Supplementary information for "Data-driven equation for drug–membrane permeability across drugs and membranes"

Arghya Dutta,[1, a)] Jilles Vreeken,[2] Luca Ghiringhelli,[3] and Tristan Bereau[1, 4]

[1)] *Max Planck Institute for Polymer Research, Mainz, Germany*

[2)] *CISPA Helmholtz Center for Information Security, Saarbrücken, Germany*

[3)] *NOMAD Laboratory, Fritz Haber Institute of the Max Planck Society, Berlin, Germany*

[4)] *Van 't Hoff Institute for Molecular Sciences and Informatics Institute, University of Amsterdam, Amsterdam 1098 XH, The Netherlands*

(Dated: 3 June 2021)

S1. DEFINITION OF THE IONIZATION CONSTANTS AND PK_a

Menichetti *et. al.*[1] followed the convention of ChemAxon[2] while defining apK_a and bpK_a. As it is a bit different from the usual of acidic and basic pK_as, here we provide the details. The ionization constant K_a and pK_a are defined as

$$\mathrm{K}_a = \frac{[\text{conjugate base}] \times [\mathrm{H}^+]}{[\text{conjugate acid}]}, \tag{1}$$

$$\mathrm{pK}_a = -\log{}_{10}\mathrm{K}_a = \mathrm{pH} + \log{}_{10}\frac{[\text{conjugate acid}]}{[\text{conjugate base}]}, \tag{2}$$

where $\mathrm{pH} = -\log{}_{10}[\mathrm{H}^+]$. In their simulations, Menichetti *et. al.* always started from a neutral compound which, depending on the pH and pK_a, can either protonate or deprotonate. We consider these two cases separately as follows.

S1.1. Deprotonation

A charge-neutral acid AH can deprotonate to release a proton and a charged conjugate base A^- by the following reaction

$$\mathrm{AH} \rightleftharpoons \mathrm{A}^- + \mathrm{H}^+.$$

The corresponding acidic pK_a, denoted as apK_a, is defined as

$$\mathrm{K}_a = \frac{[\mathrm{A}^-][\mathrm{H}^+]}{[\mathrm{AH}]}, \tag{3}$$

$$\mathrm{ap}K_\mathrm{a} = -\log_{10}\mathrm{K}_a = \mathrm{pH} + \log{}_{10}\frac{[\mathrm{AH}]}{[\mathrm{A}^-]}. \tag{4}$$

S1.2. Protonation

A charge-neutral base B can protonate and becomes a charged conjugate acid $[\mathrm{BH}^+]$ by the following reaction

$$\mathrm{BH}^+ \rightleftharpoons \mathrm{B} + \mathrm{H}^+.$$

[a)] Electronic mail: dutta@mpip-mainz.mpg.de

To comply with the unified definition of pK_a of ChemAxon—it is the ratio of conjugate acid to conjugate base—the corresponding basic pK_a, denoted as bpK_a, is defined as

$$\mathrm{K}_a = \frac{[\mathrm{B}][\mathrm{H}^+]}{[\mathrm{BH}^+]} \tag{5}$$

$$\mathrm{bp}K_\mathrm{a} = -\log_{10}\mathrm{K}_a = \mathrm{pH} + \log_{10}\frac{[\mathrm{BH}^+]}{[\mathrm{B}]} \tag{6}$$

The usefulness of this definition is that now both apK_a and bpK_a are written as $\mathrm{pK}_a = \mathrm{pH} + \log_{10}\frac{[\text{conjugate acid}]}{[\text{conjugate base}]}$.

Strong acids, as defined in the Acid–base asymptotes section of the paper, have low (ap$K_\mathrm{a} \leq 4$). At pH = 7, from Eq. 4 we find that

$$\frac{[\mathrm{AH}]}{[\mathrm{A}^-]} = 10^{\mathrm{ap}K_\mathrm{a}-7} \leq 10^{-3}. \tag{7}$$

So, with *decreasing* apK_a, the acid's concentration [AH] will keep decreasing and the conjugate base's concentration [A−] will keep increasing, as expected. Conversely, strong bases have high (bp$K_\mathrm{a} \geq 10$). At pH = 7, from Eq. 6 we get

$$\frac{[\mathrm{BH}^+]}{[\mathrm{B}]} = 10^{\mathrm{bp}K_\mathrm{a}-7} \geq 10^3. \tag{8}$$

So, with *increasing* bpK_a, the base's concentration [B] will keep decreasing and the conjugate acid's concentration [BH$^+$] will keep increasing, again as expected.

