SUPPLEMENTARY INFORMATION for Manuscript

Enhancing molecular property prediction through data integration and consistency assessment

Raquel Parrondo-Pizarro[1,2], Luca Menestrina[1], Ricard Garcia-Serna[1], Adrià Fernández-Torras[1,*], Jordi Mestres[1,2,*]

[1] Chemotargets SL, Parc Cientific de Barcelona, Baldiri Reixac 4 (TR-03), 08028 Barcelona, Catalonia, Spain.
[2] Institut de Quimica Computacional i Catalisi, Facultat de Ciencies, Universitat de Girona, Maria Aurelia Capmany 69, 17003 Girona, Catalonia, Spain.
* Correspondence to: adria.fernandez@chemotargets.com, jordi.mestres@udg.edu

Supplementary Table S1. Description of half-life individual datasets. This table describes the main features of the five half-life datasets considered in the study.

Dataset	*Nº molecules*	*Endpoint Units*	*Characteristics*	*Reference*
Fan et al. (2024)	3512	h	They collected half-life data from the ChEMBL database [doi.org/10.1093/nar/gkad1004] and Arnot's [doi.org/10.1021/es4029414] and Wang's [doi.org/10.1021/acs.jcim.9b00300] studies. A total of 71,264 compounds were initially collected. After preprocessing steps and duplicate removal, a total of 3,512 molecules remained. The base 10 logarithms of their half-life values were calculated and adopted for model training.	doi.org/10.1021/acs.jcim.3c02030
Lombardo et al. (2018)	1352	h	They collected human intravenous pharmacokinetic data on a dataset of 1,352 drugs. All in vivo data were obtained or derived from original references, either through the scientific literature or regulatory agency reports, exclusively from studies utilizing intravenous administration. They greatly expand the already existing large publicly available database they reported earlier (Obach et al., 2008), following the same criteria.	doi.org/10.1124/dmd.118.082966
e-Drug3D: Half-life	1503	h	e-Drug3D is an annotated database based on the 'Drugs@FDA Data File' released by the US Food and Drug Administration (FDA). It currently contains 2,118 molecular structures of small drugs with a molecular weight ≤ 2000 (last update: June 2024). Pharmacokinetic data, including half-life data,	doi.org/10.1093/bioinformatics/bts186

			has been manually extracted from the drug label (additional sources were the review by Obach et al. [doi.org/10.1124/dmd.108.020479], PubMed and the EMA).	
DDPD 1.0: Half-life	1697	h	DDPD 1.0 is a manually curated and standardized database of manually curated and standardized digital properties of approved drugs. Experimental physicochemical properties, pharmacokinetic/toxicokinetic properties and maximum dosages of approved small-molecule drugs from multiple text-based unstructured data resources have been manually assembled, curated, further digitized and processed into structured data deposited in the Database of Digital Properties of approved Drugs (DDPD). DDPD 1.0 contains 30,212 drug property entries, including 2,250 approved drugs and 32 properties. PK data of the small-molecule drugs were manually extracted via text reading from DrugBank [doi.org/10.1093/nar/gkx1037] (supplementary sources were the appendix of *The Pharmacological Basis of Therapeutics* [accessmedicine.mhmedical.com/content.aspx?aid=1154973575] and the human intravenous PK dataset [doi.org/10.1124/dmd.118.082966]).	doi.org/10.1093/database/baab083
Obach et al. (2008) (TDC benchmark)	670	h	They collected human intravenous pharmacokinetic data on a dataset of 670 drugs, the first largest publicly available set of human clinical pharmacokinetic data ever published. Pharmacokinetic data, including terminal phase half-life data, was obtained or derived from original references exclusively from scientific studies utilizing i.v. administration.	doi.org/10.1124/dmd.108.020479

Supplementary Table S2. Description of clearance individual datasets. This table describes the main features of the seven clearance datasets considered in the study.

