scientific data

OPEN

DATA DESCRIPTOR

A Dataset of Plausible Proton Transfer Steps for Arrow-Pushing Mechanisms

Alexander E. Dashuta[1], Ryan J. Miller[2], Pierre Baldi[2], Thomas Sander[3] & David L. Van Vranken[1]✉

Proton transfers are fundamental steps in polar reaction mechanisms. We generated a large dataset of over 51 million kinetically plausible proton transfer steps between heteroatoms from about 8,000 acids and conjugate bases with experimental aqueous pK_as, spanning pK_a values from −15 to +37. Rate factors were estimated at 25 °C using a simplified Eigen equation with pK_as but without statistical factors. Steps with estimated rate constants $\geq 10^3$ M^{-1} s^{-1} were included in the final dataset. Additionally, 5,043 proton transfer steps from carbon acids to heteroatom bases were estimated using the Eigen-Bernasconi equation based on reported intrinsic rate constants and Brønsted β values. Carbon proton transfers with rate constants $\geq 10^3$ M^{-1} s^{-1} were added to the final dataset. Each entry was encoded in SMIRKS format with electron-flow specification for machine learning compatibility. Diversity of structure was prioritized over diversity of conditions; calculated rate constants are expected to be accurate in aqueous environments. This approach and dataset should prove valuable for training models to predict stepwise mechanistic pathways.

Background & Summary

Proton transfer steps are particularly important in stepwise mechanistic pathways for organic transformations in synthesis, biochemical processes, and environmental chemistry. For example, an intermediate organic chemistry text[1] has over 70% (900 out of 1,269) diverse polar transformations that involve stepwise mechanisms with at least one proton transfer step. There are many datasets of acidic species with equilibrium pK_a values[2–5], but no large datasets of solution phase acid-base proton transfers suitable for training and learning (Table 1). For example, the Notre Dame Radiation Laboratory (NDRL) / NIST Solution Kinetics Database (through 1995) has abundant rate data, focusing on radicals[6]. Few of the 23,675 records involve proton transfers: none of the 11 records for pyridine involve proton transfer steps, and only 1 out of 26 records for acetic acid involve proton transfers. The proprietary Reaxys dataset has a large number of rate constants, but few for proton transfers[7].

Datasets have a significant advantage over general equations because exceptional examples can easily be added to a dataset, but not to an equation[8]. The most common datasets focus on transformations, without revealing the underlying stepwise mechanistic pathways. More recently, efforts have focused on databases of individual mechanistic steps[9]. The large NIST Chemical Kinetics database lists elementary reaction steps along with composite processes, mostly for gas-phase reactions[10]. Frenklach and workers created a dataset of mechanistic steps, with rate factors, from combustion processes[11,12]. Green, West, and coworkers have created a much larger dataset of mechanistic steps with rate constants for combustion processes[13]. Coley and coworkers created a large dataset of mechanistic steps and used it to train for prediction of stepwise pathways[14]. Grzybowski and coworkers generated a large dataset of carbenium ion rearrangements, with energetic parameters for prediction of rearrangement pathways[15]. Jung, Han, and coworkers also reported a large dataset of mechanistic steps to support prediction of polar reaction pathways[16]. We set out to create a large dataset of plausible proton transfer steps, each with electron-flow specification corresponding to curved arrows[17]. The term "plausible" (as opposed to "proven") accounts for wide variations in factors such as species concentrations and solvation. Our goal was to use equilibrium aqueous pK_as to estimate rate constants for proton transfer steps[18,19], and include in the dataset those proton transfer steps expected to be fast at room temperature using conservative boundaries.

[1]Department of Chemistry, University of California, Irvine, CA, 92697, USA. [2]Department of Computer Science, University of California, Irvine, CA, 92697, USA. [3]Alipheron AG, Schlossweg 63, 4143 Dornach, Switzerland. ✉e-mail: david.vv@uci.edu

Dataset Source	Proton Transfers	Entries with Rate Data	Availability	Ref
NDRL / NIST Solution Kinetics Database	Radical species with few proton transfers	NA	Public	[6]
Reaxys	2,741 Reaction Type contains "proton" or is "acid-base reaction"	21 proton transfers with subject "Rate Constant"	Proprietary	[7]
Plausible Proton Transfer Steps (this work)	51,510,157	24,890,236 with calculated or experimental log k_1	Public	

Table 1. Datasets of Proton Transfers with Rate Information.

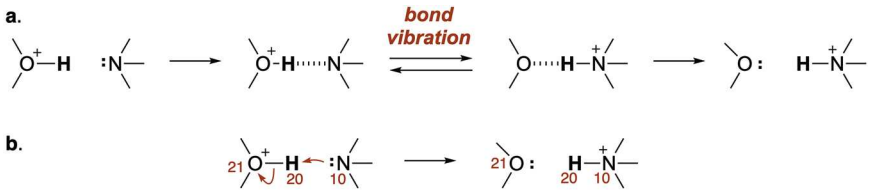


Fig. 1 Proton transfers have been represented as a stepwise or concerted process. (**a**) The 3-step Eigen mechanism for proton transfers. (**b**) The common 1-step arrow-pushing depiction of proton transfers, with SMIRKS code.

Simple, one-step representations of proton transfers belie a more complex process. The Eigen mechanism[20,21] for proton transfer involves three steps (Fig. 1a): i) formation of a hydrogen bond between acid and base, ii) transfer of proton from acid to base through a bond vibration, and iii) dissociation of the conjugate acid and the conjugate base. Following arrow-pushing convention, we treat proton transfers as a *one-step process* with no hydrogen-bonded intermediates (Fig. 1b). The exclusion of hydrogen-bonded intermediates is particularly important for consistency with Lewis representations in SMILES and SMIRKS text formats, which are commonly used for machine learning[17,22–25].

