00

	SARS-CoV-2_Mpro	651	0.80	0.81	0.00	0.00	1.00	1.00	3.00
	SARS-CoV_Mpro	77	0.95	0.76	0.00	0.00	1.00	2.00	2.00
Number of aromatic rings with heteroatoms	DB	92	1.08	0.82	0.00	0.00	1.00	2.00	3.00
	HRV_Protease	298	0.60	0.76	0.00	0.00	0.00	1.00	4.00
	IAV_M2 proton channel	68	0.47	0.53	0.00	0.00	0.00	1.00	2.00
	IAV_Neuraminidase	733	0.31	0.69	0.00	0.00	0.00	0.00	7.00
	IAV_Polymerase (PA)	151	0.95	0.68	0.00	1.00	1.00	1.00	3.00
	IBV_Neuraminidase	132	0.05	0.21	0.00	0.00	0.00	0.00	1.00
	SARS-CoV-2_Mpro	651	0.81	0.87	0.00	0.00	1.00	1.00	4.00
	SARS-CoV_Mpro	77	0.71	0.92	0.00	0.00	0.00	1.00	3.00
CLogP	DB	92	2.28	3.03	-3.29	-0.38	2.47	4.00	9.45
	HRV_Protease	298	2.61	1.82	-5.41	1.57	2.70	3.65	9.30
	IAV_M2 proton channel	68	2.63	0.86	0.57	2.13	2.76	3.08	4.96
	IAV_Neuraminidase	733	1.21	2.45	-13.97	-0.08	1.15	2.78	7.29
	IAV_Polymerase (PA)	151	2.02	1.17	-0.61	1.21	1.96	2.73	6.18
	IBV_Neuraminidase	132	0.41	1.29	-3.79	-0.20	0.53	1.12	3.48
	SARS-CoV-2_Mpro	651	2.60	1.90	-5.02	1.45	2.23	3.49	8.42
	SARS-CoV_Mpro	77	3.45	2.62	-4.33	1.73	3.35	4.59	10.25
MW	DB	92	451.36	217.69	126.00	267.25	416.90	584.03	1113.20
	HRV_Protease	298	517.80	223.36	158.16	340.46	510.55	625.49	2003.42
	IAV_M2 proton channel	68	262.46	77.58	127.23	200.80	270.40	324.40	432.36
	IAV_Neuraminidase	733	415.59	236.79	209.20	322.41	387.44	465.55	3541.62
	IAV_Polymerase (PA)	151	311.60	89.68	139.15	245.26	291.35	363.39	682.08
	IBV_Neuraminidase	132	328.11	48.24	236.23	298.14	313.25	344.18	479.58
	SARS-CoV-2_Mpro	651	458.68	117.62	151.19	358.81	484.96	530.64	901.12
	SARS-CoV_Mpro	77	445.17	124.51	205.24	336.41	434.45	549.62	685.82
Number of nitrogen atoms	DB	92	3.88	2.28	0.00	2.00	4.00	6.00	10.00
	HRV_Protease	298	3.98	2.80	0.00	2.00	4.00	5.00	19.00
	IAV_M2 proton channel	68	1.69	0.80	1.00	1.00	1.00	2.00	3.00

