dr Łukasz Paweł Kozłowski
Institute of Informatics
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw

Stefana Banacha Street 2c
02-097 Warsaw, Poland

Summary of Professional Accomplishments

Prediction of isoelectric point and pK_a dissociation constants of proteins, peptides, and amino acids

Lukasz Pawel Kozlowski

Warsaw, December 10, 2021

UNIVERSITY OF WARSAW
UW MIM
Faculty of Mathematics, Informatics and Mechanics

1 *Name:*

Łukasz Paweł Kozłowski

2 Diplomas, degrees conferred in specific areas of science including the name of the institution which conferred the degree, year of degree conferment, title of the PhD dissertation

2013 **Ph.D. degree** (**Biochemistry**), Institute of Biochemistry and Biophysics of the Polish Academy of Science (Warsaw). Thesis title: *Integrated bioinformatics platform for protein analysis. Prediction of protein domain and intrinsic protein disorder.* Defended *cum laude* 07.05.2013. Thesis advisor: prof. dr hab. Janusz M Bujnicki. Reviewers: prof. dr hab. Andrzej Koliński, prof. dr hab. Zofia Szweykowska-Kulinska
https://depot.ceon.pl/handle/123456789/15446 (*in polish*)

2007 **Bachelor's degree (Computer Science)** Jan Kochanowski University (Kielce), Institute of Physics, Division Computer Science. Thesis title: *Determining amino acid composition in proteins using genetic algorithm.* Thesis advisor: dr I. Pardyka
https://depot.ceon.pl/hanle/123456789/3287 (*in polish*)

2006 **Master's degree (Genetics)**: Jan Kochanowski University (Kielce), Institute of Biology, Department of Biochemistry and Genetics. Thesis title: *Phylogenetic analysis of the linker histones in vertebrates*. Thesis advisor: prof. dr hab. Jan Pałyga
https://depot.ceon.pl/handle/123456789/3288 (*in polish*)

2004 **Bachelor's degree (Biology)**: Jan Kochanowski University (Kielce), Institute of Biology, Department of Biochemistry and Genetics. Thesis title: *The diversity of linker histones in vertebrates.* Thesis advisor: prof. dr hab. Jan Pałyga
https://depot.ceon.pl/handle/123456789/8793 (*in polish*)

3 Information on employment in research institutes or faculties/departments

2018 –	Assistant Professor (pol. *Adiunkt*), Institute of Informatics, Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland
2015 – 2017	Postdoctoral Researcher, Quantitative and Computational Biology Group headed by dr Johannes Soeding, Max Planck Institute for Biophysical Chemistry, Gottingen, Germany
2013 – 2015	Postdoctoral Researcher, Laboratory of Bioinformatics and Protein Engineering headed by prof. dr hab. Janusz M. Bujnicki, International Institute of Molecular and Cell Biology, Warsaw, Poland
2008 – 2013	PhD student, Laboratory of Bioinformatics and Protein Engineering headed by prof. dr hab. Janusz M. Bujnicki, International Institute of Molecular and Cell Biology, Warsaw, Poland

2007 – 2008	Bioinformatician, Laboratory of Bioinformatics and Protein Engineering headed by prof. dr hab. Janusz M. Bujnicki, International Institute of Molecular and Cell Biology, Warsaw, Poland
2006 – 2007	Intership, Akademia Swietokrzyska (currently Jan Kochanowski University), Kielce, Poland

4 Description of the achievements, set out in art. 219 para 1 point 2 of the Act.

4.1 Title of the achievement:

Prediction of isoelectric point and pK_a dissociation constants of proteins, peptides, and amino acids

4.2 Cycle of scientific articles related thematically

---

(1) Kozlowski, L.P. (2016) IPC – Isoelectric Point Calculator. *Biol. Direct*, **11**, 55.

https://doi.org/10.1186/s13062-016-0159-9 Highly Cited Paper

MNiSW: **100**; IF_{2016}: **3.472**; Citations: **245** (Google Scholar), **176** (Scopus), **174** (Web of Science)