Methods

Proton Transfers to and from Heteroatoms. For proton transfers to and from heteroatoms, the second-order rate constants can often be predicted when the pK_{HA} of the acid and the pK_{HB} of the base are known[26]. Under equilibrium (or pre-equilibrium) conditions, when the acid is at least 1000 times more acidic than the conjugate acid ($pK_{HB} - pK_{HA} \geq 3$, Fig. 2), the forward proton transfer is diffusion-controlled ($k_1 = 10^9\,M^{-1}\,s^{-1}$). These diffusion-controlled proton transfer steps with rate factors $> 10^9\,M^{-1}\,s^{-1}$ should be considered plausible and are included in the dataset with $\log k_1 = 9$. When the backward proton transfer is diffusion-controlled ($-3 > \Delta pK_a$), the rate constant is readily estimated from the simplified relationship (Eq. 1), which ignores minor statistical factors for the number of acidic sites on the conjugate acid HB and the number of basic sites on the conjugate base[21].

$$\log k_1 = (pK_{HB} - pK_{HA}) + \log k_{-1} \approx \Delta pK_a + 9 \quad (1)$$

The mathematical relationship between rate and equilibrium doesn't hold for proton transfers between heteroatoms with comparable acidity ($3 > \Delta pK_a \geq -3$). The range over which ΔpK_a ceases to correlate with the rate constant depends on the heteroatoms[27], while the range is larger for thiols and carbon acids; carbon acids are dealt with separately below[28]. Even though the relationship given by Eq. (1) is unreliable for proton transfers between species of comparable acidity, proton transfers involving O, N, and S acids with comparable acidity are typically very fast in each direction ($k_1 \geq 10^6\,M^{-1}\,s^{-1}$)[29,30]. We assume the same holds true for Se (selenols)[31]. Therefore, proton transfers between heteroatom species with comparable acidity are included in the dataset, but no log k_1 is displayed.

In cases of thermodynamically *unfavorable* proton transfers where the product acid is more than 1000 times more acidic than the reactant acid ($-3 > \Delta pK_a$), the rate constants can still be calculated, but what rate constant is the cutoff for plausibility? Our goal for this dataset is to be conservative, excluding many plausible proton transfers in order to have greater confidence in the proton transfer steps in the dataset. In the standard state (25 °C), where both acid and base are present at 1 M, a second-order rate constant $k_1 = 10^{-1}\,M^{-1}\,s^{-1}$ corresponds to a half-life on the order of seconds (from $t_{1/2} = 1 / (k \bullet [HA]_0)$, where $[HA]_0 = [HB]_0$), which seems quite plausible. However, the concentrations of acidic and basic species are almost never 1 M and the solvent and other species will greatly impact proton transfer rates. To be more conservative, we include only those proton transfers predicted to be plausible at 0.1 mM where $k_1 \geq 10^3\,M^{-1}\,s^{-1}$ corresponds to $t_{1/2} < 10$ s. This would also correspond to acid-base proton transfers with $\log k_1 \geq 3$ and $\Delta pK_a \geq -6$.

Assembly of Arrow-Pushing Steps for Proton Transfers Between Heteroatoms. With these criteria in mind, we utilized the rich set of pK_{HA} values tabulated in the DataWarrior dataset made by Sander and coworkers[32]. The DataWarrior dataset contains 7,913 entries with SMILES representations, temperatures, and pK_{HA} values; 1,002 entries without temperatures were assumed to be at room temperature. About half of the structures,

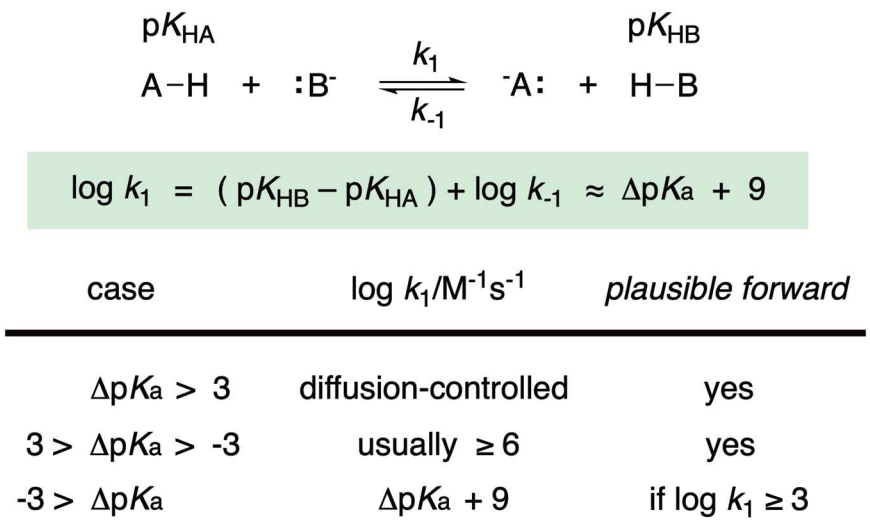


Fig. 2 Relationship between equilibrium and rate constants for proton transfers involving heteroatoms.

represented in the conjugate base form, were converted to SMILES for the conjugate acid form. Aromatic molecules, except for tropylium cation and a few others, are represented in non-Kekulized SMILES forms. Entries with temperatures near the range from room temperature (25 °C) to physiological temperature (37 °C) were included. Four examples with pK_{HA} values below 15 °C and 6 examples with pK_{HA} values above 40 °C were discarded, leaving 7,913 acids. The set was culled to remove 185 entries for which the site of protonation, tautomeric form, or other aspects of structure could not be confirmed, leading to a total set of 7,728 acids. Carbon acids were dealt with in a different way, so 209 carbon acids were removed to leave 7,519 heteroatom acids in the list. However, HCN has been shown to behave as a normal Eigen heteroatom acid[33] and was added to the dataset with a pK_a of 9.0. For 41 acids with more than one pK_{HA} value, the entry with the pK_a value closest to 7.0 was used, leaving 7,479 entries. Water (H_2O) and hydronium ion (H_3O^+) were notably absent from the DataWarrior dataset. Traditional pK_a values for H_2O (pK_a 15.7) and H_3O^+ (pK_a −1.7) have recently been revised by log(55 M), to values of 14.0 and 0, respectively[34,35]. We added H_2O and H_3O^+ to our list of acids, using the newer values. The final set of heteroatom acids and bases taken from DataWarrior contained 7,481 entries.