	IAV_Neuraminidase	733	3.59	3.14	0.00	2.00	3.00	4.00	40.00
	IAV_Polymerase (PA)	151	1.95	1.51	0.00	1.00	1.00	3.00	6.00
	IBV_Neuraminidase	132	2.93	1.13	1.00	2.00	3.00	4.00	5.00
	SARS-CoV-2_Mpro	651	3.42	1.72	0.00	2.00	3.00	4.00	12.00
	SARS-CoV_Mpro	77	2.68	1.54	0.00	2.00	2.00	4.00	6.00
Number of oxygen atoms	DB	92	4.85	2.72	0.00	3.00	5.00	7.00	16.00
	HRV_Protease	298	5.36	2.90	1.00	3.00	5.00	6.00	23.00
	IAV_M2 proton channel	68	1.26	1.45	0.00	0.00	0.00	2.00	5.00
	IAV_Neuraminidase	733	5.18	4.23	1.00	4.00	4.00	5.00	56.00
	IAV_Polymerase (PA)	151	4.06	1.49	0.00	3.00	4.00	5.00	11.00
	IBV_Neuraminidase	132	4.26	1.03	3.00	4.00	4.00	5.00	8.00
	SARS-CoV-2_Mpro	651	4.50	2.49	0.00	2.00	4.00	6.00	12.00
	SARS-CoV_Mpro	77	4.44	2.94	0.00	2.00	4.00	7.00	12.00
Number of ring systems	DB	92	3.55	2.17	0.00	2.00	3.00	5.00	10.00
	HRV_Protease	298	2.98	1.61	0.00	2.00	3.00	4.00	11.00
	IAV_M2 proton channel	68	3.49	1.54	1.00	2.00	3.50	5.00	8.00
	IAV_Neuraminidase	733	2.13	1.64	0.00	1.00	2.00	3.00	17.00
	IAV_Polymerase (PA)	151	2.65	1.05	1.00	2.00	2.00	3.00	6.00
	IBV_Neuraminidase	132	1.34	0.54	1.00	1.00	1.00	2.00	3.00
	SARS-CoV-2_Mpro	651	3.33	1.15	0.00	3.00	3.00	4.00	7.00
	SARS-CoV_Mpro	77	3.30	1.34	0.00	2.00	3.00	4.00	8.00
Number of rotatable bonds	DB	92	6.57	4.91	0.00	2.75	6.00	11.00	26.00
	HRV_Protease	298	10.37	7.92	0.00	4.00	9.00	14.00	46.00
	IAV_M2 proton channel	68	2.85	2.64	0.00	0.00	2.00	6.00	8.00
	IAV_Neuraminidase	733	8.45	7.93	0.00	6.00	8.00	10.00	116.00
	IAV_Polymerase (PA)	151	3.55	1.71	0.00	2.00	3.00	4.00	8.00
	IBV_Neuraminidase	132	6.75	1.78	3.00	6.00	7.00	8.00	14.00
	SARS-CoV-2_Mpro	651	7.55	4.12	0.00	4.00	8.00	11.00	23.00
	SARS-CoV_Mpro	77	6.27	5.09	0.00	2.00	4.00	10.00	20.00
TPSA	DB	92	122.37	48.99	20.23	92.88	113.00	157.21	267.04

	HRV_Protease	298	135.14	83.32	34.14	75.45	124.68	165.92	623.79
	IAV_M2 proton channel	68	46.79	26.61	3.24	26.02	53.11	61.35	102.45
	IAV_Neuraminidase	733	134.32	103.11	17.07	93.45	121.96	147.47	1346.48
	IAV_Polymerase (PA)	151	89.30	30.57	31.20	68.53	83.83	112.00	197.37
	IBV_Neuraminidase	132	116.48	26.78	66.40	98.66	104.46	131.19	200.72
	SARS-CoV-2_Mpro	651	111.11	50.21	6.48	67.83	114.04	142.70	280.80
	SARS-CoV_Mpro	77	98.82	63.67	7.12	52.33	83.11	143.14	303.65

^a std: standard deviation.

^b min: minimum value.

^c Q1: value under which 25% of data points are found in increasing order.

^d Q3: value under which 75% of data points are found in increasing order.

^e max: maximum value.