---

(2) Kozlowski, L.P. (2017) Proteome-*pI*: proteome isoelectric point database. *Nucleic Acids Res.*, **45**, D1112–D1116.

https://doi.org/10.1093/nar/gkw978

MNiSW: **200**; IF_{2017}: **11.561**; Citations: **175** (Google Scholar), **104** (Scopus), **105** (Web of Science)

---

(3) Kozlowski L.P. (2021) IPC 2.0: prediction of isoelectric point and pK_a dissociation constants. *Nucleic Acids Res*, **49**, W285–W292 (Published: 27 April 2021).

https://doi.org/10.1093/nar/gkab295

MNiSW: **200**; IF_{2021}: **16.971**; Citations: **4** (Google Scholar), **3** (Scopus), **2** (Web of Science)

---

(4) Kozlowski, L.P. (2021) Proteome-*pI* 2.0: Proteome Isoelectric Point Database Update. *Nucleic Acids Res.*, (Published: 28 October 2021)

https://doi.org/10.1093/nar/gkab944

MNiSW**: 200;** IF_{2021}: **16.971**; Citations: **0** (Google Scholar), **0** (Scopus), **0** (Web of Science)

All above mentioned publications are available free of charge (Open Access)

In the case of MNiSW points, for consistency, only the new act has been used (even if the publication had been published before 2017)

Highly Cited Paper *– top 1% cited in Web of Science in given discipline*

$SIF_{2\text{-years}}$**: 48.975** (average **12.244**)

Citations of the cycle: **424** (Google Scholar), **283** (Scopus), **281** (Web of Science)

4.3 Discussion of the scientific goals of the above work, the results and their possible applications

Introduction

The charge of a protein is one of its key physicochemical characteristics. It is related to the **dissociation constant pK_a** which is a quantitative measure of the strength of an acid in solution. For proteins and peptides, the ionizable groups of seven charged amino acids should be taken into account: histidine (imidazole side chains), glutamate (γ-carboxyl group), aspartate (β-carboxyl group), cysteine (thiol group), lysine (ϵ-ammonium group), tyrosine (phenyl group), and arginine (guanidinium group) (5). Furthermore, other groups can possess the charge, such as the amine and carboxyl-terminal groups of the polypeptide chain and the post-translational modifications (PTMs) that carry charged groups (e.g. phosphorylation and N-terminal acetylation). If known, the pK_a values of charged groups can be used to calculate the overall charge of the molecule at a given pH and to estimate the **isoelectric point (*pI*, IEP)**, that is, the pH at which there is an equilibrium of positive and negative charges and therefore the total net charge of the molecule is equal to zero (6).

Both pK_a and the isoelectric point are used in numerous techniques, such as two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) (7, 8), capillary isoelectric focusing (9), crystallization (10), and mass spectrometry (11, 12). It should be stressed that experimental measurements of isoelectric points (SWISS-2DPAGE (13) and PIP-DB (14)) and pK_a values (PKAD database (15)) are very limited (not more than a few thousand records), but the development of computational methods for predicting these features is possible.

Prediction of isoelectric point

The simplest approach for the computational prediction of the isoelectric point is to use the Henderson–Hasselbalch equation (6), where the charge of a macromolecule at a given pH is the sum of the negative and positive charges of the individual amino acids, given by Equations 1 and 2, respectively.

$$\sum_{i=1}^{n} \frac{-1}{1+10^{pKn-pH}} \quad \text{(Eq. 1)}$$

where pK_n is the acid dissociation constant of the negatively charged amino acid.

$$\sum_{i=1}^{n} \frac{1}{1+10^{pH-pKp}} \quad \text{(Eq. 2)}$$

where pK_p is the acid dissociation constant of the positively charged amino acid.