To better represent mechanistic intermediates for which no experimental pK_as are available, pK_a values for a range of other heteroatom acids were taken from the well-known Reich tables[36], from Guthrie's estimates of pK_as for mechanistic intermediates[37,38], plus a set of sixteen distinctive heteroatom functional groups not represented in the DataWarrior set. This additional set of 132 additional acids was added to the DataWarrior set for a total of 7,613 in the overall Heteroatom set.

For each entry, the acidic proton in the SMILES string was mapped as atom 20 and the attached heteroatom was mapped as atom 21, for subsequent use in arrow-pushing. A complementary set of 7,613 conjugate bases was generated and the basic atom was mapped as atom 10 for use in arrow-pushing specification.

From the Heteroatom set of 7,613 heteroatom acids and 7,613 heteroatom bases, all possible combinations ($7{,}613^2 = 57{,}957{,}769$) of acid-base reactions were generated. Combinations are represented in SMIRKS format along with electron-flow specification[17] where the basic atom mapped as 10 deprotonates the acidic proton mapped as 20 and the attached heteroatom mapped as 21. Each entry included fields for pK_{HA}, pK_{HB}, and ΔpK_a. Transformations with $\Delta pK_a \geq -6.00$ (89%) were deemed plausible and were retained in the final dataset of 51,505,065 proton transfer steps between heteroatoms in SMIRKS format, ready for use in applications such as deep learning of stepwise mechanisms. Doubly-charged species are common in the Heteroatom dataset (6.3% of acids have net charge $> +1$ and 6.6% of bases have net charge < -1) so it is important for users to be aware that steps that form dications or dianions are likely to be less plausible in non-aqueous solvents.

Using a precise cutoff of $\log k_1 \geq 3.00$ may seem arbitrary because few of the tabulated pK_as were determined with ±0.01 accuracy. However, the cutoff removes only 6,452,704 (11%) out of the nearly 58 million proton transfer steps. The precise cutoff results, for example (Fig. 3), in inclusion of a proton transfer from a *Z* oxime to a quinuclidone (Eq. 2, $\log k_1 = 3.00$), but exclusion of a proton transfer from the same *Z* oxime to an enolate (Eq. 3, $\log k_1 = 2.99$). Interestingly, the same proton transfer involving the more acidic *E* oxime (pK_{HA} 10.75)[39] is included in the dataset (Eq. 4, $\log k_1 = 3.57$). Fortunately, both the acid and the base in the excluded entry (Eq. 3) are well-represented in this combinatorial dataset with 4,473 and 7,242 occurrences of each, respectively.

Proton Transfers Between Heteroatoms and Carbon. Rates of proton transfers to and from carbon are generally slower than corresponding rates for proton transfers between heteroatoms[27]. We set aside proton transfers between carbon atoms, with the expectation that carbon-to-carbon proton transfers will be much slower than proton transfers involving at least one heteroatom[40–43]. Rate constants for deprotonation of carbon acids by heteroatom bases have been shown to obey the general Brønsted relation in Eq. (5)[44], where *p* and *q* are statistical factors relating to the number of equivalent H^+ sites on the acid and the number of basic sites on the base, respectively.

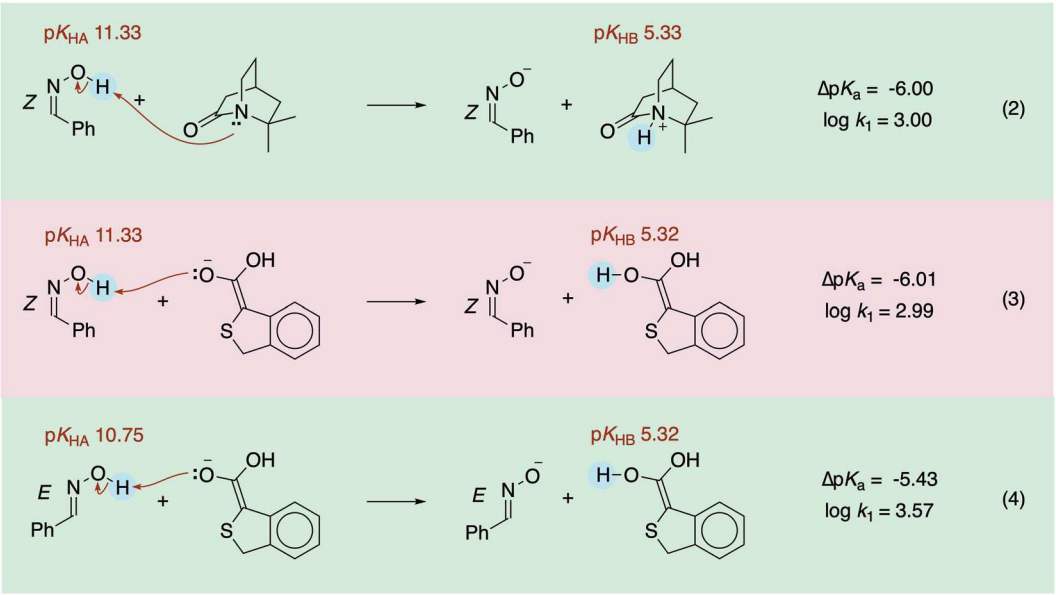


Fig. 3 The precise cutoff of log $k_1 \geq 3.00$ leads to exclusion of some proton transfer steps yet each acid and base in Eq. 3 is still represented thousands of times in the dataset.