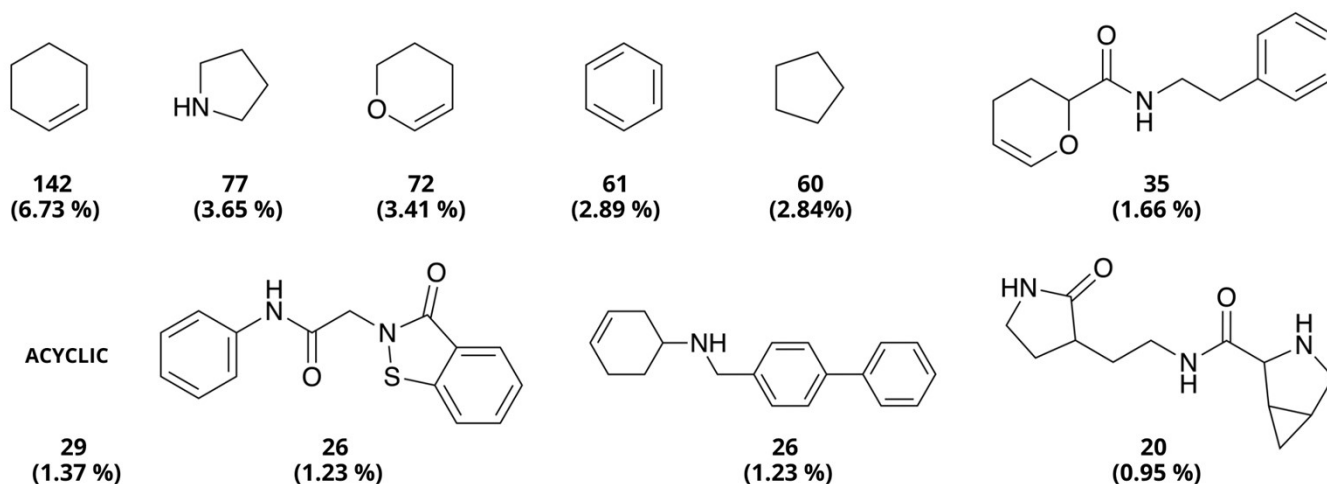


Figure S2. Most frequent scaffolds presented in the data set utilized for constructing the ML models.

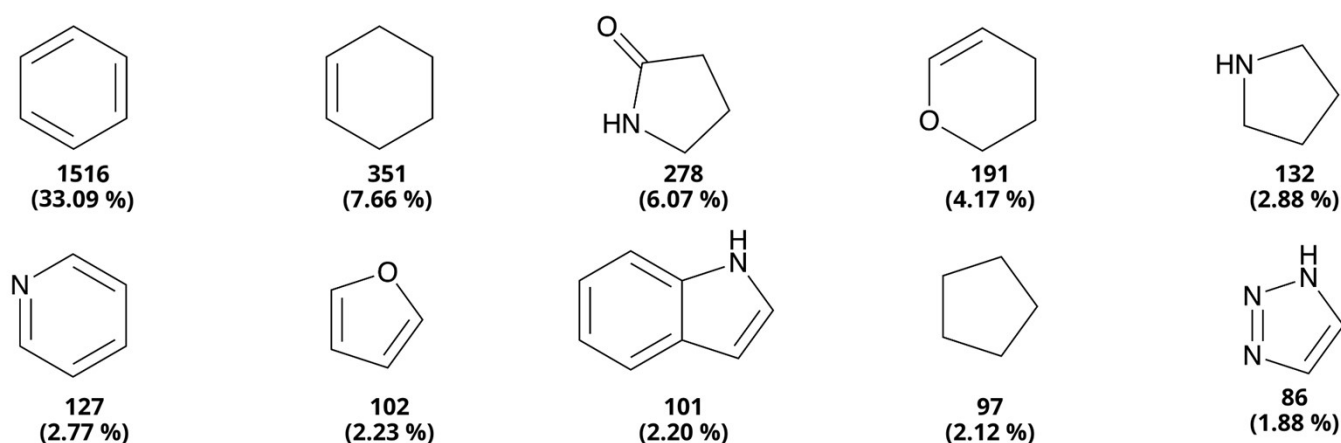


Figure S3. Most frequent ring systems presented in the data set utilized for constructing the ML models.