As one can easily notice, the only variables are pK_a values (namely pK_n and pK_p), and, by iteratively changing the pH, it is possible to find the isoelectric point. Therefore, in a gross approximation, the isoelectric point can be estimated by counting the number of charged amino acids in the protein/peptide sequence. Admittedly, in reality, the situation is more complicated than this, as many proteins are chemically modified (e.g. amino acids can be phosphorylated, methylated, acetylated), which can change their charge. For instance, the occurrence of cysteines (negative charge), which may oxidize and lose charge when they form disulfide bonds in the protein, can be problematic. Moreover, one must consider the charged residue's exposure to solvent, charge–dipole interactions (hydrogen bonds), dehydration (the Born effect), and charge–charge interactions. Nevertheless, the most critical factors are pK_a values, and these can be obtained experimentally. However, many different pK_a sets have been reported, depending on the experimental setup (e.g. the buffer, amino acids surrounding the charged group). The most commonly used pK_a values for the ionizable groups of proteins are presented in **Table 1**.

The isoelectric point calculation algorithm is simple and can be summarized as finding the charge of zero (NQ) given partial sums related to the number of charged amino acids (Equation 3):

$$NQ=QN1+QN2+QN3+QN4+QN5+QP1+QP2+QP3+QP4 \quad \textit{(Eq. 3)}$$

where QN1 is carboxyl-terminal charge, QN2 is aspartic acid charge, QN3 is glutamate charge, QN4 is cysteine charge, QN5 is tyrosine charge, QP1 is N-terminal charge, QP2 is histamine charge, QP3 is lysine charge, and QP4 is arginine charge, as calculated using Equations 1 and 2.

As stated previously, this can be done iteratively by setting the pH to 0 and then changing the pH by 0.01 (or any other precision). Although the formula is fairly simple and can be computed relatively fast, the iterative approach is ineffective from the algorithmic point of view. The problem can be solved more effectively using a bisection algorithm which in each iteration halves the search space (initially, the pH is set to 7) and then moves higher or lower by 3.5 (half of 7) depending on the charge. In the next iteration, the pH is changed by 1.75 (half of 3.5), and so on. This process is repeated until the algorithm reaches the desired

precision. The bisection algorithm can improve the speed of isoelectric point estimation by 3–4 orders of magnitude, and usually, after only approximately a dozen iterations, the algorithm converges with 0.001 precision.

When the Applicant started working on this problem, there were over a dozen different experimentally derived pK_a sets and methods using them. However, nobody performed systematic studies to assess which was better for proteins or peptides (some attempts to build more sophisticated methods, such as those using genetic algorithms (16), artificial neural networks (17), and support vector machines (18), should be acknowledged here). Therefore, the very first step was to assemble the datasets that could be used for benchmarking the pK_a sets' usefulness for isoelectric point prediction (**Tables 2** and **3**).

In 2015, the first data from the literature were gathered (**Table 2**). After combining various sources and cleansing the data of duplicates and obvious errors, it was possible to compose two major datasets: one for proteins and one for peptides. The protein dataset was derived from SWISS-2DPAGE (13) and PIP-DB (14): 2,324 proteins in total. The peptide dataset was basad on two studies (Gauci et al. 2008 (19) and Heller et al. 2005 (20)): 16,882 peptides in total. The separation of the data into two datasets was necessary bacause the tasks of predicting of isoelectric point for proteins and peptides differ significantly. The intact proteins (analyzed, for instance, with 2D-PAGE) have multiple charged residues that can possess many PTMs, and in general the number of charged groups is high; therefore, the isoelectric point estimation using a simple model, in which merely the number and type of charged residues are considered, would be fraught with high risk. The contrary is true if we consider short peptides (for improved resolution in mass spectrometry, the proteins are digested by trypsin or any other protease into short fragments), for which the number of charged residues is limited and the macromolecules lack any 3D structure.

