Supplementary Information

A fast and interpretable deep learning approach for accurate electrostatics-driven $\mathrm{p}K_\mathrm{a}$ predictions

Pedro B.P.S. Reis,[*,†] Marco Bertolini,[‡] Floriane Montanari,[‡] Walter Rocchia,[¶] Miguel Machuqueiro,[*,†] and Djork-Arné Clevert[*,‡]

†*BioISI - Biosystems & Integrative Sciences Institute, Faculty of Sciences, University of Lisboa, Campo Grande, 1749-016 Lisboa, Portugal*

‡*Bayer A.G., Machine Learning Research, Berlin, Germany*

¶*CONCEPT Lab, Istituto Italiano di Tecnologia, Via E. Melen 83, 16152 - Genova, Italy*

E-mail: pdreis@ciencias.ulisboa.pt; machuque@ciencias.ulisboa.pt; djork-arne.clevert@bayer.com

List of Figures

List of Tables

Results

Table S1: PDB identification codes of the proteins in experimental test set.

135l	1a2p	1a6m	1a91	1a93	1ans	1b2v	1bcx
1beo	1bhc	1bi6	1bni	1bpi	1bus	1bvi	1cdc
1d0d	1div	1duk	1dwr	1eh6	1epg	1era	1ert
1eru	1ex3	1ey0	1fks	1fna	1goa	1gs9	1h4g
1hho	1hng	1hpx	1hrc	1i0v	1igd	1igv	1jbb
1kxi	1l54	1lni	1lys	1lz1	1mbc	1nfn	1nzp
1p5f	1pnt	1poh	1ppf	1ppo	1ptd	1qh7	1qlp
1rga	1rgg	1sap	1stg	1stn	1trs	1trw	1ubq
1wla	1xnb	1ymb	1yph	1ypi	1ypt	2bca	2bus
2cpl	2hnp	2igd	2lzm	2lzt	2ovo	2sni	2tga
2trx	2zta	3ebx	3egf	3fx5	3icb	3nbs	3rn3
3srn	3ssi	4icb	4lzt	4ma9	4mbn	4pti	6gst
6lyz							

Table S2: Performance comparison between the Null model (RMSE) and pKAI (RMSE; percentage of errors below 0.5 pH units). Information about the distribution of residue pK_{a} shifts (ΔpK_{a}) and relative solvent accessible surface area (SASA$_r$) in the test data is also shown. The Null model was calculated with ΔpK_{a} equal to zero.

Residue	Abundance (%)	Null	pKAI	Error	ΔpK_{a}		SASA$_r$	
		RMSE	RMSE	< 0.5 (%)	Avg	Stdev	Avg	Stdev
GLU	24.9	1.42	0.44	84.7	-0.7	1.2	0.43	0.24
LYS	22.5	1.04	0.32	92.1	0.6	0.9	0.47	0.23
ASP	21.9	1.74	0.50	80.5	-1.0	1.4	0.40	0.26
TYR	13.9	3.14	0.69	67.5	2.4	2.1	0.19	0.20
HIS	9.4	1.92	0.67	73.1	-1.0	1.6	0.29	0.25
CYS	3.9	3.30	0.82	56.6	2.8	1.8	0.11	0.17
NTR	1.7	0.74	0.28	94.2	-0.3	0.7	0.75	0.27
CTR	1.8	0.88	0.35	92.5	-0.2	0.9	0.74	0.27
All	100.0	1.89 (1.24[a])	0.52 (0.31[a])	81.2	0.0	1.9	0.38	0.27

[a] Mean Absolute Error (MAE)

Table S3: Execution time comparison between PypKa and pKAI. This benchmark was executed on a machine with a single Intel Xeon E5-2620 processor.

Protein	Number of residues / titratable	Execution Time (s)		Speedup Factor	Time per residue / titratable (s)	
		PypKa	pKAI		PypKa	pKAI
4LZT	129/21	26.5	0.8	33×	0.21/ 1.26	0.006/0.038
4K5C	341/100	92.0	1.2	76×	0.27/ 0.92	0.004/0.012
7C8J	902/249	2898.2	2.3	1260×	3.21/11.64	0.003/0.009

Table S4: Experimental pK_a benchmark of several methods on a data set of 736 residues from 97 proteins. For each method, we report their RMSE, the mean absolute error (MAE), the 0.9 quantile, the error percentage below 0.5 pK units, and the coefficient of determination (R^2). The null model values have been taken from.[1,2]

	RMSE	MAE	Quantile 0.9	Error < 0.5 (%)	R^2
Null	1.09	0.72	1.51	52.3	0.84
PypKA	1.07	0.71	1.48	52.6	0.85
PROPKA	1.11	0.73	1.58	51.1	0.84
pKAI	1.15	0.75	1.66	49.3	0.82
pKAI+	0.98	0.64	1.37	55.0	0.87

Table S5: Comparison between Null model and pKAI+ RMSE values. The Null model is defined as the pK_a values of the residues in water taken from Reference 1.