Dataset	*Measured Endpoint*	*Nº original molecules*	*Nº preprocessed molecules*	*Endpoint Units*	*Characteristics*	*Reference*
Lombardo et al. (2018)	Human plasma clearance	1352	1352	mL/min/kg	They collected human intravenous pharmacokinetic data on a dataset of 1,352 drugs. All in vivo data were obtained or derived from original references, either through the scientific literature or regulatory agency reports, exclusively from studies utilizing intravenous administration. They greatly expand the already existing large publicly available database they reported earlier (Obach et al., 2008), following the same criteria.	doi.org/10.1124/dmd.118.082966
TDC (AstraZeneca)	Human intrinsic clearance (from microsomes and hepatocytes)	1102	744	uL/min/10^6 cells	This database contains experimental in vitro DMPK and physicochemical data, including human intrinsic clearance data derived from assays performed on human liver microsomes [pubchem.ncbi.nlm.nih.gov/bioassay/1159394] and hepatocytes [pubchem.ncbi.nlm.nih.gov/bioassay/1159396], determined at Astrazeneca on a set of publicly disclosed compounds.	doi.org/10.6019/CHEMBL3301361
		1245	289	mL/min/g (uL/min/mg)		
Iwata et al. (2022)	Human total body clearance	741	741	mL/min/kg	They obtained 741 sets of human total body clearance data from JCP2013 [doi.org/10.1177/0091270012440282] and ChEMBL23 [doi.org/10.1093/nar/gkr777]. The gathered data was preprocessed by applying logarithmic transformation and	doi.org/10.1021/acs.jcim.2c00318

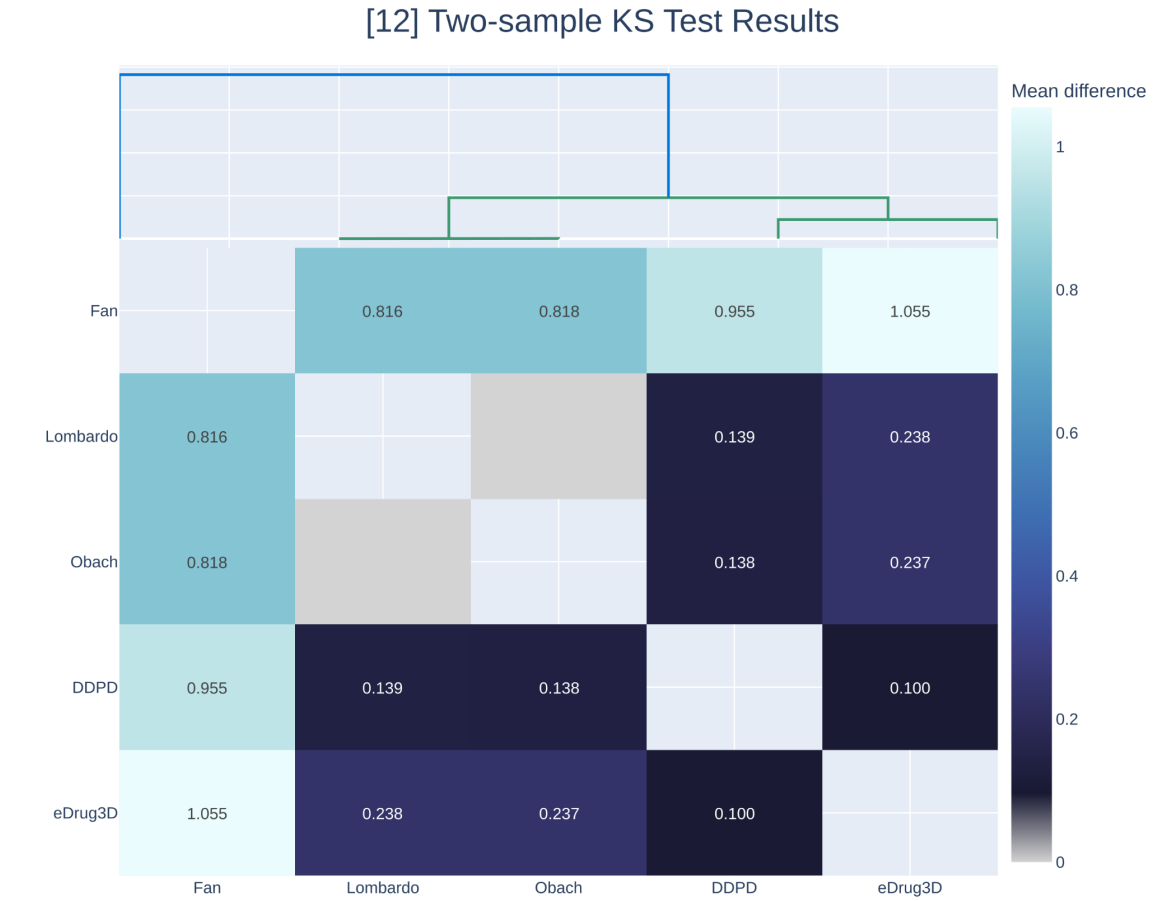
					standardization before applying ML modeling.	
Obach et al. (2008)	Human plasma clearance	670	670	mL/min/kg	They collected human intravenous pharmacokinetic data on a dataset of 670 drugs, the first largest publicly available set of human clinical pharmacokinetic data ever published. Pharmacokinetic data, including plasma clearance data, was obtained or derived from original references exclusively from scientific studies utilizing i.v. administration.	doi.org/10.1124/dmd.108.020479
Gombar and Hall (2013)	Systemic clearance	525	525	mL/min/kg	The training set was derived from the human intravenous clearance data compiled by Obach et al. [doi.org/10.1124/dmd.108.020479], in many instances through digitization of the published time−concentration plots. The available data were further cleaned and harmonized, leading to a final dataset of 525 compounds. The data were log transformed before modeling.	doi.org/10.1021/ci400001u
Varma et al. (2009)	Total body clearance	391	391	mL/min/kg	A database of 391 compounds was developed after carefully and exhaustively mining the scientific literature for relevant human pharmacokinetic parameters. The human total body clearance was taken from the literature	doi.org/10.1021/jm900403j