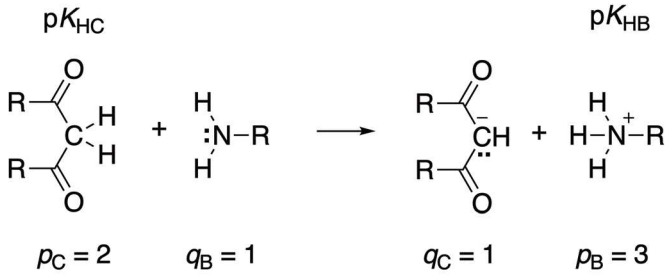
$$\log k_1 = \beta\left[(\mathrm{p}K_{\mathrm{HB}} - \mathrm{p}K_{\mathrm{HC}}) + \log\left(\frac{p_{\mathrm{B}}\, q_{\mathrm{C}}}{q_{\mathrm{B}}\, p_{\mathrm{C}}}\right)\right] + \log k_o + \log(q_{\mathrm{B}} p_{\mathrm{C}}) \quad (5)$$

Statistical factors were included because they were relatively easy to assign for the small set of carbon acids. For planar enolates, each face of the enolate should be considered a statistically different site of protonation ($q_{\mathrm{C}} = 2$) but for this work, we treated planar enolates as having a single basic site ($q_{\mathrm{C}} = 1$). Diastereotopic protons were treated as equivalent for XH_2 acids such as CH_2. The value of the statistical terms in Eq. (5) correspond to $\log [(p_B q_C)\beta \bullet (q_B p_C)_{(1-\beta)}]$ and ranged from −0.69 to +0.80. Only 2,142 out of 20,859 (10%) of proton transfers from carbon acids involved a statistical term $\geq \log(3) = 0.48$. Without the inclusion of the statistical factors, 4,258 (20.4%) of the proton transfers from carbon were above the log $k_1 = 3.00$ cutoff. When the statistical factors were included, 5,043 were above the cutoff; of those, 730 (14%) were raised above the cutoff by a statistical term of log(3) or less. In a typical reaction, the uncertainties in concentrations will have a greater impact on rate $= k_1$[acid][base] than the statistical factors that affect k_1. These statistical effects are small when considering rate constant ratios on the order of thousands, millions, or billions.

Assembly of Arrow-Pushing Steps for Proton Transfers Between Carbon and Heteroatoms. Grzybowski and coworkers have trained a system to accurately predict a wide range of equilibrium pK_as for a wide range carbon acids[45]. Yet, intrinsic rate constants (k_o) and Brønsted β values needed for Eq. (5) are only available for a much smaller number of carbon acids, typically with $\mathrm{p}K_{\mathrm{HC}} < 15$ and a number of classes of bases such as 1°, 2°, and 3° amines, HO^-, and carboxylate anions. The carbon acids are typically substituted with one or more anion-stabilizing groups such as nitro, benzoyl, acetyl, carboxyalkyl, cyano, phenyl, pyridinium, phosphonium, and sulfonium. Subsets of bases and $\mathrm{p}K_{\mathrm{HB}}$ values were extracted from the Heteroatom set of 7,613 bases: for example, a set of 1,048 carboxylate anions, 439 primary amines, 380 secondary amines, 857 tertiary amines, 42 thiolate anions, 637 aryloxide anions, and 59 alkoxide anions, including HO^- that has the revised pK_a value of 14.0 for its conjugate acid. For carbon acids, log k_o and β are dependent on the structure of the base; 81 different pairings were created from 7 heteroatom base classes and 65 specific carbon acids.

We then generated 20,859 combinatorial variations and calculated log k_1 according to the Eigen-Bernasconi equation (Eq. 5, Fig. 4): $\log k_1 = \beta\,[(\mathrm{p}K_{\mathrm{HB}} - \mathrm{p}K_{\mathrm{HC}}) + \log(p_{\mathrm{B}}\, q_{\mathrm{C}}/q_{\mathrm{B}}\, p_{\mathrm{C}})] + \log k_o + \log(q_{\mathrm{B}}\, p_{\mathrm{C}})$. We included in the dataset proton transfer steps with log $k_1 \geq 3$. The resulting set of 5,043 proton transfer steps, from carbon acids to heteroatom bases, were included in the total dataset. Application of this cutoff led to exclusion of proton transfer steps involving 6 out of 65 carbon acids, mostly simple nitro compounds. As a result of the conservative cutoff, there were no remaining examples of tertiary amines deprotonating carbon acids. The final set of proton transfers from carbon acids contained 71 different combinations of the 6 heteroatom base classes with the 59 specific carbon acids.

Experimental determination of rate constants is laborious, and even more so for intrinsic rate constants (k_o) and Brønsted values. Due to the paucity of parameters for proton transfer to carbon bases, the dataset does not contain rate constants calculated for combinatorial variations of carbon bases. To fill this gap, we identified in the literature 49 proton transfers from heteroatom acids to carbon bases (mostly enol ethers and nitronate



$$\log k_1 = \beta\left[\left(pK_{HB} - pK_{HC}\right) + \log\left(\frac{p_B q_C}{q_B p_C}\right)\right] + \log k_o + \log(q_B \cdot p_C)$$

Fig. 4 The Eigen-Bernasconi equation relates rate constants for proton transfers (from carbon) to equilibrium pK_as, intrinsic rate constants, and statistical factors. Log k_o is dependent on the structure of each carbon acid and each family of heteroatom bases.

