Supporting Information for Manuscript titled

"Rapid and Accurate Prediction of pK_a Values of C-H Acids Using Graph Convolutional Neural Networks"

by Rafał Roszak[1,3], Wiktor Beker[1,3], Karol Molga[1], Bartosz A. Grzybowski[1,2,3]*

[1] Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, 01-224 Warsaw, Poland

[2] Institute for Basic Science, Center for Soft and Living Matter, Ulsan 44919, South Korea

[3] Allchemy, Inc., 2145 45th Street #201, Highland, IN 46322, USA

CONTENTS:

Section S1. Overview of other methods for pK_a prediction 1
Section S2. Quantum-chemical benchmark calculations 2
Section S3. Details on the composition of the training set 7
Section S4. Details of neural networks' setup 9
Section S4.1 Fingerprint based 9
Section S4.2 Embedded fingerprint 9
Section S4.3 Chemistry-aware NN 10
Section S4.4 Graph convolutional neural network (GCNN) 11
Section S5. Transfer learning experiment 12
Section S6. Other factors influencing reaction outcomes 14
Section S7. Supplementary references 16

Section S1. Overview of other methods for pK_a prediction

Methods of pK_a prediction can be divided into two categories: i) ab initio calculations and ii) statistical modeling. Ab initio calculations allow to estimate absolute pK_a from Gibbs free energy of dissociation reaction in a given solvent, (ΔG_{aq}): $HA = H^+ + A^-$. This value can be obtained from a thermodynamic cycle[7a]: $\Delta G_{aq} = G(A^-_{gas}) - G(AH_{gas}) + \Delta G_s(A^-) + \Delta G_s(H^+) - \Delta G_s(AH)$, where G denotes free energy of a given species and ΔG_s is the corresponding solvation energy. Since the proton solvation free energy cannot be obtained directly from quantum-chemical calculations, it has to be estimated using experimental data. Although this protocol allows for de novo estimation of pK_a of any molecule, values obtained in this manner "can be a few pK_a units away from the experimentally measured values, even for the simplest, most rigid molecules.".[6] To circumvent this problem and improve the prediction quality, a set of empirical correction parameters tailored for specific molecules/motifs may be used. An example of such an approach is the pK_a prediction module of the Jaguar software developed by Schrodinger.[7b]

A common approximation taken in *ab initio* approaches is the continuum (implicit) solvent model, which averages the internal states of the solvent and characterizes it with some global parameters (usually the dielectric constant). Absolute pK_a prediction with atomic resolution and accounting for the conformational flexibility of both the compound of interest and its surroundings is possible with *ab initio* molecular dynamics (MD) simulations (either Car-Parinello[7c] or Born-Oppenheimer[7d] scheme of computation). However, these approaches involve significant computational cost, which is probably the reason behind rather small number of such studies reported to date.[7e]

It is also worth to briefly narrate techniques useful for the correction of systematic errors in theoretical pK_a calculations, namely establishing the relationship between calculated parameters and the dissociation free energy (strictly proportional to pK_a). The most common type of this approach is the so-called Linear Free Energy Relationship (LFER), whereby one assumes linear dependence between pK_a and calculated dissociation energy: $pK_a = a*\Delta G+b$. Empirical constants *a* and *b* are obtained from linear regression to experimental data.[7f] This methodology, when properly tuned,[7h] yields highly accurate pK_a values, but its application is limited only to homologous series as regression parameters are not transferable between different groups of compounds (e.g., phenols and carboxylic acids would require different constants).

The last group of pK_a prediction methods is focused on identifying statistical correlations between molecular features and experimentally reported pK_a values. Historically, the first such model was based on Hammett and Taft parameters.[8a] Other descriptors were later developed leading to modern QSAR methods.[8b] However, such approaches were usually intended for rather narrow groups of compounds (e.g. phenols,[8c] amines,[8d] etc.). More recently, machine learning methods were applied to pK_a prediction as well,[2a] yet still restricted to functional groups ionizable in water solutions (i.e., not C-H acids). Quality of these models depends strongly on the number of available experimental data. Considering typical classes of pharmaceutical substances, tens of thousands of experimental pK_a values are available, allowing to achieve mean absolute error (MAE) in the range 0.5 – 0.9, depending on the software used.[5]

Section S2. Quantum-chemical benchmark calculations

Because the quality of the LFER approximation depends strongly on the configuration of QM calculations, we performed a comprehensive validation of the corresponding settings. Due to computational cost, we decided to limit ourselves to Density Functional Theory (DFT) methods and to not try any post-Hartree-Fock approaches, like Coupled Cluster or Møller–Plesset perturbation theory. Looking for a configuration providing the best correlation between calculated ΔG and the experimental pK_a values of 26 sulfone compounds, we independently evaluated the effects of the choice of density functional (Table S1), atomic radii (defining molecular cavity in implicit solvent model, Table S2) and the basis set (Table S3). All calculations were performed using Gaussian09 software.[S1]

First, a series of 50 density functionals (Table S1) was examined using 6-31+G(d) basis set with integral-equation-formalism variant of polarizable continuum model (IEF-PCM) of DMSO solvent (with UFF atomic radii scaled by 1.1). Each one of these DFT methods gives a very good correlation with experimental data, with coefficient of determination $R^2 > 0.95$ and mean absolute error (MAE) in the range from 0.65 to 1.16. The worst correlation was observed for functionals with empirical dispersion correction (e.g. wB97XD, APFD, B97D), whereas the best result was obtained for those including a PBE correlation term. Among all considered methods, the best result – that is, the lowest MAE and the highest R^2– was obtained for HISSbPBE functional.

In the continuum model of solvent, a molecule is placed inside a cavity within a continuous dielectric medium. The influence of the environment on the molecule is then modeled by point charges

induced on the surface of the cavity by molecular electric field. The cavity itself is constructed from a set of overlapping spheres centered on molecule's atoms, with the radius of each one defined as an atom-type-dependent parameter multiplied by a scaling factor. In order to find an optimal cavity definition, three types of atomic radii were tested (Bondi, UFF, Pauling) with different scaling factors (Table S2). This test was performed with B3LYP functional with 6-31+G(d) basis set. The best correlation was obtained using UFF atomic radii: the lowest MAE was observed with scaling factor 1.4, whereas the highest R^2 was obtained with scaling factor 1.6. We decided to take the midpoint value of the scaling factor (1.5) for the final model.