Residue	Abundance (%)	Null RMSE	pKAI+ RMSE	Error < 0.5 (%)	ΔpK_a		SASA$_r$	
					Avg	Stdev	Avg	Stdev
GLU	29.6	0.77	0.81	58.3	-0.5	0.9	0.45	0.24
LYS	14.4	0.74	0.68	60.4	0.3	0.6	0.55	0.21
ASP	29.2	1.30	1.08	59.5	-0.6	0.9	0.45	0.25
TYR	2.4	1.23	0.95	38.9	0.5	0.7	0.33	0.25
HIS	19.4	1.14	0.97	42.0	-0.5	1.1	0.39	0.22
CYS	1.2	3.39	3.43	0.0	-0.1	1.5	0.11	0.09
NTR	1.5	0.59	0.47	63.6	-0.3	0.8	0.74	0.20
CTR	2.2	0.41	0.56	75.0	-0.1	0.7	0.77	0.23
TOTAL	100.0	1.09	0.98	55.0	-0.4	1.0	0.46	0.25

[a] Mean Absolute Error (MAE)

Table S6: One hot encoding classes of all atoms used.

Atom Name	Residue	Atom Classes
N	Main Chain	N
O	Main Chain	O
NE2	GLN	N_AMIDE
ND2	ASN	N_AMIDE
OE1	GLN	O_AMIDE
OD1	ASN	O_AMIDE
NE	ARG	NE_ARG
NH1/NH2	ARG	NH_ARG
NZ	LYS	NZ_LYS
N	NTR	NZ_LYS
OXT	CTR	O_COOH
OD1/OD2	ASP	O_COOH
OE1/OE2	GLU	O_COOH
OG	SER	OG_SER
OG1	THR	OG1_THR
ND1	HIS	ND1_HIS
NE2	HIS	NE2_HIS
NE1	TRP	NE1_TRP
OH	TYR	OH_TYR
SG	CYS	SG_CYS
SD	Methionine	SD_MET

Table S7: RMSE improvement by adding a solvent exposure-related extra feature to the input layer. Different ways of measuring solvent exposure were tested: Half-sphere exposure (HSE), Coordination Number, Residue Depth, and relative solvent accessible surface area ($SASA_r$). HSE is a 2D measure and be subdivided into an upper (side chain facing, HSE^{up}) and lower sphere (backbone facing, HSE^{down}) half-spheres. Furthermore, two residues can be used as reference C_α (HSE_α) and C_β (HSE_β). Residue depth is the average distance of all residue's atoms to the molecular surface, and $\text{Residue Depth}_{C\alpha}$ is the atom depth of C_α.

Feature	RMSE Improvement
HSE_α^{up}	0.01
HSE_α^{down}	0.01
HSE_β^{up}	0.01
HSE_β^{down}	0.02
Coordination Number	0.02
Residue Depth	0.01
$\text{Residue Depth}_{C\alpha}$	0.01
$SASA_r$	0.02

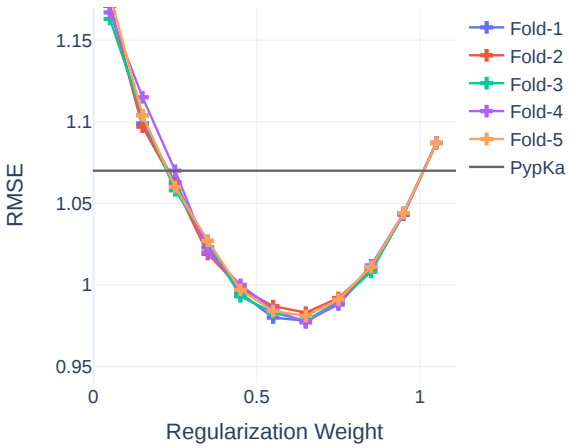

Figure S1: Performance of pKAI+ with different regularization weights on 5 folds of the experimental test set.