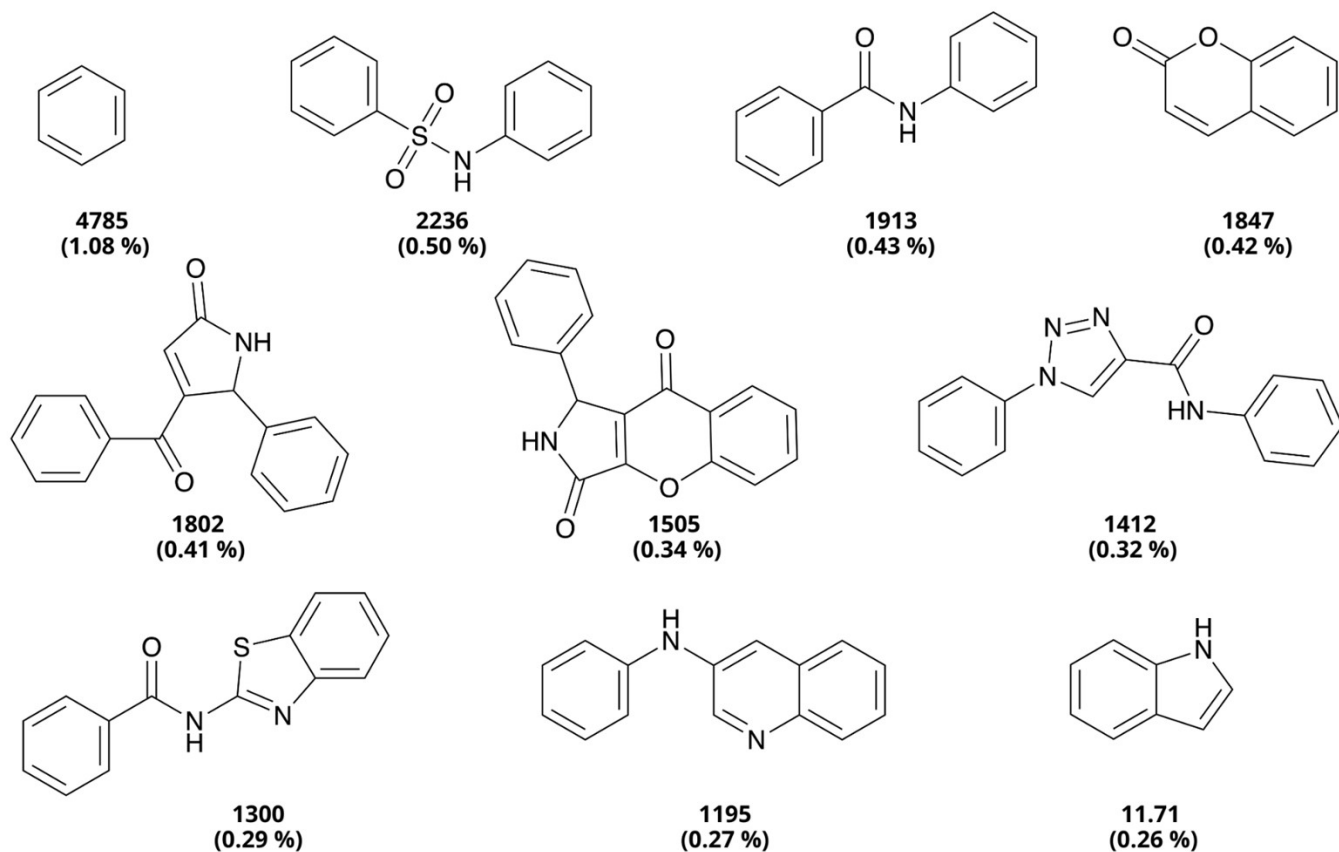


Figure S4. Most frequent scaffolds identified in the newly designed antiviral-focused VS data set library.

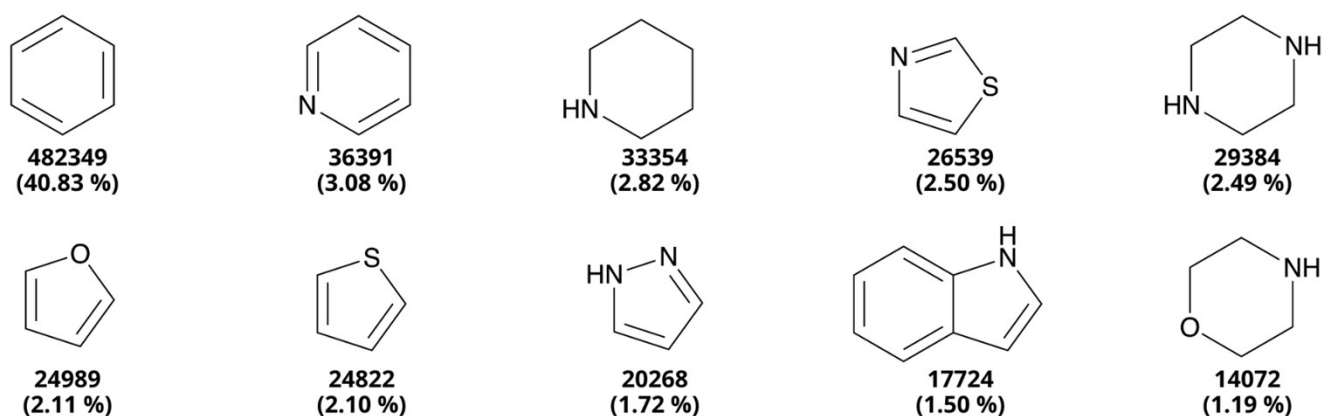


Figure S5. Most frequent ring systems identified in the newly designed antiviral-focused VS data set library.