					compilation reported by Obach et al. [doi.org/10.1124/dmd.108.020479].	
Varma et al. (2010)	Total body clearance	309	309	mL/min/kg	Pharmacokinetic literature was extensively mined to identify a human data set of 309 compounds. Intravenous clearance data was derived from the scientific literature compilation released by Obach et al. [doi.org/10.1124/dmd.108.020479].	doi.org/10.1021/jm901371v

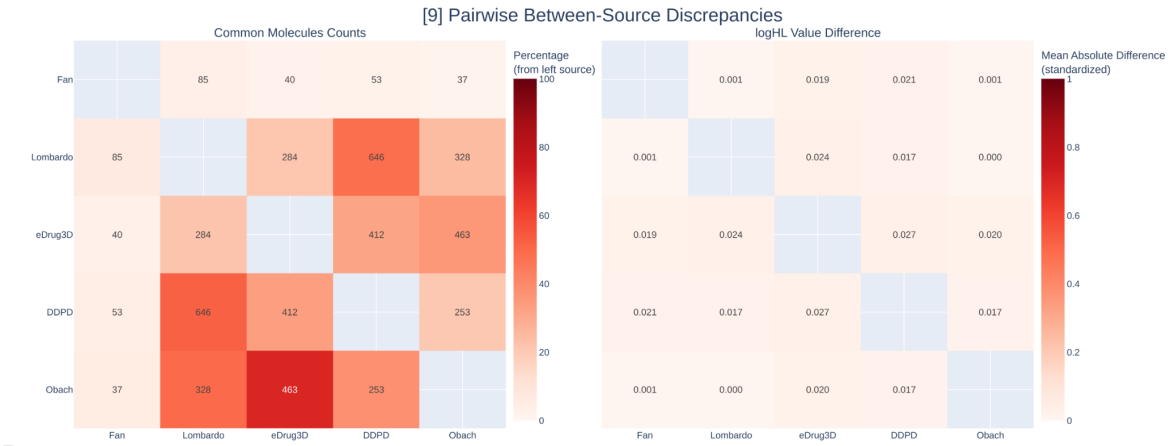
Supplementary Table S3. Hyperparameter search space for the ML algorithms.