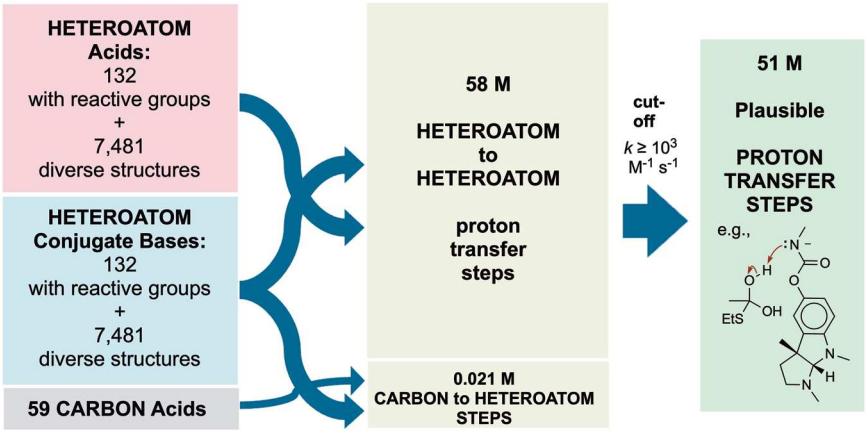


Fig. 5 The DataWarrior dataset of structurally diverse heteroatom acids was curated and then complemented with a smaller set of functionally diverse heteroatom acids. Conjugate bases were generated from these acids. Acids and bases were combined to generate 58 M proton transfer steps. A rate factor cutoff was then applied to yield the final set of 51 M Plausible Proton Transfer Steps.

anions) for which experimentally measured rate constants were available with $k_1 \geq 10^3\,M^{-1}\,s^{-1}$ ($\log k_1 \geq 3$). The advantage of datasets over rules is that additional examples are easily appended.

Data Records

Structuring of Data Records. All the datasets are available for download as a single zipped file (https://doi.org/10.6084/m9.figshare.30875087)[46]. One folder contains two types of acidity and basicity data that were used to generate millions of raw heteroatom-heteroatom proton transfer steps. They consisted of 7,613 heteroatomic acids (Acid.csv) and 7,613 conjugate bases (ConBase.csv) with structures in SMILES format with pK_as. Another folder contains eight types of acidity and basicity data (in CSV format) used to generate thousands of raw carbon to heteroatom proton transfer steps. The carbon acids data (Carbon_Acids.csv) consists of names, SMILES, pK_as, statistical factors p_C and q_C, intrinsic rate constants (k_o), Brønsted β parameters, and literature references. The data for the seven different classes of bases, each in separate file, consisted of SMILES, pK_as, statistical factors p_B and q_B, and literature references (Fig. 4).

These acidity and basicity data were used to generate combinatorial variations of proton transfer steps: i) 51 M proton transfer steps (51M_Heteroatom.csv) from heteroatom acids to heteroatom bases, with SMIRKS, calculated log k_1, and pK_as, ii) 5 K proton transfer steps from carbon acids to heteroatom bases (5KCarbonPT.csv), with SMIRKS, calculated log k_1, and pK_as, Brønsted β values, statistical factors (q_B, p_B, q_C, p_C) intrinsic rate constants (k_o), iii) 49 proton transfers from heteroatom acids to carbon bases (49ExperimentalCarbonPT.csv) in SMIRKS format, with experimentally measured log k_1, and literature references. Representative samples of the 51 M heteroatom to heteratom proton transfer steps are also included as CSV files (100K_Heteroatom.csv and 100_Heteroatom.csv).

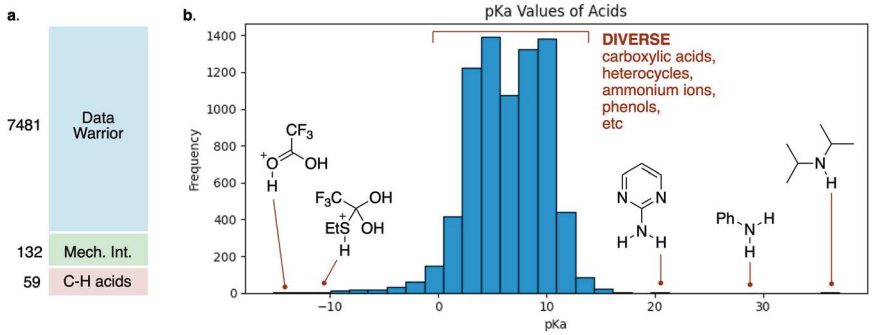


Fig. 6 The dataset of proton transfer steps incorporates over 7,600 acidic species with pK_as spanning a range of over 50 orders of magnitude. (**a**) The proton transfer steps in the dataset involve acids from several sources. (**b**) The set of acids in the proton transfer steps cover a large range of pK_as, mostly heteroatom acids with experimentally determined pK_as in the range of 0 to 14.

HA (pK_{HA})	B (pK_{HB})	log k_1 (calc)	log k_1 (exp)	Δ log k_1	Ref
NH_4^+ (9.2)	$CH_3CO_2^-$ (4.7)	4.5	4.32	+0.18	[48]
HCN (9.0)	CH_3ONH_2 (4.62)	4.62	5	−0.38	[49]
NH_4^+ (9.2)	$CH_3CH_2CO_2^-$ (4.83)	4.63	4.3	+0.33	[48]
NH_4^+ (9.2)	*t*-$BuCO_2^-$ (5.04)	4.84	4.58	+0.26	[48]
$PrNH_3^+$ (10.65)	imidazole (7.07)	5.42	6.15	−0.73	[30]
CH_3CO_2H (4.7)	$Cl_2CHCO_2^-$ (1.48)	5.78	6.76	−0.98	[30]
CH_3CO_2H (4.7)	$ClCH_2CO_2^-$ (2.85)	7.15	7.38	−0.23	[30]
imidazoleH$^+$ (7.07)	pyridine (5.26)	7.19	7.43	−0.24	[30]
imidazoleH$^+$ (7.07)	α-picoline (5.98)	7.91	7.81	+0.10	[30]
malonic acid (5.7)	$CH_3CO_2^-$ (4.7)	8.0	7.79	+0.21	[30]

Table 2. Comparison of Calculated log k_1 for Heteroatom to Heteroatom Proton Transfer Steps in the Dataset with Reported Experimental Values.