Finally, we evaluated how the choice of the basis set affected model's accuracy. The calculations were performed with B3LYP functional and continuum model of DMSO using IEF-PCM formalism with UFF atomic radii (scaling factor=1.1). Four families of basis sets were selected for tests (Table S3), namely: i) Pople (entries 1-9), ii) Duning (entries 9-15), iii) Jensen (16-22) and iv) Weigend (entries 23-27). We decided to not include computationally demanding quadruple zeta (or higher) basis sets. In the Pople's basis set series, mean absolute error of pK_a obtained with a widely used double-zeta 6-31+G(d) basis set was 0.82, yet even higher MAE was observed with corresponding triple-zeta 6-311+G(d) basis set. Any improvement in this family requires inclusion of significant number of polarization functions, like 6-31+G(2df,p). However, even such augmented basis sets gave slightly higher MAE than Jensen's double-zeta basis sets (pc1, pcseg-1 or pcSseg-1) comprised of significantly fewer functions (ca. half). In the Dunning family, MAE of double-zeta cc-pVDZ basis set is rather high, namely 0.85; incorporation of diffusion functions (entries 10-14) does not improve the correlation in the tested series of compounds. In this series, MAE was lower only in the case of rather large triple zeta cc-pVTZ, with the value of 0.61. Within Jansen's basis set family, the result does not depend strongly on the inclusion of diffusion functions, whereas extension to triple-zeta reduces MAE by only 0.06. Double-zeta basis sets from Weigend, def2svp and def2svpp, provide high MAE of 3.26 and 2.53, respectively, which is significantly higher than any other tested double-zeta basis set. On the other hand, triple-zeta def2tzvp gives low MAE of 0.68. To summarize, among all tested basis sets, the best results were obtained with the Jensen family, namely 0.57 in the case of pc-s-seg-2 (triple-zeta class) and 0.63 for pc-seg-1 (double-zeta class).

The choice of the basis set involves a trade-off between the computational cost (increasing with the size of the basis set) and accuracy (increasing with the basis set size). Hence, we decided to compare selected basis sets using production setup (HISSbPBE functional and UFF radii with scaling factor 1.5 for solute cavity definition). For the final comparison, we decided to choose two basis sets: pcseq-1, which was the best among tested double-zeta basis sets, and triple-zeta def2tzvp (Table S4). The def2tzvp basis set was ultimately chosen as a compromise between accuracy and size – this basis set gives slightly higher MAE that triple zeta basis sets from Jansen and Duning but contains ~15% less functions.

Table S1: Mean absolute errors (MAE) and coefficients of determination (R^2) for different density functionals. Calculations were performed for 26 sulfones, $PhSO_2$-R, using 6-31+G* basis set and integral equation formalism variant of polarizable continuum model of DMSO solvent (ε = 46.826, UFF radii with scaling factor 1.1)

Entry	Functional	R^2	MAE
1	B3LYP	0.96795	0.81711
2	HCTC/406	0.96623	0.76564
3	HCTH/147	0.96630	0.73145
4	HCTH/93	0.96526	0.73766
5	BLYP	0.95730	0.90410
6	tHTCH	0.96801	0.73107

Entry	Functional	R^2	MAE
7	BMK	0.95635	1.05251
8	tHTCHhyb	0.96973	0.76594
9	BPBE	0.96716	0.71087
10	MPWPBE	0.96695	0.74513
11	PW91PBE	0.96736	0.74998
12	wB97	0.95842	1.01437
13	wB97X	0.96168	0.98052
14	wB97XD	0.94982	1.15367
15	M06	0.96056	0.98845
16	M06HF	0.95208	1.15923
17	M06L	0.96297	0.84693
18	M11	0.95889	1.03419
19	M11L	0.95836	0.83824
20	MN12L	0.95744	0.98311
21	MN12SX	0.96652	0.87897
22	N12	0.96469	0.77204
23	N12SX	0.97529	0.65423
24	SOGGA11	0.96170	0.84898
25	SOGGA11X	0.97219	0.79256
26	M062X	0.96712	0.88886
27	X3LYP	0.96784	0.82808
28	B1B95	0.97602	0.69945
29	B3P86	0.97520	0.65900
30	B971	0.96944	0.79930
31	B972	0.97479	0.66320
32	B97D	0.95413	1.01965
33	Bb98	0.96853	0.81366
34	BhandH	0.97470	0.78554
35	BhandHLYP	0.96676	0.90800
36	**HISSbPBE**	**0.97830**	**0.64696**
37	HSEH1PBE	0.97570	0.66561
38	OHSE1PBE	0.97542	0.66832
39	OHSE2PBE	0.97622	0.65782
40	mPW1LYP	0.96611	0.87323
41	mPW1PW91	0.97564	0.65565
42	mPW3PBE	0.97494	0.65749
43	PBE1PBE	0.97622	0.65496
44	PBEh1PBE	0.97603	0.66767
45	PBEPBE	0.96702	0.73907
46	PW91PW91	0.96729	0.75778
47	TPSSh	0.97224	0.70149
48	TPSSTPSS	0.96813	0.74869
49	APF	0.97583	0.64786
50	APFD	0.95902	0.97976

Table S2: Mean absolute errors (MAE) and coefficients of determination (R^2) for different density functionals and for different definition of atom radii defining molecular cavity. Calculations were performed for 26 sulfones, $PhSO_2$-R, using B3LYP functional with 6-31+G* basis set and integral equation formalism variant of polarizable continuum model of DMSO solvent (ε = 46.826).