ML Model	*Hyperparameter*	*Values*
XGBoost Regressor	Number of estimators	[50, 100, 200, 500]
	Maximum depth	[3, 5, 8, 12, 15, 20]
	Learning rate	[0.001, 0.01, 0.05, 0.1, 0.2]
	Gamma	[0, 0.1, 0.25, 0.5, 1]
	Minimum child weight	[1, 3, 5]
	Subsample	[0.25, 0.5, 0.7, 1.0]
	Column sample by tree	[0.25, 0.5, 0.7, 1.0]
	L1 regularization (alpha)	[0.00001, 0.0001, 0.001, 0.01, 0.1, 1]
	L2 regularization (lambda)	[0.00001, 0.0001, 0.001, 0.01, 0.1, 1]
Support Vector Regressor	Kernel	['linear', 'rbf']
	Gamma	['auto', 0.001, 0.01, 0.1]
	C	[0.1, 1, 10, 100, 1000]
	Epsilon	[0.01, 0.05, 0.1, 0.2]
Random Forest Regressor	Number of estimators	(10, 250)
	Maximum depth	(1, 50)
	Maximum features	(0.1, 0.999)
	Minimum samples per split	(2, 25)
K-Nearest Neighbors Regressor	Number of neighbors	(5, 100)
	Metric	['minkowski', 'euclidean', 'manhattan']
	Weights	[‘uniform’, ‘distance’]

Supplementary Figure S1. Additional analysis of the half-life dataset using DataReporting. (A) Heatmap displaying the pairwise results of the two-sample Kolmogorov–Smirnov (KS) test. Mean differences are shown for statistically significant pairwise comparisons. A side dendrogram illustrates the clustering of the datasets. (B) Combined plot integrating two complementary pairwise heatmaps: the left heatmap shows discrepancies between data sources based on the proportion of common molecules, while the right one depicts differences in property values for these shared compounds. In both the upper and lower triangular matrices, the proportion of common molecules and standardized difference values are calculated using the data source on the ordinate as the reference.