HA (pK_{HA})	B (pK_{HB})	log k_1 (calc)	log k_1 (exp)	Δ log k_1	Ref
$PhSCH_2NO_2$ (6.67)	piperazine (9.93)	3.16	3.02	+0.14	[50]
$PhCOCH(Ph)NO_2$ (5.04)	$MeO_2CCH_2CH_2S^-$ (9.33)	3.25	3.02	+0.23	[51]
$PhCOCH_2NO_2$ (4.67)	PhO^- (9.9)	3.54	3.51	+0.03	[51]
$PhSO_2CH_2$(4-Py$^+$)Me (11.54)	*n*-$BuNH_2$ (10.7)	3.90	3.57	+0.33	[44]
indanedione (6.35)	*n*-$BuNH_2$ (10.7)	4.64	4.16	+0.48	[52]
$AcCH_2PPh_3^+$ (7.83)	$MeOCH_2CH_2NH_2$ (9.67)	5.02	5.16	−0.14	[44]
indanedione (6.35)	piperidine (11.24)	5.39	4.89	+0.50	[52]

Table 3. Comparison of Calculated log k_1 for Carbon to Heteroatom Proton Transfer Steps in the Dataset with Reported Experimental Values.

Data Overview

A large dataset was obtained through combinatorial assembly and application of a conservative rate cut-off (Fig. 5). The few examples of carbon bases are important but did not significantly expand the size of the dataset.

The resulting dataset of proton transfer steps incorporates over 7,600 acidic species with pK_as spanning a range from −15 to + 37 (Fig. 6). The majority of the acids come from the DataWarrior dataset, which is a structurally diverse set of heteroatom N, O, and S acid species but covers a relatively limited range of pK_as. A smaller set of about 100 heteroatom acids was added to ensure that mechanistic intermediates, including species with very high or very low estimated pK_as, were included. The set of carbon acids is much smaller because the necessary intrinsic rate constants and Brønsted parameters are not widely available. The carbon acids cover a limited range of pK_as from 5 to 20.

The total dataset of 51,510,157 plausible acid-base proton transfer steps was created combinatorially, using equilibrium aqueous pK_as to estimate rates of proton transfer. The proton transfer steps are encoded in SMIRKS format with electron-flow specification, which is particularly suitable for machine learning. The majority of the dataset (51,505,065) contains proton transfers to and from heteroatoms; 24,885,144 (48%) of the 51,505,065 entries have log k_1 values calculated by applying pK_as in the Heteroatom set to Eq. (1). The dataset contains a smaller number (5,043) of proton transfers from carbon acids to groups of heteroatom bases, each with log k_1

from Eq. (5). The dataset also includes a small number (49) of proton transfer steps from heteroatom acids to carbon bases, each with log k_1 values. A conservative cutoff was used to determine plausible steps: log $k_1 \geq 3$ for proton transfers between heteroatoms and for proton transfers to or from carbon Table 1.

Technical Validation

Validation of Rate Constants - Comparison of Calculated log k_1 with Experimentally Measured log k_1. Numbers in the dataset generated by application of Eq. (1) were compared to published experimental values (Table 2) for proton transfers in protic solvents. The rate constants k_1 in the dataset are generally in good agreement with experimental values, within a factor of 10 M^{-1} s^{-1} (one log unit), for proton transfers spanning $k_1 = 10^6$ M^{-1} s^{-1} (log $k_1 = 6$).

Numbers in the dataset that were generated by application of Eq. (5) to proton transfers from carbon acids were compared to published experimental values (Table 3) in aqueous solvents or aqueous solvent mixtures from 20-25 °C. As with transfers between heteroatoms, most of the calculated log k_1 values were within an order of magnitude of the experimental values.

Values for log k_1 are reported to the hundredths decimal place (like most of the literature pK_a values), but the calculated log k_1 clearly lacks this high level of precision. For users training a system to distinguish plausible proton transfer from implausible proton transfers, the levels of accuracy afforded by the Eigen relationship are sufficient.

Data availability

All the datasets are available for download as a single zipped file in the PMechDB section at DeepRXN[47] (https://deeprxn.ics.uci.edu/pmechdb/download) and at figshare (https://doi.org/10.6084/m9.figshare.30875087) under a CC-BY license[46].

Code availability

The codes used to generate the proton transfer steps are publicly available at (https://github.com/rjmille3/combinatorial-proton-transfer.git).