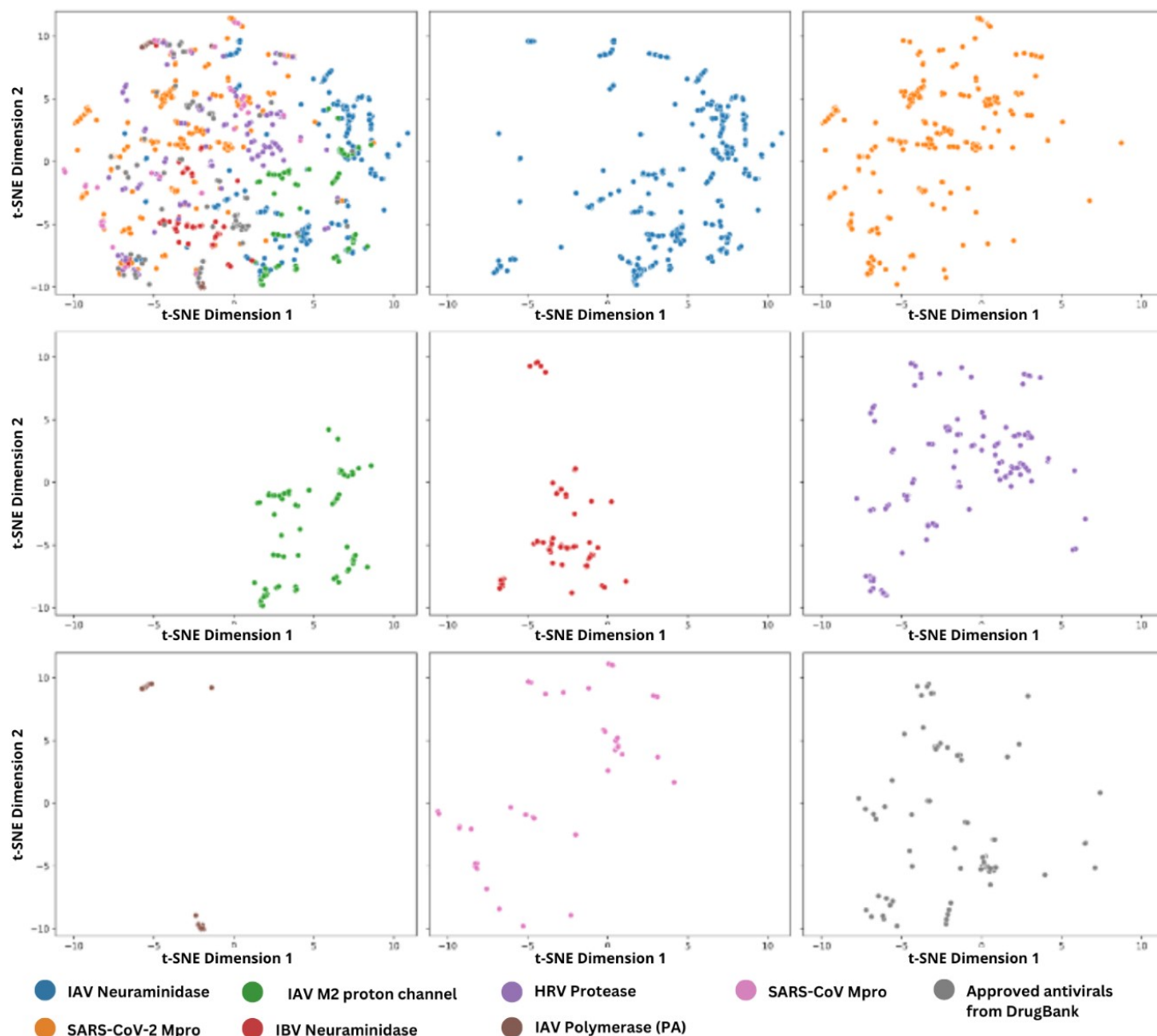


Figure S6. Chemical multiverse visualization of seven antivirals data sets focused on different targets as compared with approved antivirals from DrugBank. The visualization is done with t-SNE using MACCS keys (166-bit) fingerprint. On the upper left are illustrated superimposed data sets, followed by individual data sets using the same coordinates for all of them.