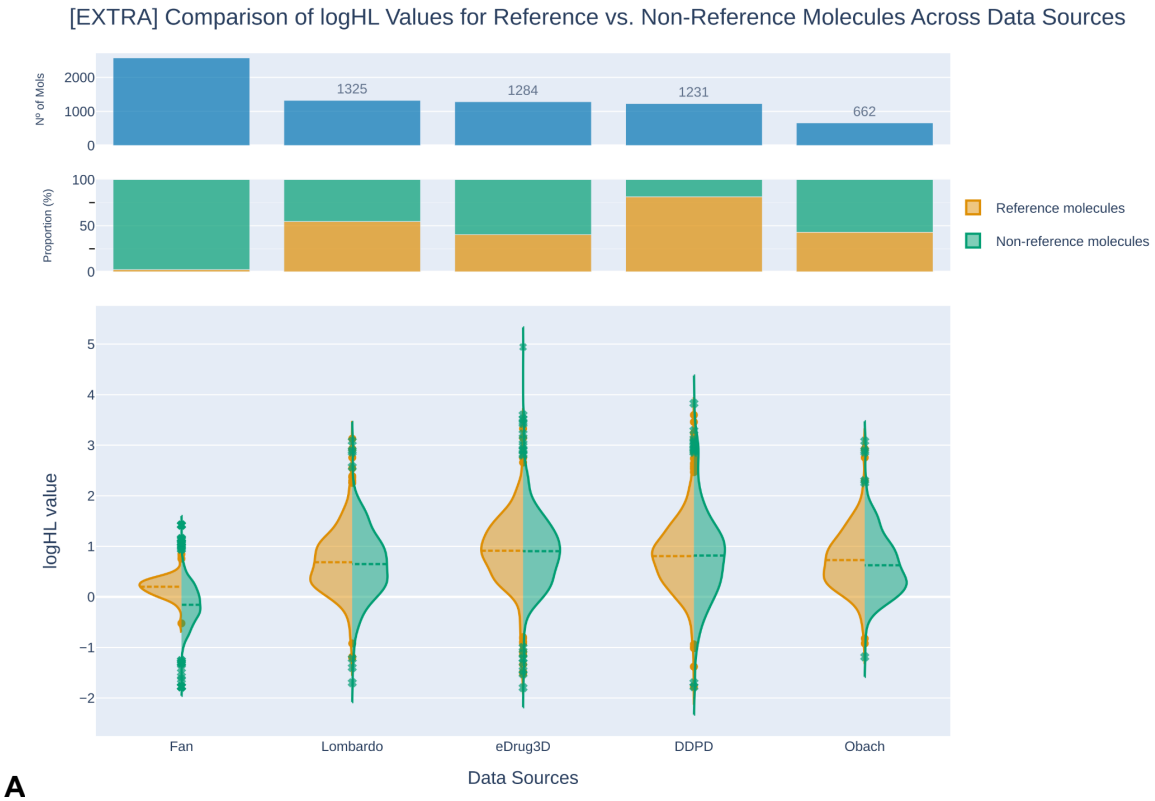


A

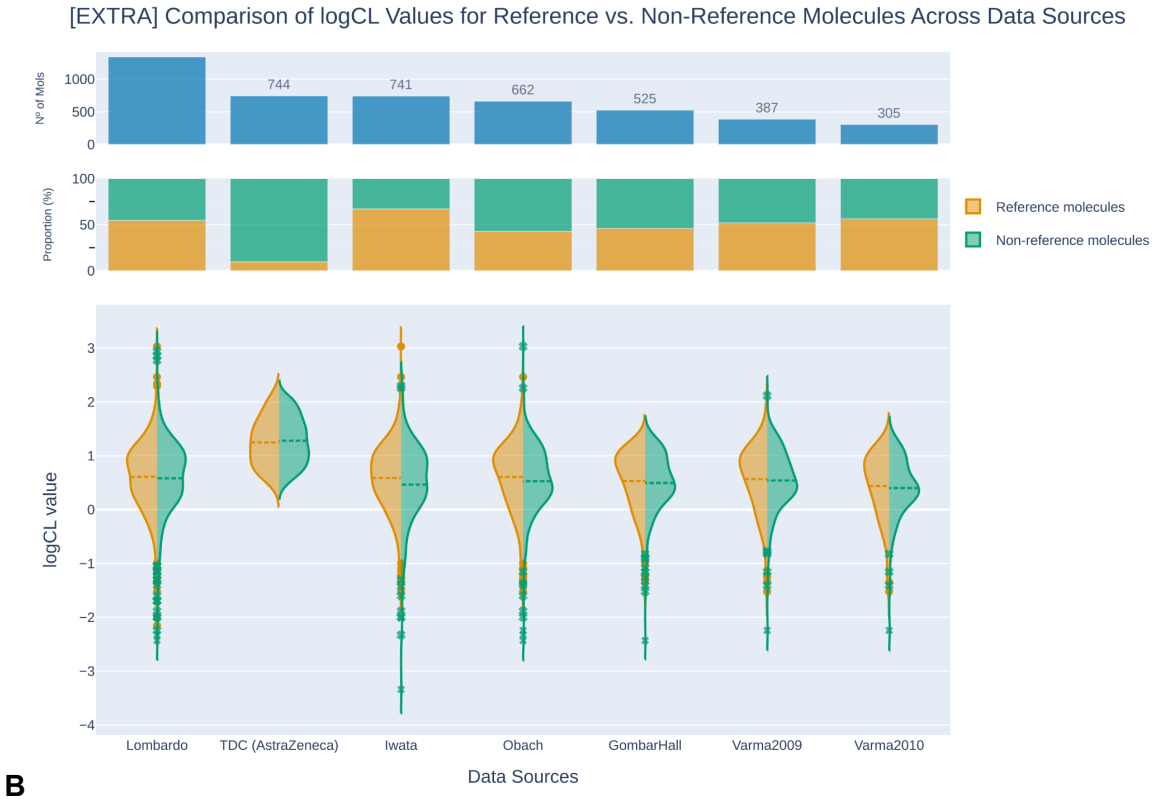


B

Supplementary Figure S2. AssayInspector analysis of drug vs. non-drug molecules for both (A) half-life and (B) clearance datasets. Violin plots showing the distribution of property values for drug (yellow) and non-drug (green) molecules across different data sources. Two additional bar charts at the top display the number of molecules and the proportion of compounds included in the reference set for each dataset.