Received: 1 September 2025; Accepted: 17 December 2025;
Published online: 10 January 2026

References

1. Clayden, J., Greeves, N. & Warren, S. *Organic Chemistry*. 2nd edn (Oxford University Press, 2012).
2. Reijenga, J., van Hoof, A., van Loon, A. & Teunissen, B. Development of Methods for the Determination of pK_a Values. *Anal. Chem. Insights.* **8**, 53–71 (2013).
3. Settimo, L., Bellman, K. & Knegtel, R. M. A. Comparison of the Accuracy of Experimental and Predicted pKa Values of Basic and Acidic Compounds. *Pharm Res* **31**, 1082–1095 (2014).
4. AAT Bioquest, Inc. Quest Database™ pKa and pKb Reference Table. AAT Bioquest. https://www.aatbio.com/data-sets/pka-and-pkb-reference-table (accessed on 15 October 2025).
5. Pahari, S., Sun, L. & Alexov, E. PKAD: a database of experimentally measured pKa values of ionizable groups in proteins. *Database* **2019**, baz024 (2019).
6. Huie, R. NDRL/NIST Solution Kinetics Database on the WEB, NIST Standard Reference Database 40, Data version 2003, National Institute of Standards and Technology, Gaithersburg, Maryland, https://kinetics.nist.gov/solution/ (accessed on 15 October 2025).
7. Reaxys, Elsevier Information Systems GmbH, https://www.reaxys.com, (accessed on 15 October 2025).
8. Musen, M. A. & van der Lei, J. Of Brittleness and Bottlenecks: Challenges in the Creation of Pattern-Recognition and Expert-System Models. *Mach. Intell. Pattern Recognit.* **7**, 335–352 (1988).
9. Lowe, D. *Chemical reactions from US patents* (1976–Sep 2016), https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873/1 (accessed 16 July 2025).
10. Manion, J. A. *et al*. NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.6.8, Data version 2024, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899-8320., https://kinetics.nist.gov (accessed on 16 July 2025).
11. Feeley, R., Seiler, P., Packard, A. & Frenklach, M. Consistency of a Reaction Dataset. *J. Phys. Chem. A* **108**, 9573–9583 (2004).
12. Smith, G. P. *et al*. GRI 3.0 Mechanism. Gas Research Institute. http://www.me.berkeley.edu/gri_mech/ (accessed on 16 July 2025).
13. Johnson, M. S. *et al*. RMG Database for Chemical Property Prediction. *J. Chem. Inf. Model.* **62**, 4906–4915 (2022).
14. Joung, J. F. *et al*. Reproducing Reaction Mechanisms with Machine-Learning Models Trained on a Large-Scale Mechanistic Dataset. *Angew. Chem. Int. Ed.* **63**, E202411296 (2024).
15. Klucznik, T. *et al*. Computational prediction of complex cationic rearrangement outcomes. *Nature* **625**, 508–515 (2024).
16. Chen, S., Babazade, R., Kim, T., Han, S. & Jung, Y. A large-scale reaction dataset of mechanistic pathways of organic reactions. *Sci. Data* **11**, 863 (2024).
17. Chen, J. H. & Baldi, P. No electron left behind: a rule-based expert system to predict chemical reactions and reaction mechanisms. *J. Chem. Inf. Model.* **49**, 2034–2043 (2009).
18. Kütt, A. *et al*. pK_a values in organic chemistry – Making maximum use of the available data. *Tetrahedron Lett* **59**, 3738–3748 (2018).
19. Tshepelevitsh, S. *et al*. On the Basicity of Organic Bases in Different Media. *Eur. J. Org. Chem.* **40**, 6735–6748 (2019).
20. Eigen, M. Fast Reactions and Primary Processes in Chemical Kinetics. *Nobel Symposium* **5**, 245–252 (1967).
21. Eigen, M. Proton Transfer, Acid-Base Catalysis, and Enzymatic Hydrolysis. Part I: ELEMENTARY PROCESSES. *Angew. Chem., Int. Ed. Engl.* **3**, 1–19 (1964).
22. Komp, E., Janulaitis, N. & Valleau, S. Progress towards machine learning reaction rate constants. *Phys. Chem. Chem. Phys.* **24**, 2692–2705 (2022).
23. Kayala, M. A., Azencott, C.-A., Chen, J. H. & Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **51**, 2209–2222 (2011).
24. Tavakoli, M., Mood, A., Van Vranken, D. & Baldi, P. Quantum mechanics and machine learning synergies: graph attention neural networks to predict chemical reactivity. *J. Chem. Inf. Model.* **62**, 2121–2132 (2022).
25. Fooshee, D. *et al*. Deep learning for chemical reaction prediction. *Mol. Sys. Des. Eng.* **3**, 442–452 (2018).
26. Crooks, J. E. Proton Transfer to and From Atoms Other Than Carbon. *Compr. Chem. Kinet.* **8**, 197–250 (1977).
27. Murdoch, J. R. Rate-equilibrium relations and proton-transfer reactions. *J. Am. Chem. Soc.* **94**, 4410–4418 (1972).