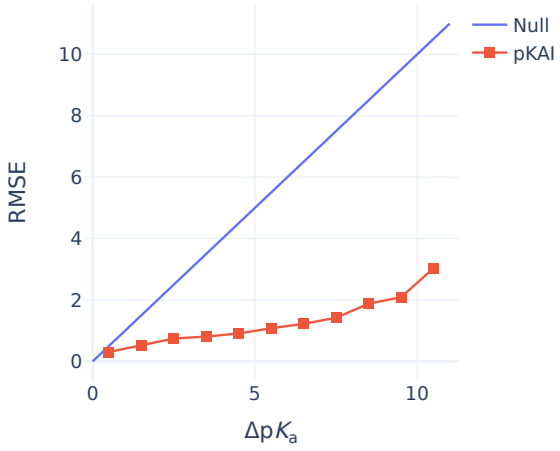

Figure S2: RMSE variation versus the magnitude of the pK_a shift (ΔpK_a). The calculations were performed for pKAI and Null model using the PypKa predictions as reference.

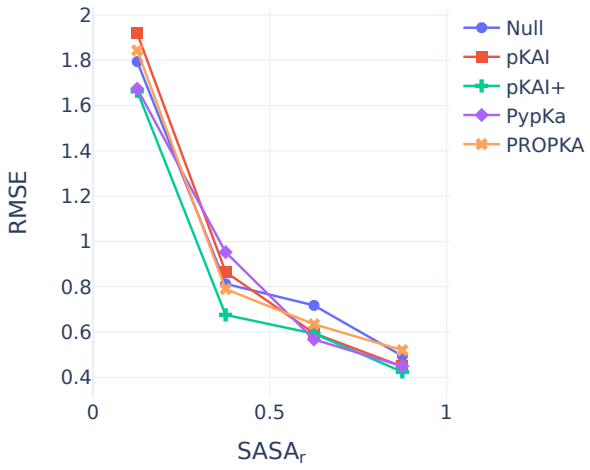

Figure S3: pKAI+ performance at predicting experimental pK_a values dependency on the magnitude of solvent exposure (SASA) of the residues.

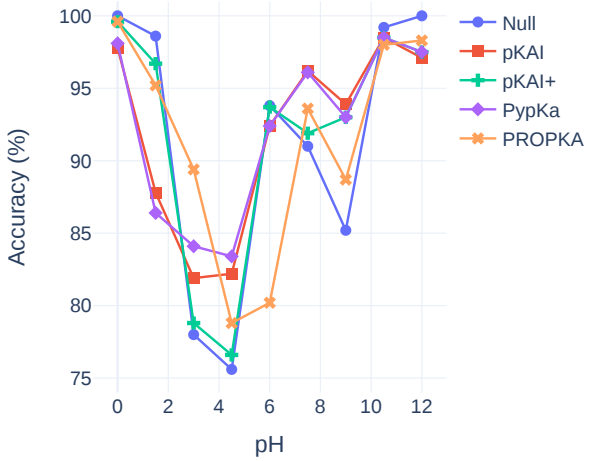

Figure S4: Accuracy of several methods at predicting the most representative protonation states derived from experimental pK_a values.

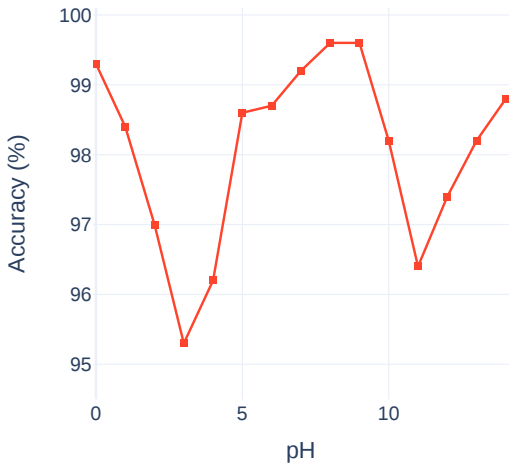

Figure S5: pKAI accuracy at predicting PypKa-derived protonation states.