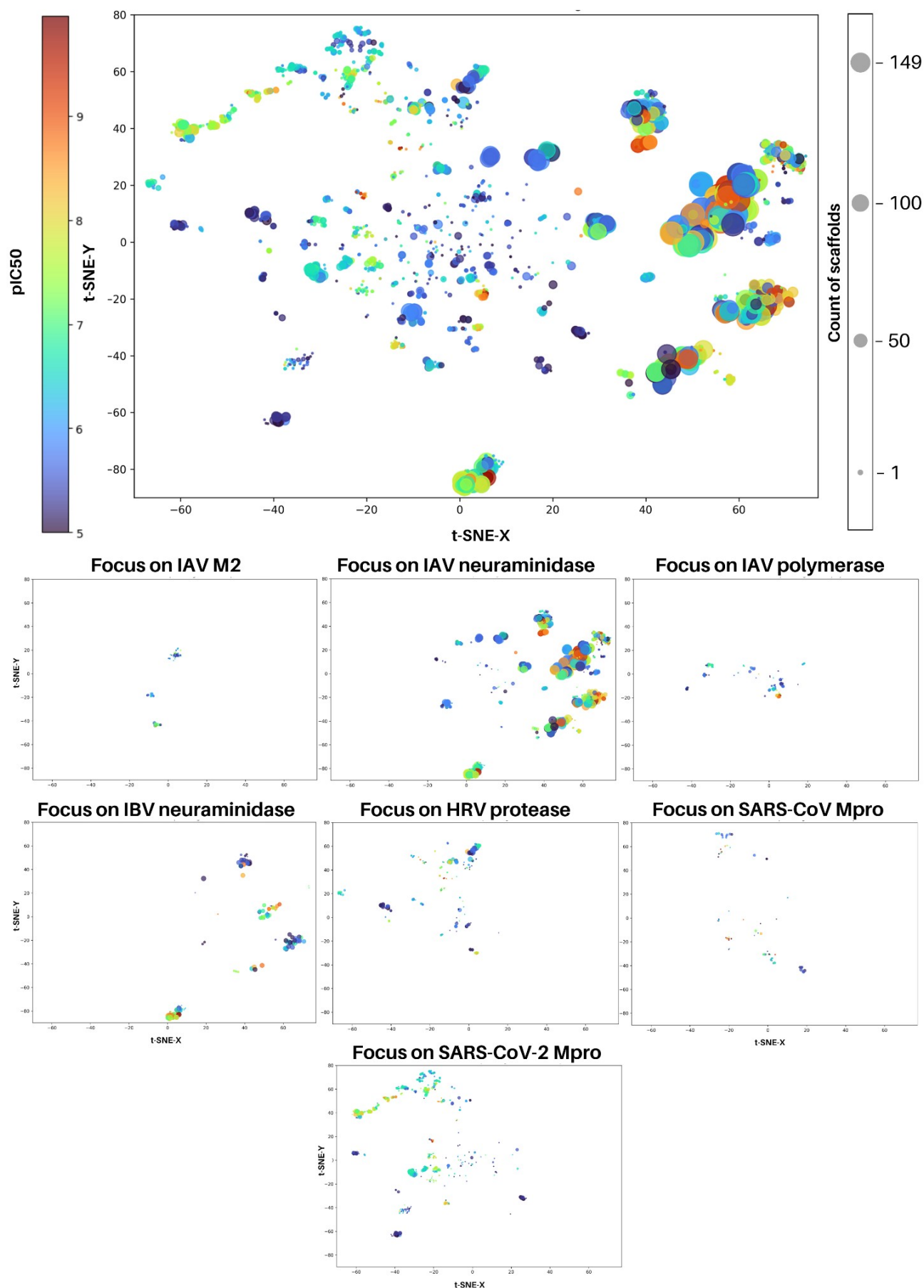


Figure S7. Constellation plot of scaffolds from the data set utilized for constructing the ML models, categorized by target. The x and y axes represent the two dimensions of the t-SNE projection. Each data point corresponds to a unique chemical scaffold, with the size of the point proportional to the number of compounds associated with that scaffold. The color scale indicates pIC_{50} values obtained from ChEMBL, where red represents higher pIC_{50} values and blue represents lower values. This visualization highlights the distribution and activity trends of scaffolds within the chemical data set.

Table S6. MCC values calculated for different validation phases.