S2. DISTRIBUTION OF COMPOUNDS ACROSS PERMEABILITY

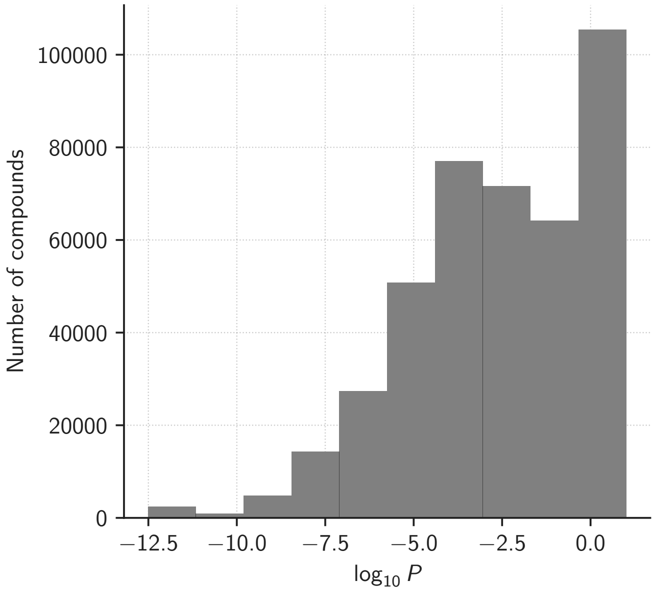

FIG. S1: Distribution of the small molecules considered in this work across the range of permeability values.

S3. TABLE 1 WITH ERROR VALUES

TABLE S1: Table I from the main text along with the standard errors shown in parentheses.

Model			c_0	c_1	c_2	c_3	RMSE	MaxAE	r^2
f^{Hyd}	$= c_0^{\mathrm{Hyd}}$	$+c_1^{\mathrm{Hyd}} \beta \Delta G_{\mathrm{W \to Ol}}$	-3.444	-0.648			1.53	11.82	0.64
			(± 0.004)	(± 0.001)			(± 0.00)	(± 0.00)	(± 0.00)
f^{1D}	$= c_0^{\mathrm{1D}}$	$+c_1^{\mathrm{1D}} (\mathrm{ap}K_{\mathrm{a}} - \mathrm{bp}K_{\mathrm{a}} - 2\beta \Delta G_{\mathrm{W \to Ol}})$	-5.419	0.163			1.40	6.35	0.70
			(± 0.007)	(± 0.000)			(± 0.00)	(± 0.02)	(± 0.00)
f^{2D}	$= c_0^{\mathrm{2D}}$	$+c_1^{\mathrm{2D}} (\sqrt[3]{\beta \Delta G_{\mathrm{W \to Ol}}} + \beta \Delta G_{\mathrm{W \to Ol}} - \mathrm{ap}K_{\mathrm{a}})$	-5.753	-0.487	-0.017		1.06	8.28	0.83
		$+c_2^{\mathrm{2D}} (\mathrm{ap}K_{\mathrm{a}}^2 + \mathrm{bp}K_{\mathrm{a}}^2)$	(± 0.007)	(± 0.001)	(± 0.000)		(± 0.00)	(± 0.03)	(± 0.00)
f^{3D}	$= c_0^{\mathrm{3D}}$	$+c_1^{\mathrm{3D}} (\beta \Delta G_{\mathrm{W \to Ol}} - \mathrm{ap}K_{\mathrm{a}})$	-7.101	-0.614	-0.001	-0.018	0.94	8.19	0.86
		$+c_2^{\mathrm{3D}} (\mathrm{bp}K_{\mathrm{a}}^2 (\mathrm{ap}K_{\mathrm{a}} + \mathrm{bp}K_{\mathrm{a}}))$	(± 0.007)	(± 0.002)	(± 0.000)	(± 0.000)	(± 0.00)	(± 0.02)	(± 0.00)
		$+c_3^{\mathrm{3D}} (\mathrm{ap}K_{\mathrm{a}}^2 + (\beta \Delta G_{\mathrm{W \to Ol}})^2)$							

S4. ONE-DIMENSIONAL DESCRIPTORS

TABLE S2: Best one-dimensional descriptors for ten training sets as predicted by SISSO. Each column corresponds to a particular training set. The data demonstrates the robustness of the predictions across training sets—only twelve unique descriptors are present the best ten descriptors from all training sets. The top three descriptors do not change. The best one-dimensional descriptor $((\mathrm{ap}K_{\mathrm{a}} - \beta \Delta G_{\mathrm{W \to Ol}}) - (\mathrm{bp}K_{\mathrm{a}} + \beta \Delta G_{\mathrm{W \to Ol}}))$ and the baseline hydrophobicity descriptor have been highlighted for reference.