Entry	Radius type	Scaling factor	R^2	MAE
1		1.0	0.9603	0.9681
2		1.1	0.9679	0.8175
3		1.2	0.9703	0.7422
4		1.3	0.9725	0.6769
5	UFF	1.4	0.9736	**0.6711**
6		1.5	0.9739	0.6930
7		1.6	**0.9742**	0.7055
8		1.7	0.9739	0.7333
9		1.8	0.9735	0.7497
10		1.0	0.6156	2.3946
11		1.1	0.9518	1.0697
12		1.2	0.9637	0.9002
13	Bondi	1.3	0.9691	0.7947
14		1.4	0.9709	0.7377
15		1.5	0.9728	0.6853
16		1.6	0.9739	0.6714
17		1.1	0.9436	1.3589
18		1.2	0.9527	1.1716
19	Pauling	1.3	0.9615	0.9445
20		1.4	0.9661	0.8244
21		1.5	0.9675	0.7687

Table S3: Mean absolute errors (MAE) and coefficients of determination (R^2) for different basis sets. Calculations were performed for 26 sulfones, $PhSO_2$-R, using B3LYP functional and integral equation formalism variant of polarizable continuum model of DMSO solvent (ε = 46.826, UFF radii with scaling factor 1.1)

entry	Family	Basis set	R^2	MAE
1		6-31+G(d)	0.9679	0.8171
2		6-31++G(d)	0.9677	0.8234
3		6-31+G(2d)	0.9737	0.7153
4		6-31+G(2d,p)	0.9739	0.6980
5	Pople	6-31+G(2df,p)	**0.9760**	**0.6536**
6		6-31+G(2df,2p)	0.9756	0.6629
7		6-311+G(d)	0.9678	0.8367
8		6-311+G(d,p)	0.9688	0.8135
9		cc-pVDZ	0.9648	0.8554
10		apr-cc-pVDZ	0.9623	0.9324
11	Dunning	may-cc-pVDZ	0.9622	0.9332
12		jun-cc-pVDZ	0.9623	0.9324

entry	Family	Basis set	R^2	MAE
13		jul-cc-pVDZ	0.9622	0.9149
14		aug-cc-pVDZ	0.9619	0.9162
15		cc-pVTZ	**0.9750**	**0.6143**
16	Jensen	pc-1	0.9742	0.6564
17		pc-seg-1	0.9753	0.6344
18		pc-s-seg-1	0.9747	0.6384
19		aug-pc-seg-1	0.9699	0.8411
20		pc-2	0.9776	0.5746
21		pc-seg-2	0.9773	0.6013
22		pc-s-seg-2	**0.9779**	**0.5702**
23	Weigend	def2svp	0.9636	3.2601
24		def2svpp	0.9641	2.5265
25		def2tzvp	**0.9769**	**0.6840**
26		def2tzvpp	0.9765	1.2438

Table S4: Mean absolute errors (MAE) and coefficients of determination (R^2) for different basis sets. Calculations were performed for 15 nitriles using HISSbPBE functional and integral equation formalism variant of polarizable continuum model of DMSO solvent (ε = 46.826, UFF radii with scaling factor 1.5)

entry	Basis set	R^2	MAE
1	pc-seq-1	0.9726	0.9002
2	def2tzvp	**0.9823**	**0.7358**

Table S5: Mean absolute errors (MAE) and coefficients of determination (R^2) for different series of compounds. Calculations were performed at the HISSbPBE level with def2tzvp basis set and integral equation formalism variant of polarizable continuum model of DMSO solvent (ε = 46.826, UFF radii with scaling factor 1.5)

entry	Number of compounds	Compounds type	R^2	MAE
1	27	R-SO_2Ph	0.982	0.6232
2	15	R-CN	0.9823	0.7358
3	14	R-NO_2	0.9555	0.7866
4	11	R-COOEt	0.9207	1.0680

Section S3. Details on the composition of the training set

Table S6.Molecular make-up of the dataset. Each compound was assigned to class/group according to the patterns provided in the middle column. In case of multiple matches, the one higher in the Table was used for classification. Group 'Other' includes compounds like DMSO, methane, etc.

Class	Group Name	Group Pattern	No. of exp. examples	No. of theoretical examples
I	Nitroalkanes	[C!H0][NX3+](=[O])[O-]	14	56
II	[SX3+]	[C!H0][SX3+,SX4+]	4	0
III	**Total**		**261**	**156**
	1,3-diketones	[#6][CX3](=[OX1])[CX4!H0][CX3](=[OX1])[#6]	14	0
	Cyclic ketones	[CX4][CX3R]([CX4])=[OX1]	13	0
	Acetophenones	c1cc([*])ccc1[CX3](=[OX1])[CX4H3]	19	0
	Phenylketones	c1ccccc1[CX3](=[OX1])[CX4H2][*]	14	0
	Alkyl ketones	[CX4H3][CX3](=[OX1])[CX4H2][*]	8	0
	Other ketones	[C!H0][CX3]=[OX1]	16	0
	Esters	[C!H0][CX3](=[OX1])[OX2]	31	72
	Amides	[C!H0][CX3](=[OX1])[NX3H0]	8	0
	Thioamides	[C!H0][CX3](=[SX1])[NX3]	3	0
	Nitriles	[C!H0]C#N	55	42
	Sulfones	[C!H0][SX4](=[OX1])(=[OX1])	63	42
	Sulfoximides	[C!H0][SX4](=[N])(=[OX1])	7	0
	[PX4+]	[C!H0][PX4+]	10	0
IV	**Total**		**119**	**0**
	Thioethers	[C!H0][SX2H0]	22	0
	[PX3]	[C!H0][PX3]	1	0
	[OX1]=[PX4]	[C!H0][PX4]=[OX1]	4	0
	[NX4+]	[C!H0][NX4+]	2	0
	Ethers	[C!H0][OX2H0]	7	0
	Benzylic	[C!H0][c]	78	0
	Allylic	[CX4!H0][CX3]=[CX3]	5	0
V	Arenes	[cH]	6	194
Other	Compounds not belonging to either of above groups		10	2

Table S7. Similarity between groups in the dataset. In all comparisons, carbon atom with the lowest pKa value was marked as C-14 isotope before calculation of ECFP4 fingerprint. In this way, we intended to enforce comparisons of active sites between molecules. Average Tanimoto similarities within each group, as well as between the group and the rest of the dataset, are provided.