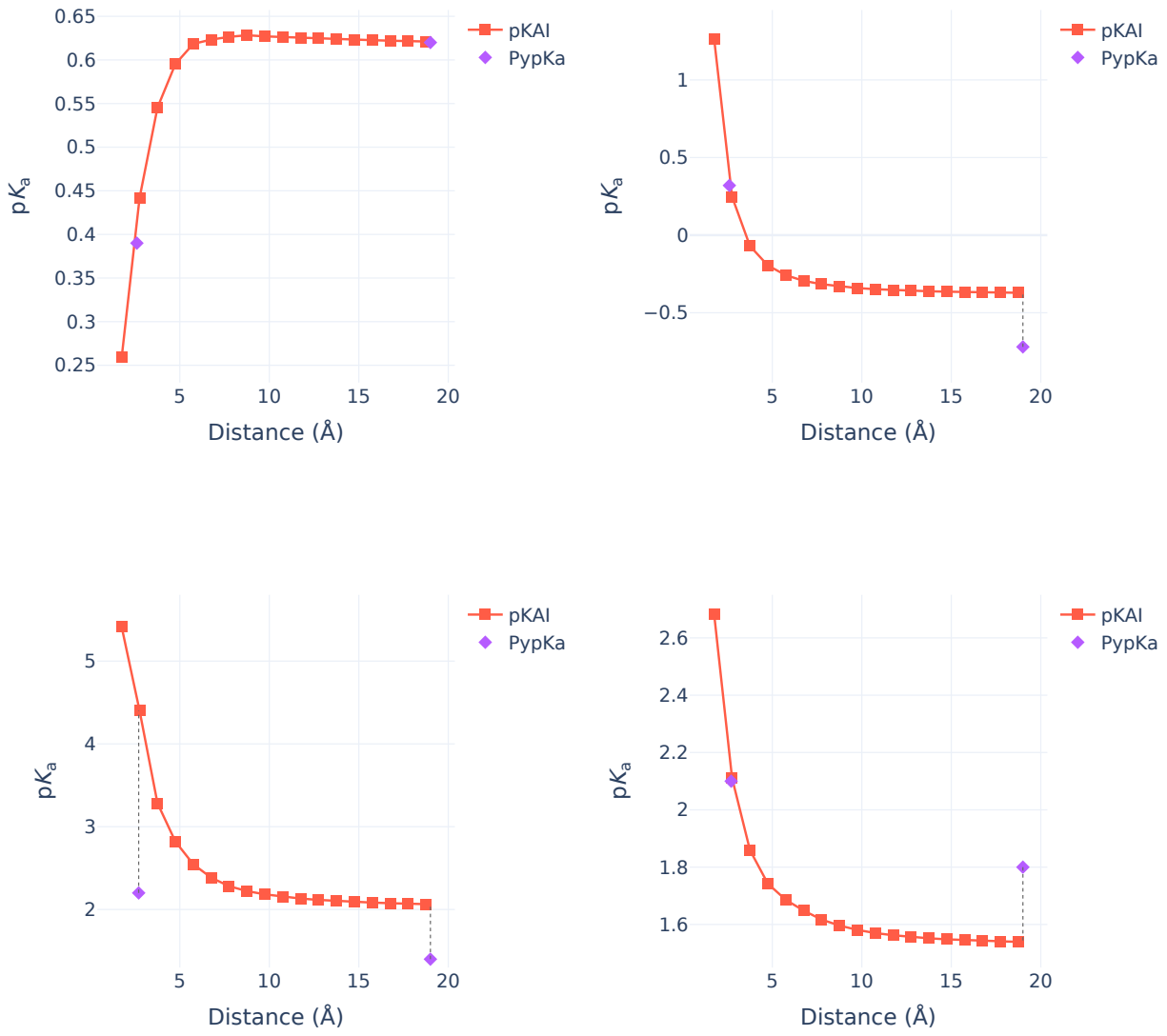

Figure S6: Impact of changing the distance of the closest atom on pKAI's predictions for: residue GLU-154 from structure 6FT4 (A); residue LYS-118 from structure 2HRK (C); residue TYR-98 from structure 6FT4 (C); residue LYS-55 from structure 2BJU (D). For reference, we have included PypKa's predictions of the same residue in the state presented in the experimental structure and in an modified structure in which the closest atom is absent.

References

(1) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups of proteins. *Protein Sci.* **2006**, *15*, 1214–1218.

(2) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci.* **2009**, *18*, 247–251.