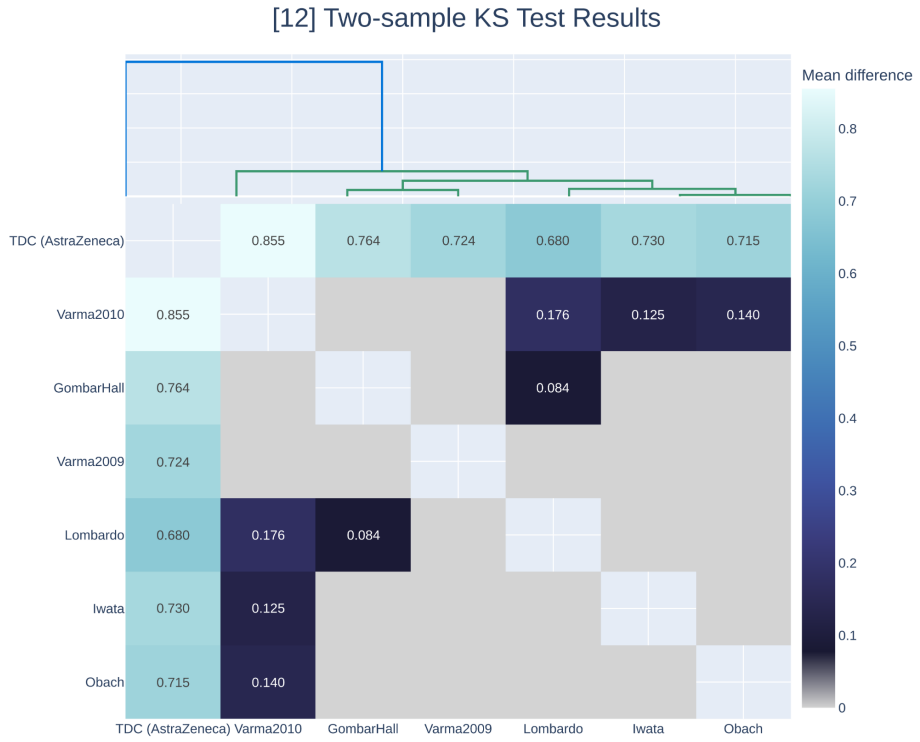


A

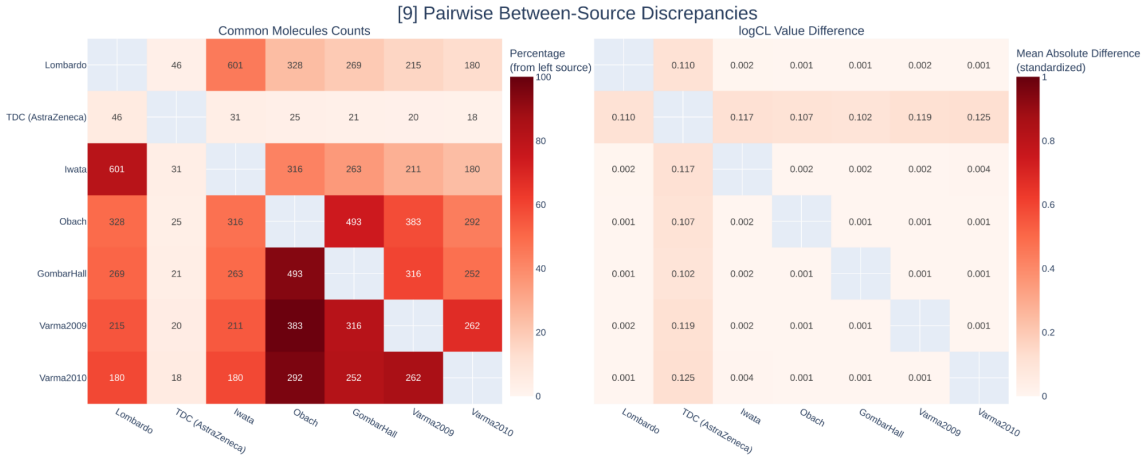


B

Supplementary Figure S3. Additional analysis of the clearance dataset using DataReporting. (A) Heatmap displaying the pairwise results of the two-sample Kolmogorov–Smirnov (KS) test. Mean differences are shown for statistically significant pairwise comparisons. A side dendrogram illustrates the clustering of the datasets. (B) Combined plot integrating two complementary pairwise heatmaps: the left heatmap shows discrepancies between data sources based on the proportion of common molecules, while the right one depicts differences in property values for these shared compounds. In both the upper and lower triangular matrices, the proportion of common molecules and standardized difference values are calculated using the data source on the ordinate (y-axis) as reference. Note that while counts are symmetrical, the color gradients reflect the percentage of shared molecules relative to the reference dataset.



A



B

Supplementary Table S4. Additional performance evaluation metrics of ML models for half-life prediction. This table includes the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Spearman correlation for each model trained on different aggregated datasets and evaluated on different sources across folds (median and median absolute deviation).