Class	Group Name	Average similarity within group	Average similarity to the rest of the dataset	No. of members
I	**Nitroalkanes**	**0.285**	**0.060**	**70**
II	**[SX3+]**	**0.244**	**0.104**	**4**
III	**Whole class**	**0.136**	**0.088**	**417**
	Ketones	0.201	0.106	84
	Esters	0.326	0.077	103
	Amides	0.240	0.103	8
	Thioamides	0.183	0.100	3
	Nitriles	0.216	0.071	97
	Sulfones	0.318	0.129	105
	Sulfoximide	0.430	0.128	7
	[PX4+]	0.353	0.125	10
IV	**Whole class**	**0.161**	**0.098**	**119**
	Thioethers	0.186	0.103	22
	[PX3]	Not applicable	0.136	1
	[OX1]=[PX4]	0.387	0.083	4
	[NX4+]	0.045	0.074	2
	Ethers	0.319	0.127	7
	Benzylic	0.192	0.122	78
	Allylic	0.050	0.048	5
V	**Arenes**	**0.145**	0.056	200
Other	**Whole class**	**0.098**	**0.086**	**12**

Section S4. Details of neural networks' setup

All neural networks were built in Python using Keras library with TensorFlow backend. Molecular descriptors and fingerprints were obtained with RDKit library (versions 2018.09 and 2017.09). The details of the training of NNs shown in the main-text Figure 4 (panels a-c) are described in the following subsections.

Section S4.1 Fingerprint based

Fingerprint-based input vector has 604 dimensions and consists of three parts:

1. Description of proton-donating atom, four features: Gasteiger charge of the neutral atom, Gasteiger charge of atom in anionic form (after proton dissociation), number of lone electron pairs and number of pi electrons on the atom.
2. Fingerprint representation of the core atom: ECFP2 fingerprint of radius 1[i] centered at the proton-donating atom. Such fingerprint was then cast into a bit vector of length 300.
3. Fingerprint representation of proton-donating atom neighbors: for each one, a bit vector was created in the same fashion as above; then, all those vectors were summed up.

Note that such a representation effectively describes environment of a proton-donating atom within topological distance of 2, since i) the nearest neighbors of center atom are explicitly described in the feature vector; ii) each of these neighbors is characterized by their environment of radius 1; and iii) the atoms within the topological distance of 2 from the center atom are simultaneously within distance of 1 of central atom's nearest neighbors. After the network optimization with 5-fold cross-validation, following hyper-parameters gave the lowest MAE: dropout = 0.05, L2 regularization = 0.005, learning rate = 0.002, batch size = 5 and ELU as a activation function.

We have also tested other input vectors (all having 604 dimensions).

1. The same as described in the paragraph above but radius of 2 was used for all fingerprints (thus the effective radius of proton-donating atom's environment covered by the model was 3). Such representation leads to higher MAE of 3.8 which is not surprising because of very significant overlap between the environments of central and neighboring atoms. Optimal network architecture consists of two hidden layers with 100 and 20 neurons respectively, dropout = 0.005, L2 regularization = 0.002, learning rate = 0.002, batch size = 5 and ELU as a activation function).
2. A representation sharing the same four features for proton-donating atom (Gasteiger charges, numbers of pi electrons and the lone pairs), but having only a ECFP4 fingerprint of proton donating atom (cast into a 600-bit vector). Best model of dense neural network with such an input gives MAE of 4.5 pK_a units. We also tested radii of value 3 and 4, which led to MAE 4.7 and 4.9, respectively.

Section S4.2 Embedded fingerprint

Here, we tested a fingerprint embedding based on the concept of word embedding (WE), which is currently a standard technique in natural language processing (NLP). The basic idea of WE is to assign a high-dimensional vector to each word in such a way that the distance between two words corresponds

[i] In the Extended Connectivity Fingerprints (ECFP; ref 17 in the main text), the last digit corresponds to the maximum diameter of neighborhood considered during computation. Diameters of value 4 and 6 (corresponding to the neighborhood radii 2 and 3, respectively) are most often used. Here, we implemented the radius 1, corresponding to the diameter 2 – thus, the proper name is ECFP2.

to the probability of their simultaneous occurrence in the same context. In the mol2vec approach,[17b] a molecule is converted into a list of ECFP4 identifiers (unique numerical representations of molecular fragments). Then, each identifier is treated as a 'word', whereas the whole molecule corresponds to a 'sentence'. Identifiers are ordered according to the occurrence of atoms in canonical SMILES. Such representation is then used for embedding with word2vec, which is an unsupervised WE method requiring to be trained on a large text corpus. Here, we use 19 million compounds taken from Zinc 15 database as such corpus. The embedding was trained with the skipgram method with 10 words window and the dimensionality of the resulting vector space set to 300. Each word that occurred in the corpus less than three times was replaced by string 'UNSEEN'. Embedded fingerprints for atoms were built in a following way: all identifiers of radius 1 from ECFP4 fingerprints (which describe just the atom and its immediate neighborhood) were embedded into the aforementioned 300-D space and the resulting vectors were summed up. Input vector for the NN was constructed in the same way as in previous section, but instead of ECFP4 fingerprints, the embedded fingerprints were used.
After the network optimization with 5-fold cross-validation, the following hyper-parameters gave the lowest MAE: dropout = 0.01, L2 regularization = 0.0005, learning rate = 0.002, batch size = 5 and ELU as an activation function.