		MCC		
		Training		Retraining
Target	Architecture	Cross-validation	External validation	Cross-validation
IAV_Polymerase (PA)	<i>SVM - Linear Kernel</i>	0.81	0.73	0.74
	<i>Extra Trees Classifier</i>	0.71	0.79	0.68
	<i>Gradient Boosting Classifier</i>	0.72	0.68	0.65
	Consensus	-	0.75	-
HRV_Protease	<i>Gradient Boosting Classifier</i>	0.74	0.50	0.68
	<i>Ada Boost Classifier</i>	0.64	0.40	0.44
	<i>Extreme Gradient Boosting</i>	0.68	0.36	0.63
	Consensus	-	0.49	-
IAV_M2 proton channel	<i>Quadratic Discriminant Analysis</i>	0.44	0.00	0.30
	<i>Dummy Classifier</i>	0.34	0.08	0.00
	<i>K Neighbors Classifier</i>	0.20	0.17	0.30
	Consensus	-	-0.07	-
SARS-CoV_Main protease (Mpro)	<i>Linear Discriminant Analysis</i>	0.54	0.44	0.51
	<i>SVM - Linear Kernel</i>	0.48	0.44	0.28
	<i>K Neighbors Classifier</i>	0.53	0.31	0.42
	Consensus	-	0.53	-
SARS-CoV-2_Main protease (Mpro)	<i>Light Gradient Boosting Machine</i>	0.65	0.46	0.64
	<i>Decision Tree Classifier</i>	0.65	0.40	0.55
	<i>Extreme Gradient Boosting</i>	0.73	0.30	0.64
	Consensus	-	0.43	-
IAV_Neuraminidase	<i>Random Forest Classifier</i>	0.76	0.40	0.77
	<i>Extra Trees Classifier</i>	0.76	0.32	0.75
	<i>Logistic Regression</i>	0.79	0.29	0.79

	Consensus	-	0.31	-
IBV_Neuraminidase	<i>Ada Boost Classifier</i>	0.64	0.31	0.82
	<i>K Neighbors Classifier</i>	0.57	0.21	0.46
	<i>Extra Trees Classifier</i>	0.64	0.11	0.40
	Consensus	-	0.21	-

Table S7. Descriptive statistics of Jaccard and Euclidean distances computed for training sets by target.

Distance	Data set	Size	mean	std	min	Q1	Q2	Q3	max
Jaccard (Morgan fingerprints)	HRV_Protease	389	0.84	0.30	0.78	0.82	0.84	0.86	0.93
	IAV_M2 proton channel	92	0.87	0.04	0.79	0.84	0.88	0.90	0.93
	IAV_Neuraminidase	1123	0.82	0.06	0.70	0.77	0.82	0.86	0.94
	IAV_Polymerase (PA)	256	0.84	0.03	0.78	0.82	0.83	0.85	0.92
	IBV_Neuraminidase	202	0.73	0.05	0.63	0.70	0.72	0.76	0.87
	SARS-CoV-2_Mpro	815	0.81	0.06	0.70	0.76	0.80	0.86	0.94
	SARS-CoV_Mpro	197	0.86	0.03	0.81	0.84	0.86	0.89	0.93
Euclidean (Physicochemical properties)	HRV_Protease	389	4.28	1.13	3.19	3.60	4.01	4.64	11.88
	IAV_M2 proton channel	92	2.92	0.49	2.32	2.57	2.72	3.13	4.28
	IAV_Neuraminidase	1106	3.74	1.58	2.79	3.22	3.41	3.74	26.02
	IAV_Polymerase (PA)	256	4.69	1.17	3.40	4.01	4.40	5.04	14.67
	IBV_Neuraminidase	200	4.97	0.98	3.91	4.36	4.68	5.36	10.62
	SARS-CoV-2_Mpro	814	3.95	0.75	2.96	3.51	3.77	4.22	13.44
	SARS-CoV_Mpro	197	4.55	0.97	3.51	3.91	4.27	4.85	8.67