28. Pearson, R. G. & Dillon, R. L. Rates of ionization of pseudo acids. IV. relation between rates and equilibria. *J. Am. Chem. Soc.* **75**, 2439–2443 (1953).
29. Barroso, M., Arnaut, L. G. & Formosinho, S. J. Absolute rate calculations. Proton transfers in solution. *J. Phys. Chem. A* **111**, 591–602 (2007).
30. Ahrens, M.-L. & Maass, G. Elementary Steps in Acid-Base. *Catalysis. Proton Transfer Reactions in Aqueous Solutions. Angew. Chem., Int. Ed.* **7**, 818–819 (1968).
31. Reich, H. J. & Hondal, R. J. Why nature chose selenium. *ACS Chem. Biol.* **11**, 821–841 (2016).
32. Sander, T., Freyss, J., von Korff, M. & Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **55**, 460–473 (2015).
33. Bednar, R. A. & Jencks, W. P. Is hydrocyanic acid a normal acid? Proton transfer from hydrocyanic acid to bases and small inhibition of proton exchange by acid. *J. Am. Chem. Soc.* **107**, 7117–7126 (1985).
34. Silverstein, T. P. & Heller, S. T. pK_a Values in the Undergraduate Curriculum: What Is the Real pK_a of Water? *J. Chem. Educ.* **94**, 690–695 (2017).
35. Neils, T. L., Silverstein, T. P. & Schaertel, S. H_2O(aq) Does Not Exist: Critique of a Proof-of-Concept Derivation. *J. Chem. Educ.* **100**, 1676–1679 (2023).
36. Reich, H. J. "pK_a Values in Water" in *Hans Reich's Collection. Bordwell pK_a Table*. ACS Division of Organic Chemistry. https://organicchemistrydata.org/hansreich/resources/pka/#ka-water (Accessed 17 Jan 2025).
37. Guthrie, J. P. Hydration of thioesters. Evaluation of the free-energy changes for the addition of water to some thioesters, rate-equilibrium correlations over very wide ranges in equilibrium constants, and a new mechanistic criterion. *J. Am. Chem. Soc.* **100**, 5892–5904 (1978).
38. Guthrie, J. P., Barker, J., Cullimore, P. A., Lu, J. & Pik, D. C. The tetrahedral intermediate from the hydration of N-methylformanilide. *Can. J. Chem.* **71**, 2109–2122 (1993).
39. Brady, O. L. & Goldstein, R. F. "CCLIV.-The Isomerism of the Oximes. Part XXV. The Dissociation Constants of Some Isomeric Aldoximes. *J. Chem. Soc.* **129**, 1918–1924 (1926).
40. Bernasconi, C. F. & Ni, J. X. Proton Transfer from Carbon Acids to Carbanions. 2. Reaction of Phenylnitromethane with Carbanions, Enolate, and Nitronate Ions in 90% Me_2SO-10% Water. Carbon to Carbon or Carbon to Oxygen Proton Transfer? Test of the Marcus Equation. *J. Org. Chem.* **59**, 4910–4916 (1994).
41. Bernasconi, C. F., Wenzel, P. J., Keeffe, J. R. & Gronert, S. Intrinsic Barriers and Transition State Structures in the Gas Phase Carbon-to-Carbon Identity Proton Transfers from Nitromethane to Nitromethide Anion and from Protonated Nitromethane to *aci*-Nitromethane. An *ab Initio* Study. *J. Am. Chem. Soc.* **119**, 4008–4020 (1997).
42. Bernasconi, C. F. & Wenzel, P. J. Carbon-to-Carbon Identity Proton Transfers from Propyne, Acetimide, Thioacetaldehyde, and Nitrosomethane to Their Respective Conjugate Anions in the Gas Phase. An ab Initio Study. *J. Org. Chem.* **66**, 968–979 (2001).
43. Bernasconi, C. F. & Wenzel, P. J. Proton Transfers from Carbon Acids Activated by π-Acceptors. Changes in Intrinsic Barriers and Transition State Imbalances Induced by a Cyano Group. An ab Initio Study. *J. Org. Chem.* **68**, 6870–6879 (2003).
44. Bernasconi, C. F. *et al*. Kinetics of Proton Transfer from Cationic Carbon Acids in Water and Aqueous DMSO. Effect of Activating Groups and Solvent on Intrinsic Rate Constants. *J. Org. Chem.* **70**, 7721–7730 (2005).
45. Roszak, R., Beker, W., Molga, K. & Grzybowski, B. A. Rapid and Accurate Prediction of pK_a Values of C–H Acids Using Graph Convolutional Neural Networks. *J. Am. Chem. Soc.* **141**, 17142–17149 (2019).
46. Dashuta, A. E., Miller, R. J., Baldi, P., Sander, T. & Van Vranken, D. L. Plausible Proton Transfer Data Files. *figshare* https://doi.org/10.6084/m9.figshare.30875087 (2025).
47. Tavakoli, M. & Baldi, P. DeepRXN: A Platform of Deep Learning for Chemical Reactions. https://deeprxn.ics.uci.edu/ (Accessed 6 Mar 2025).
48. Chang, K. C. & Grunwald, E. Water Participation in Proton-Transfer Reactions of Glycine and Glycine Methyl Ester. *J. Phys. Chem.* **80**, 1422–1425 (1976).
49. Bednar, R. A. & Jencks, W. P. Direct Proton Transfer between HCN and Nitrogen and Oxygen Bases Direct Proton Transfer between HCN and N and O Bases. *J. Am. Chem. Soc.* **107**, 7126–7134 (1985).
50. Bernasconi, C. F. & Kittredge, K. W. Carbanion Stabilization by Adjacent Sulfur: Polarizability, Resonance, or Negative Hyperconjugation? Experimental Distinction Based on Intrinsic Rate Constants of Proton Transfer from (Phenylthio)nitromethane and 1-Nitro-2-phenylethane. *J. Org. Chem.* **63**, 1944–1953 (1998).
51. Bernasconi, C. F. & Montañez, R. L. Kinetics of Proton Transfer from Benzoylnitromethane and 1,2-Diphenyl-2-nitroethanone to Various Bases. Resonance, Inductive, Solvation, Steric, and Transition State Hydrogen-Bonding Effects on Intrinsic Rate Constants. *J. Org. Chem.* **62**, 8162–8170 (1997).
52. Bernasconi, C. F. & Paschalis, P. Kinetics of Ionization of 1,3-indandione in Me_2SO-Water Mixtures. Solvent Effect on Intrinsic Rates and Bronsted Coefficients. *J. Am. Chem. Soc.* **108**, 2969–2977 (1986).

Acknowledgements

This work was made possible by the generous support of NSF CHE 1955811.

Author contributions

A.D. and D.V.V. conceived the project. D.V.V. and P.B. supervised the project. T.S. provided the initial DataWarrior set. D.V.V. adapted data to the Eigen and Eigen-Bernasconi relationships. A.D. and R.M. refined the methods and performed the experiments. A.D. and D.V.V. analyzed the chemical results. A.D. and D.V.V. wrote the initial draft, and all authors participated in editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.L.V.V.