Section S4.3 Chemistry-aware NN

Input vector for this network consists of 4 parts: the first one describes proton-donating atom whilst the rest characterize the neighbors. As previously, the proton-donating atom was described by four features: Gasteiger charge of the neutral atom, Gasteiger charge of the atom in anionic form (after proton dissociation), number of lone electron pairs, and the number of pi electrons on the atom. Each of substituents was assigned with a nine-element vector:

- Gasteiger charge of the atom
- Difference between Pauling electronegativity of the atom and electronegativity of carbon
- number of lone pairs on the atom
- number of pi electrons
- distance to the nearest functional group with known Hammett constant as defined in [S2]
- *F* parameter describing field/induction effect of substituent as defined in [S2]
- *R* parameter describing resonance effect of substituent as defined in [S2]
- Hammett substituent constant σ for *meta* position
- Hammett substituent constant σ for *para* position

When proton-donating atom had less than three substituents (hydrogen was not counted as a substituent) corresponding vector was filled with zeros.
After the network optimization with 5-fold cross-validation following hyper-parameters gave the lowest MAE: droupout = 0.1, L2 regularization = 0.001, learning rate = 0.002, batch size = 5 and ELU as an activation function.

Section S4.4 Graph convolutional neural network (GCNN)

Below, we include the details of the best GCNN architecture examined in the study:

- Input dense layer
  256 hidden features, activation = ReLU, L2 regularization = 0.005, dropout = 0.2
- Graph convolution part
  4 x 256 hidden features, local addition layers, activation = ReLU, L2 regularization=0.005, dropout = 0.2
- Output dense layer
  1 hidden feature, activation = ReLU, L2 regularization = 0.005

Table S8 conatins results of tests with architectures obtained by variation of the following parameters:
-number and type of Graph convolutional layers,
-number of hidden features,
-L2 regularization,
-dropout probability.

Table S8. Cross-validation results involving random splits.

Type of convolutional filter	No. of convolutional layers	Hidden units	L2 regularization	Drop out	CV MAE	Test MAE	MAE difference
Local addition layers	4	256	$5*10^{-3}$	0.2	2.07 ± 0.08	2.18	0.11
	6		$5*10^{-3}$	0.2	2.11 ± 0.06	2.23	0.12
	6	320	10^{-3}	0.1	2.3 ± 0.1	2.7	0.4
	4		10^{-3}	0.2	2.4 ± 0.1	2.9	0.5
	6		10^{-5}	0.1	2.4 ± 0.1	3.2	0.8
	4		10^{-5}	0.1	2.4 ± 0.1	2.6	0.2
	6		10^{-5}	0.2	2.4 ± 0.1	2.4	0.0
	6		10^{-3}	0.2	2.5 ± 0.1	2.6	0.1
	4		10^{-5}	0.2	2.5 ± 0.1	2.7	0.2
	4		10^{-3}	0.1	2.5 ± 0.1	2.4	-0.1
Second order Chebyshev polynomial	2		10^{-3}	0.2	2.6 ± 0.1	2.8	0.2
	3		10^{-3}	0.2	2.6 ± 0.1	2.7	0.1

In order to estimate the importance of features used to describe atoms in molecules, we performed an ‘elimination test’ for each descriptor. This procedure involved removal of the corresponding column from atomic feature matrix and subsequent 5-fold cross-validation of the model with the same configuration (that is, number of layers, hidden features, and regularization parameters). We decided to keep the atomic number, hybridization and the total number of attached hydrogen atoms as the ‘base’ model, and to focus on the evaluation of remaining features (Figure S1).

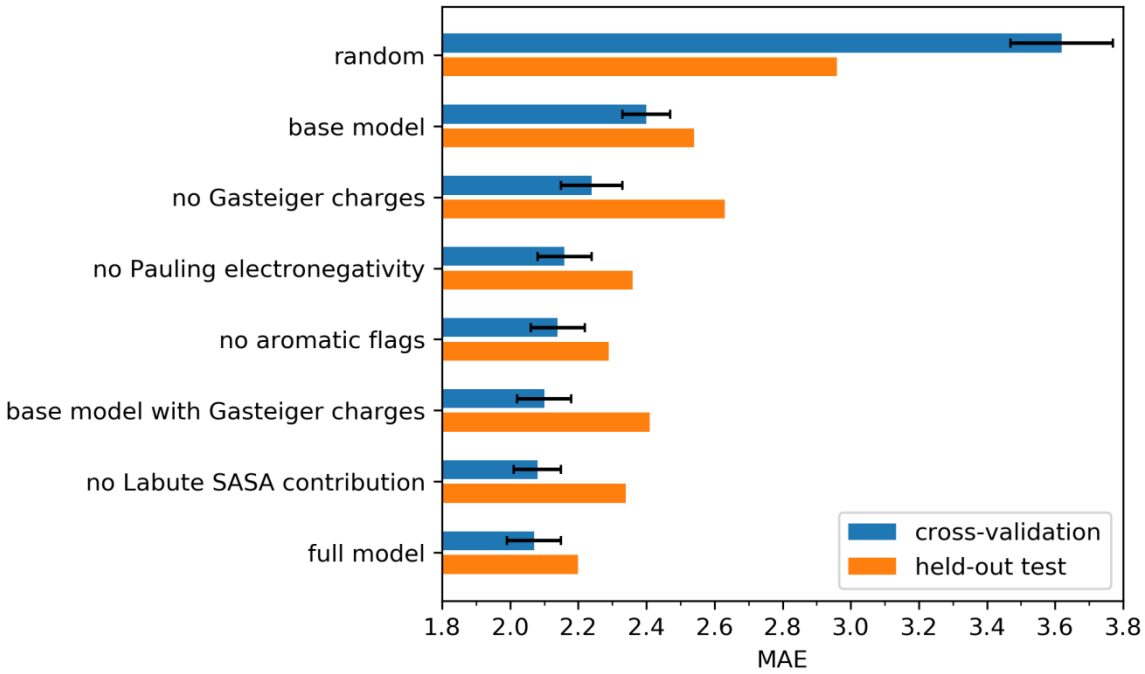

Figure S1. Performance of the best GCNN architecture upon removal of input descriptors. Labels on the y-axis describe models trained with the same architecture as the final model, but with a different set of features used to describe the input data. For the sake of comparison, three other models are presented as well: Base model = model including atomic number, hybridization and number of attached hydrogen atoms; random descriptors = model with random vector assigned to each distinct chemical element (see in the main text); full model = the best model discussed in the study. Blue bars represent the results from five-fold cross-validation on 90% of data (with error bars indicating the standard error of the resulting mean), whereas the orange series depict results obtained on the 10% held-out test set.