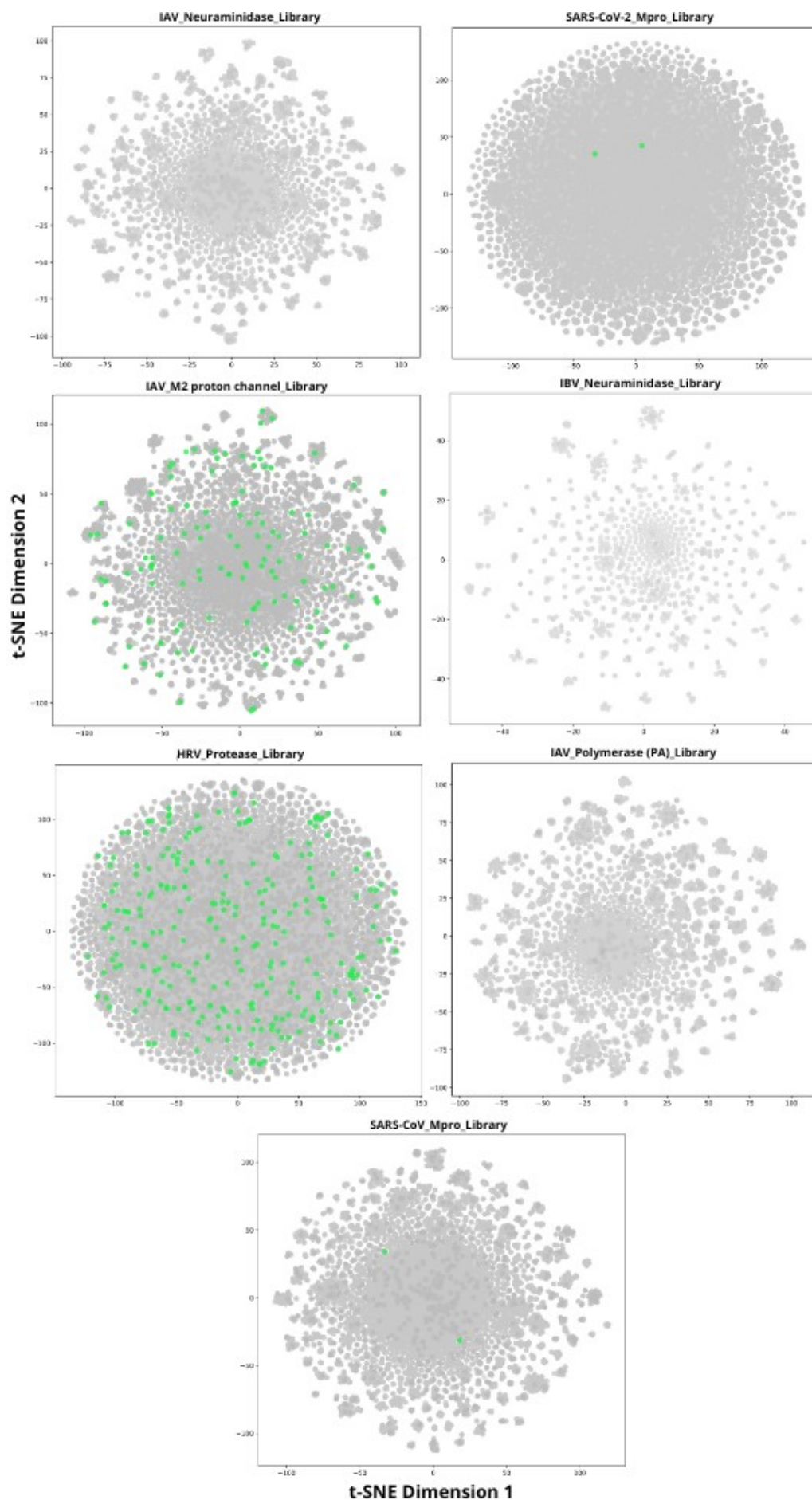


Figure S8. Chemical space visualization of active compounds generated with t-SNE based on the Morgan Chiral of radius 2 (2048-bit) fingerprint for each newly designed library. Green dots represent compounds classified as Q1. Of note, for a few molecular targets there are no compounds within the top quartile (Table 9).

References

1. S. Katz, Overview of viral respiratory infections,
<https://www.msdmanuals.com/professional/infectious-diseases/respiratory-viruses/overview-of-viral-respiratory-infections>, (accessed 21 January 2025).
2. About respiratory Illnesses, About, <https://www.cdc.gov/respiratory-viruses/about/index.html#:~:text=Every%20year%2C%20respiratory%20viruses%20such,fall%20and%20winter%20virus%20season>, (accessed 21 January 2025).