Subsequently, having established the datasets with experimentally verified isoelectric points for proteins and peptides, it was possible to turn the problem around and design new, computationally optimized pK_a sets. In practice, this can be brought down to finding nine variables (seven charged groups of amino acids and the COO^- and NH^+ terminal groups). Checking all possible combinations is not manageable, as even for nine variables in a pH range of 3 (i.e. ±1.5 pH units of the average for a given amino acid pK_a) with 0.01 precision gives 1.97×10^{22} possibilities. However, there is a huge number of optimization algorithms suitable for these tasks. In this particular case, basinhopping optimization with a truncated Newton algorithm was used (21, 22), as implemented in SciPy (*scipy.optimize.basinhopping*

and *scipy.optimize.minimize(method='TNC')*) (23). In the nutshell, the basinhopping algorithm is an iterative search procedure, with each cycle composed of the following steps: random perturbation of the coordinates; local minimization; and acceptance or rejection of the new coordinates based on the minimized function value of the Metropolis criterion of the standard Monte Carlo algorithm. As an initial 'seed,' previously published pK_a values were used. To limit the search space, a truncated Newton algorithm was used with 2-pH-unit bounds for pK_a variables (e.g. if the starting point for the pK_a for cysteine was 8.5, the solution was allowed in the interval [6.5, 10.5]). The resulting pK_a sets are presented in **Table 1** (IPC_protein and IPC_peptide sets implemented in Isoelectric Point Calculator). As shown in **Table 4**, the new optimized pK_a sets yielded improved accuracy in all performance metrics; for instance, the root mean square deviation (RMSD) of 0.874 for IP_protein versus 0.934 for Toseland and of 0.251 for IPC_peptide versus 0.255 for Solomon.

From the algorithmic point of view, isoelectric point prediction based on the Henderson–Hasselbalch equation with optimized pK_a sets can be considered to be a highly simplified approach. Therefore, in the second attempt around 2020, more advanced machine learning approaches were used. First, other optimization algorithms were tested. Here, differential evolution was used instead basinhopping, as it performed significantly better (*scipy.optimize.differential_evolution(popsize=50)* (24); the resulting pK_a sets are denoted IPC2_protein and IPC2_peptide in **Table 1**). In the next step, the machine learning approach adopted differed depending on the task. In the case of proteins, there was only a relatively small set of 2,324 experimentally validated measurements, so only a simple algorithm could be used. As there were already multiple methods based on the Henderson–Hasselbalch equation with different pK_a sets, it was straightforward to design an ensemble method (here, support vector regression [SVR] with radial basis function [RBF] kernel and *GridSearchCV* parameter optimization; *sklearn.svm.SVR*) (25). The input vector, in this case, was composed of 19 isoelectric point values predicted by these methods (IPC2.protein.svr.19 and IPC2.protein.svr.19 in **Table 5**). For peptides, a larger dataset (119,092 cases) was available than for proteins, so algorithms that were more data-hungry could be used. Here, it was possible to start from simple dense networks (multi-layer perceptrons [MLPs]) with different numbers of dense layers and neurons interconnected with dropout and with different activation layers (preferably *selu* and *elu*). Subsequently, increasingly advanced architectures (e.g. VVG16-like, Inception-like) were tested. After thorough tests, the final architecture for peptide isoelectric point prediction was based on the stacking of separable convolution layers (**Figures 4** and **5**).

The input was integrated as a four-channel 'image.' In the first channel, the sequence was stored in one-hot-encoding format. The length of the peptide was up to 60 amino acids (with padding if needed). There were 20 amino acids plus X for unknown and 0 for padding, giving 22 in total. In the second channel, the most informative features from AAindex (26) were encoded (univariate feature selection with regression [*f_regression*] and mutual information [*mutual_info_regression*] with recursive feature elimination (*RFE*) from the Scikit-learn package) (27)). Both channels stored the information by amino acids. In the third and fourth channels, the amino acid counts and predictions from methods based on the Henderson–Hasselbalch equation with pK_a sets were stored. The scalars in these two channels were extended into vectors (an individual row corresponds to a single prediction; e.g. prediction by Dawson*60). This made it possible to share information about isoelectric point predictions across many filters during the convolution. The input shaped as described was processed by separable convolution filters (the use of separable filters was crucial here), followed by average pooling (here, better than the frequently used *maxpooling*). The initial filters had a size of 22 × 5 to allow efficient amino-acid-related motif discovery in the first and second channels. After two rounds of convolution and pooling, everything was flattened and processed by three dense layers. In all layers, *selu* activation was used (**Figure 5**). The final model, as summarized in **Table 6**, provided superior accuracy on all benchmark datasets.