Section S5. Transfer learning experiment

We performed additional experiments to test whether a more robust model could be obtained with a *transfer learning* technique[S3]. The idea is to initially train a model on a large dataset different from the one considered here, and then to use its weights as a starting point in training our pKa model. We test this approach with the QM9 dataset[S4], which is one of popular benchmark sets for GCNNs[S5]. The set contains ~134,000 small molecules (up to 9 heavy atoms) for which several quantum-mechanically calculated quantities (HOMO/LUMO energy, dipole moment, polarizability, heat of formation, zero-point energy, etc.) were reported at B3LYP/6-31G(2df,p) level of quantum chemistry. It seemed possible that features useful in predicting these quantities would be as effective in prediction of pK_a. We found the presented architecture to perform well on this dataset (**Table S9**), comparably to models referred in the MoleculeNet benchmark[S5].

Table S9. Mean Absolute Error comparison of our GCNN (GC) model and benchmark models reported in MoleculeNet paper[S5] on the QM9 dataset[S4]. Only quantities exhibiting correlation with pK_a of relevant compounds are shown. Minimum values in each row are indicated by bold font.

Quantity	our GC	benchmark GC	benchmark DTNN	benchmark MPNN
Dipole moment	0.475	0.583	**0.244**	0.358
HOMO	**0.00384**	0.00716	0.00388	0.00541
LUMO	**0.00442**	0.00921	0.00513	0.00623
Gap	**0.0055**	0.0112	0.0066	0.0082
R2	39.9	35.9	**17**	28.5
U0	2.32	3.41	2.43	**2.05**
ZPVE	**0.00169**	0.00299	0.00172	0.00216

Next, we transferred the weights form this model to the one modified for node regression (the modification involves simple replacement of the last, node-gathering layer with ReLU unit) and trained the whole neural net on the pK_a values. Because QM9 set contains only second-row elements, we restricted the pK_a dataset to entries containing only H, C, N, O and F for consistency. Furthermore, we trained two baseline models: (i) one with random initialization of weights and (ii) one with transferred weights frozen during training (only parameters of the last ReLU unit were optimized). Results presented in **Figure S2** contradict the initial assumption about features learned by the model on the QM9 dataset. First, the behavior of the model with frozen weights suggests that hidden features useful in prediction of chosen 'quantum' properties can hardly be adapted to pK_a prediction. Second, the difference between model initialized randomly and the one with pre-trained weights seems negligible at the end of training. Although we could not establish whether the information from QM9 is erased or rebuilt during the first 10 epochs of training, this final convergence suggests that the transfer learning attempt did not provide any improvement.

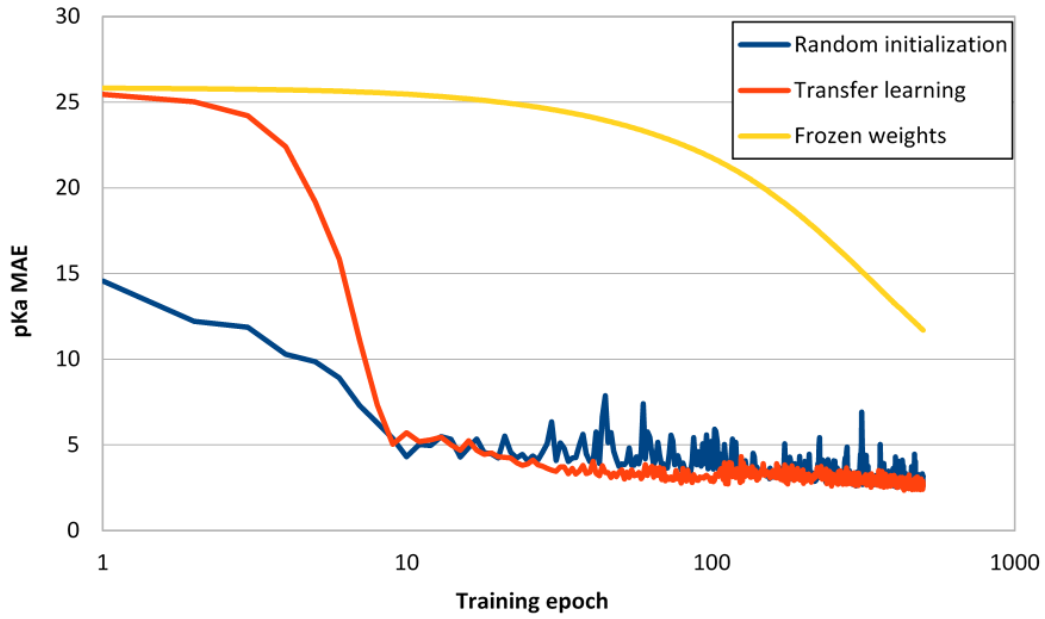

Figure S2. Results of transfer learning experiment. Curves present evolution of pK_a mean absolute error (calculated by 5-fold cross-validation with random splits) with training epochs. 'Random initialization' – GCNN model with randomly initialized weights, 'Transfer learning' – GCNN model initialized with weights pre-trained on QM9, 'Frozen weights' – GCNN model with fixed weights from QM9 model (optimization involves only last ReLU unit). Please note that the X-axis is in logarithmic scale.

Section S6. Other factors influencing reaction outcomes.