1D descriptor	Rank in training set number									
	1	2	3	4	5	6	7	8	9	10
$((\mathrm{ap}K_{\mathrm{a}} - \beta \Delta G_{\mathrm{W \to Ol}}) - (\mathrm{bp}K_{\mathrm{a}} + \beta \Delta G_{\mathrm{W \to Ol}}))$	1	1	1	1	1	1	1	1	1	1
$((\mathrm{bp}K_{\mathrm{a}})^2 + (\mathrm{ap}K_{\mathrm{a}} \cdot \beta \Delta G_{\mathrm{W \to Ol}}))$	2	2	2	2	2	2	2	2	2	2
$(\sqrt[3]{(\mathrm{bp}K_{\mathrm{a}})} + \beta \Delta G_{\mathrm{W \to Ol}})$	3	3	3	3	3	3	3	3	3	3
$((\mathrm{ap}K_{\mathrm{a}} - \beta \Delta G_{\mathrm{W \to Ol}}) - \mathrm{bp}K_{\mathrm{a}})$	4	5	4	6	4	4	5	6	5	6
$(\beta \Delta G_{\mathrm{W \to Ol}} + (\mathrm{bp}K_{\mathrm{a}} + \beta \Delta G_{\mathrm{W \to Ol}}))$	5	4	6	7	5	5	4	4	4	7
$(\sqrt[3]{(\mathrm{ap}K_{\mathrm{a}})} - \beta \Delta G_{\mathrm{W \to Ol}})$	6	7	7	5	7	7	7	7	7	5
$(\beta \Delta G_{\mathrm{W \to Ol}} + \sqrt[3]{(\beta \Delta G_{\mathrm{W \to Ol}})})$	7	6	5	4	6	6	6	5	6	4
$(\sqrt[3]{(\beta \Delta G_{\mathrm{W \to Ol}})} - \sqrt[3]{(\mathrm{ap}K_{\mathrm{a}})})$	8	8	8	8	8	8	8	8	8	8
$((\mathrm{ap}K_{\mathrm{a}})^{-1} + \beta \Delta G_{\mathrm{W \to Ol}})$	9	-	10	-	10	-	-	-	-	-
$(\beta \Delta G_{\mathrm{W \to Ol}})$	10	9	9	9	9	9	9	9	9	9
$(\sqrt[3]{(\beta \Delta G_{\mathrm{W \to Ol}})} + (\mathrm{bp}K_{\mathrm{a}} + \beta \Delta G_{\mathrm{W \to Ol}}))$	-	10	-	-	-	10	-	10	10	10
$((\mathrm{ap}K_{\mathrm{a}})^{-1} - \beta \Delta G_{\mathrm{W \to Ol}})$	-	-	-	10	-	-	10	-	-	-

S5. INPUT SCRIPT

To learn the permeability equations, we used the following SISSO[3] input script. It trains on 10% (= 41 897) of the compounds that map to a two-bead Martini representation (there are 418 971 such compounds) in the data provided by Menichetti *et al.*[1].

```
!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
! keywords for the target properties
!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
ptype=1  ! property type 1: continuous for regression, 2: categorical for classification
ntask=1  ! number of tasks (properties or maps) 1: single-task learning, >1: multi-task learning
nsample=41897  ! number of samples for each task (separate the numbers by comma for ntask >1)
task_weighting=1  ! 1: no weighting (tasks treated equally) 2: weighted by #sample_task_i/total_sample
desc_dim=3  ! dimension of the descriptor (<=3 for classification)
restart=.false.  ! set .true. to continue a job that was stopped but not yet finished

!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
! keywords for feature construction and sure independence screening
! implemented operators:(+)(-)(*)(/)(exp)(exp-)(^-1)(^2)(^3)(sqrt)(cbrt)(log)(|-|)(scd)(^6)(sin)(cos)
! scd: standard Cauchy distribution
!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
nsf=3   ! number of scalar features (one feature is one number for each material)
rung=2  ! rung (<=3) of the feature space to be constructed (times of applying the opset recursively)
opset='(+)(-)(*)(/)(exp)(log)(^-1)(^2)(^3)(sqrt)(cbrt)'  ! ONE operator set for feature transformation
maxcomplexity=10  ! max feature complexity (number of operators in a feature)
dimclass=  ! group features according to their dimension/unit; those not in any () are dimensionless
maxfval_lb=1e-3  ! features having the max. abs. data value <maxfval_lb will not be selected
maxfval_ub=1e5  ! features having the max. abs. data value >maxfval_ub will not be selected
subs_sis=500  ! size of the SIS-selected (single) subspace for each descriptor dimension

!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
! keywords for descriptor identification via a sparsifying operator
!>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
method='L0'  ! sparsification operator: 'L1L0' or 'L0'; L0 is recommended!
L1L0_size4L0= 1  ! If method='L1L0', specify the number of features to be screened by L1 for L0
fit_intercept=.true.  ! fit to a nonzero intercept (.true.) or force the intercept to zero (.false.)
metric='RMSE'  ! for regression only, the metric for model selection: RMSE,MaxAE
nm_output=100  ! number of the best models to output
```

[1] R. Menichetti, K. H. Kanekal, and T. Bereau, "Drug–membrane permeability across chemical space," ACS Central Science **5**, 290–298 (2019).

[2] *pK_a calculation* (Accessed October 10, 2020).

[3] R. Ouyang, "SISSO," `https://github.com/rouyang2017/SISSO` (2017).