Training dataset	*Fan et al.*			*Obach et al. (TDC benchmark)*			*Lombardo et al.*			*e-Drug3D: Half-life*			*DDPD 1.0: Half-life*		
	MAE	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*
All sources	0.24 ± 0.01	0.32 ± 0.01	0.73 ± 0.02	0.35 ± 0.02	0.48 ± 0.02	0.60 ± 0.03	0.42 ± 0.01	0.58 ± 0.02	0.48 ± 0.04	0.51 ± 0.03	0.69 ± 0.07	0.44 ± 0.02	0.51 ± 0.01	0.70 ± 0.05	0.41 ± 0.02
Benchmark source (Obach et al.)	0.94 ± 0.02	1.07 ± 0.01	-0.20 ± 0.04	0.35 ± 0.01	0.48 ± 0.02	0.60 ± 0.03	0.42 ± 0.01	0.56 ± 0.01	0.52 ± 0.02	0.44 ± 0.01	0.59 ± 0.03	0.53 ± 0.02	0.46 ± 0.02	0.62 ± 0.02	0.47 ± 0.03
Divergent source (Fan et al.)	0.16 ± 0.00	0.20 ± 0.00	0.86 ± 0.01	0.71 ± 0.02	0.96 ± 0.05	-0.02 ± 0.04	0.80 ± 0.02	1.02 ± 0.02	-0.14 ± 0.03	1.00 ± 0.04	1.21 ± 0.07	-0.20 ± 0.04	0.90 ± 0.07	1.13 ± 0.10	-0.16 ± 0.04
Homogeneous sources (all but Fan et al.)	1.14 ± 0.02	1.29 ± 0.03	-0.24 ± 0.02	0.33 ± 0.01	0.46 ± 0.03	0.67 ± 0.02	0.38 ± 0.01	0.51 ± 0.02	0.62 ± 0.01	0.39 ± 0.02	0.53± 0.03	0.64 ± 0.02	0.42 ± 0.01	0.58 ± 0.03	0.58 ± 0.03

Supplementary Table S5. Additional performance evaluation metrics of ML models for clearance prediction. This table includes the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Spearman correlation for each model trained on different aggregated datasets and evaluated on different sources across folds (median and median absolute deviation).