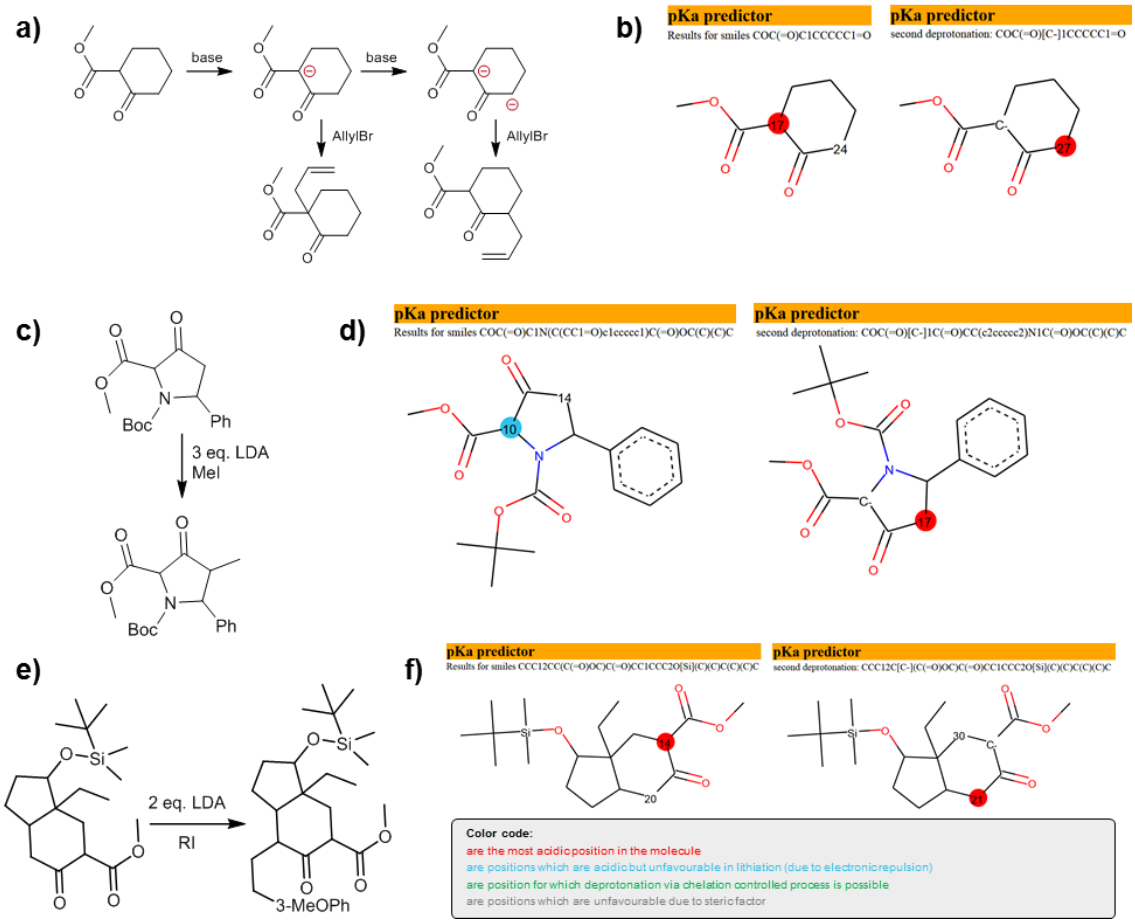

Figure S3. Additional examples for the prediction of the second deprotonation site made available by the use of two base equivalents and of a dianion. **a)** Experimentally observed outcomes of allylation of cyclohexanone when one or two equivalents of base are used[S6-S7]. **b)** Screenshots of the https://pka.allchemy.net/ web-app illustrating correct prediction of the first (*left*) and the second (*right*) deprotonation sites. **c,d)** Similar comparison for the alkylation of pyrrolidinone via dianion.[S8] **e,f)** Another example of alkylation of a cyclic ketoester from Corey's synthesis of Desogestrel.[S9]

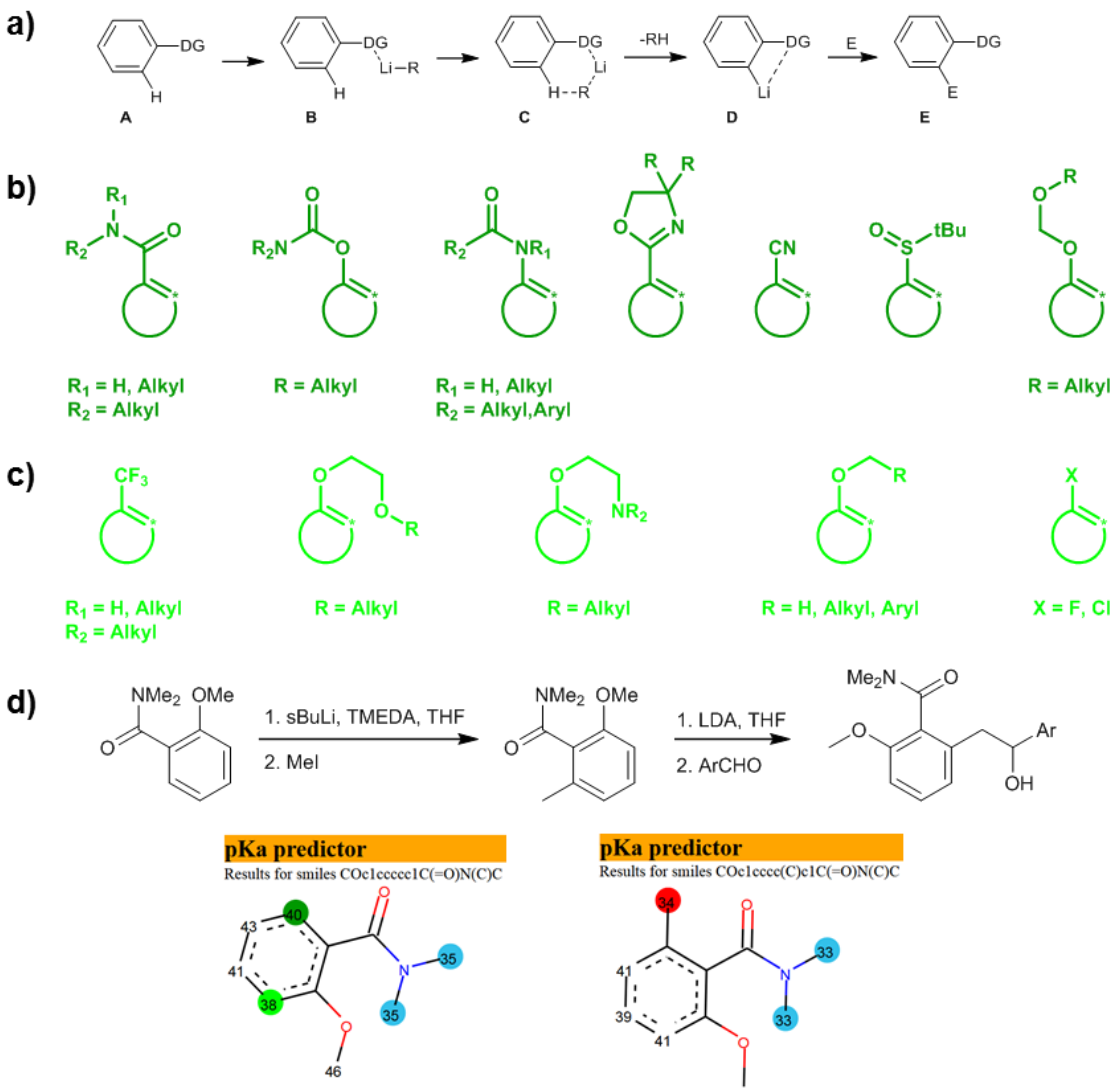

Figure S4. Deprotonation controlled by pre-coordination. **a)** The CH abstraction step (C→ D) occurs in an intramolecular fashion after pre-coordination of the Lewis-acidic metal (here, lithium) with the Lewis-basic directing group, DG. Additionally, the directing group stabilizes the obtained organolithium compound before the reaction with an electrophile. **b)** Strong and **c)** moderate directing groups controlling deprotonation. **d)** Initial steps in the total synthesis of Hydrangenol and Phyllodulcin[S10] are controlled by dimethylamide which is a stronger directing group than an aryl methyl ether.

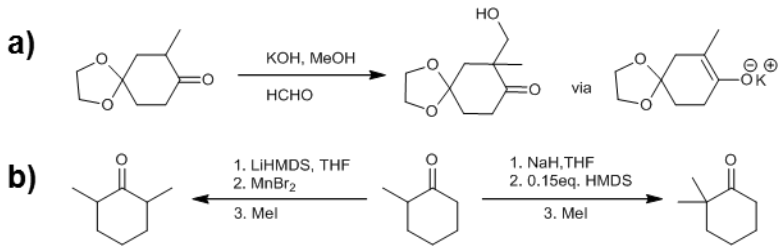

Figure S5. Selective functionalization of ketones via kinetic and thermodynamic enolates. **a)** Hydroxymethylation of less acidic position performed during total synthesis of Jiadifenin[S11] occurs via more stable tetrasubstituted enolate. **b)** Depending of the conditions used, the same substrate may deliver selectively one of the two possible products. For example, methylation of 2-methycyclohexanone leads to α,α-dimethyl cyclohexanone when treated with NaH in THF and equilibrated with 0.15 eq. of HMDS (in Baran's synthesis of Maoecrystal V reported in [S12]) or α,α'–dimethyl isomer[S13] when deprotonated with LiHMDS and converted to Mn enolate prior to addition of electrophile.

Section S7. Supplementary references

(S1) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; ; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09; Gaussian Inc.: Wallingford, CT, 2009.

(S2) Hansch, C.; Leo, A.; Taft, R. W. "A Survey of Hammett Substituent Constants and Resonance and Filed Parameters. *Chem. Rev.* **1991**, 91, 165-195.

(S3) Pan, S. J.; Yand, Q. A. Survey on Transfer Learning. *IEE Trans. Knowl. Data End.* **2010**, *22*, 1345-1359.

(S4) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **2014**, *1*,140022.

(S5) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswingd K; Pande V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.*, **2018**, *9*, 513-530.

(S6) Aburel, P. S.; Rømming, C.; Ma, K.; Undheim, K. Synthesis of α-Hydroxy and α-Oxospiranes Through Ruthenium(II)-Catalyzed Ring-Closing Metathesis. *J. Chem. Soc. Perkin Trans. 1* **2001**, *12*, 1458–1472.

(S7) Xue, F.; Seto, C. T. A Comparison of Cyclohexanone and Tetrahydro-4H-Thiopyran-4-One 1,1-Dioxide as Pharmacophores for the Design of Peptide-Based Inhibitors of the Serine Protease Plasmin. *J. Org. Chem.* **2005**, *70*, 8309–8321.

(S8) Davis, F. A.; Bowen, K. A.; Xu, H.; Velvadapu, V. Synthesis of Polysubstituted Pyrroles from Sulfinimines (N-Sulfinyl Imines). *Tetrahedron* **2008**, *64*, 4174–4182.

(S9) Corey, E. J.; Huang, A. X. A Short Enantioselective Total Synthesis of the Third-Generation Oral Contraceptive Desogestrel. *J. Am. Chem. Soc.* **1999**, *121*, 710–714.

(S10) Watanabe, M.; Sahara, M.; Kubo, M.; Furukawa, S.; Billedeau, R. J.; Snieckus, V. Ortho-Lithiated Tertiary Benzamides. Chain Extension via *o*-Toluamide Anion and General Synthesis of Isocoumarins Including Hydrangenol and Phyllodulcin. *J. Org. Chem.* **1984**, *49*, 742–747.

(S11) Carcache, D. A.; Cho, Y. S.; Hua, Z.; Tian, Y.; Li, Y.-M.; Danishefsky, S. J. Total Synthesis of (±)-Jiadifenin and Studies Directed to Understanding Its SAR: Probing Mechanistic and Stereochemical Issues in Palladium-Mediated Allylation of Enolate-Like Structures. *J. Am. Chem. Soc.* **2006**, *128*, 1016–1022.

(S12) Cernijenko, A.; Risgaard, R.; Baran, P. S. 11-Step Total Synthesis of (−)-Maoecrystal V. *J. Am. Chem. Soc.* **2016**, *138*, 9425–9428.

(S13) Reetz, M. T.; Haning, H. α-Methylation of Ketones via Manganese-Enolates: Absence of Undesired Polyalkylation. *Tetrahedron Lett.* **1993**, *34*, 7395–7398.