Training dataset	*TDC (Divergent source)*			*Lombardo et al.*			*Iwata et al.*			*Obach et al.*			*Gombar and Hall*			*Varma et al. (2009)*			*Varma et al. (2010)*		
	MAE	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*
All sources	0.36 ± 0.02	0.47 ± 0.03	0.39 ± 0.08	0.40 ± 0.01	0.54 ± 0.02	0.54 ± 0.02	0.39 ± 0.03	0.53 ± 0.04	0.59 ± 0.03	0.37 ± 0.02	0.51 ± 0.02	0.63 ± 0.02	0.33 ± 0.01	0.44 ± 0.01	0.61 ± 0.04	0.35 ± 0.02	0.48 ± 0.01	0.62 ± 0.03	0.37 ± 0.03	0.50 ± 0.04	0.62 ± 0.08
Benchmark source (Obach et al.)	0.79 ± 0.02	0.93 ± 0.02	0.14 ± 0.03	0.44 ± 0.02	0.60 ± 0.02	0.47 ± 0.03	0.40 ± 0.01	0.53 ± 0.03	0.56 ± 0.05	0.38 ± 0.02	0.51 ± 0.06	0.57 ± 0.03	0.34 ± 0.03	0.49 ± 0.05	0.56 ± 0.03	0.35 ± 0.01	0.48 ± 0.03	0.60 ± 0.02	0.36 ± 0.01	0.49 ± 0.02	0.57 ± 0.03
Divergent source (TDC)	0.34 ± 0.01	0.42 ± 0.02	0.45 ± 0.06	067 ± 0.03	0.87 ± 0.05	0.12 ± 0.04	0.69 ± 0.06	0.89 ± 0.11	0.17 ± 0.06	0.66 ± 0.04	0.88 ± 0.06	0.18 ± 0.06	0.63 ± 0.06	0.82 ± 0.05	0.26 ± 0.07	0.62 ± 0.02	0.78 ± 0.03	0.18 ± 0.11	0.72 ± 0.09	0.89 ± 0.11	0.20 ± 0.05
Homogeneous sources (all but TDC)	0.74 ± 0.03	0.88 ± 0.04	0.20 ± 0.03	0.41 ± 0.02	0.57 ± 0.04	0.57 ± 0.04	0.38 ± 0.02	0.52 ± 0.03	0.62 ± 0.04	0.37 ± 0.02	0.50 ± 0.04	0.63 ± 0.04	0.34 ± 0.03	0.47 ± 0.06	0.60 ± 0.04	0.34 ± 0.02	0.46 ± 0.02	0.64 ± 0.03	0.35 ± 0.01	0.47 ± 0.02	0.59 ± 0.03

Supplementary Table S6. Additional performance evaluation metrics of ML models for half-life prediction applying prior scaling. This table includes the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Spearman correlation for each model trained on different datasets derived from the aggregation of specific sources across folds (median and median absolute deviation).

Training datasets	*Homogenous sources*			*Divergent source (Fan et al.)*		
	MAE	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*
Homogenous sources	0.41 ± 0.01	0.57 ± 0.02	0.63 ± 0.03	1.19 ± 0.02	1.32 ± 0.02	-0.24 ± 0.05
All sources	0.44 ± 0.01	0.61 ± 0.02	0.54 ± 0.02	0.18 ± 0.00	0.24 ± 0.01	0.80 ± 0.02

Supplementary Table S7. Additional performance evaluation metrics of ML models for clearance prediction applying prior scaling. This table includes the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Spearman correlation for each model trained on different datasets derived from the aggregation of specific sources across folds (median and median absolute deviation).

Training datasets	*Homogenous sources*			*Divergent source (TDC)*		
	MAE	*RMSE*	*Spearman*	*MAE*	*RMSE*	*Spearman*
Homogenous sources	0.39 ± 0.02	0.54 ± 0.03	0.59 ± 0.03	0.72 ± 0.03	0.84 ± 0.03	0.19 ± 0.04
All sources	0.39 ± 0.01	0.55 ± 0.03	0.62 ± 0.03	0.31 ± 0.01	0.39 ± 0.01	0.47 ± 0.03