Prediction of pK_a dissociation constants

The prediction of pK_a dissociation constants is a different issue, involving many other problems. First, the dataset of experimentally measured pK_as is quite small (PKAD database (15), 1337 entries; **Table 3**), and, until this work, there were no other methods that could predict pK_a dissociation constants directly from the sequence (the methods such as MCCE (28), H++ (29), Propka (30), and Rosetta pKa (31) require the protein structure or model with 3D coordinates of the atoms). Therefore, the proposed approach was entirely novel. In the case of the charged groups for which the prediction of pK_a dissociation constants is required, the focus was on the single charged amino acid and the few surrounding it. Therefore, information related to kmers of different sizes was used. With the increasing size of the kmer (from 3 to 15), the sequence (one-hot encoding) and the amino acid scores for the most informative features from AAindex were used. This information was used later as an input for the MLP unit (three dense layers separated by dropout layers). Next, to boost the performance, an ensemble of nine models was used to build the final SVR model. To benchmark the resulting model(s), the same testing set as for the Rosetta pKa program was

used (**Table 7**). Although the IPC2_pKa model was not always better than Rosetta (for histidine, Rosetta pKa achieved an RMSD of 0.82, whereas IPC2_pKa achieved 0.85; for tyrosine, it was 0.78 vs 0.84 respecively), on average the pK_a scores were improved significantly from the RMSD of 0.83 (Rosetta site repack model) to 0.58 for IPC2_pKa. More significantly, the prediction was possible based entirely on the sequence, and it was very fast (the structure-based methods need hours to make predictions for a single protein, whereas the model presented here can perform predictions in fractions of a second).

Databases of predicted isoelectric points and pK_a dissociation constants

In the past, much work was put into creating databases with an experimentally measured isoelectric points (13, 14) and pK_a dissociation constants (15), yet none of these databases contained more than 5,000 values, which is very few compared with the protein sequence data currently available (counted in hundreds of millions). Therefore, the Proteome-*pI* and Proteome-*pI* 2.0 databases were an attempt to decrease this gap. Furthermore, the design of fast models like IPC2_protein and IPC2_pKa enabled proteome-wide predictions and enriched our knowledge regarding isoelectric points and charges of proteins in a high-throughput manner.

Before the creation of the *Proteome-pI* database, there was only one significant effort in this respect. Kiraga and co-workers in 2007 analyzed 1,784 proteomes using one algorithm (the Bjellqvist method, as implemented in Compute pI/Mw tool at the ExPASy server) (32).

However, this was a one-time analysis conducted over a decade ago, and the raw data behind it are not directly available. Consequently, in 2017, the Applicant developed the Proteome-*pI* database (http://isoelectricpointdb.org), which contains the predictions for 21,721,250 sequences from 5,029 so called reference proteomes from UniProt database (33) with isoelectric points predicted using 18 algorithms. After developing IPC 2.0, consisting of improved models for isoelectric point prediction and a novel, sequence-based pK_a dissociation constant predictor, it was possible to update the database (Proteome-*pI* 2.0, http://isoelectricpointdb2.org), which now contains predictions for 20,115 proteomes (61,329,034 proteins) using 21 algorithms (for more details, see **Tables 8–11**).

Another significant qualitative improvement presented in the Proteome-*pI* 2.0 database was the availability of pK_a dissociation constant predictions. As mentioned previously, methods other than IPC2_pKa need protein structures or at least a decent protein model, so even very recent efforts have been limited to the Protein Data Bank. Using their structure-based

algorithm, PypKa, Reis and co-workers made predictions of isoelectric points and pK_a values with increased throughput (34). The resulting database, pKPDB, contained predictions for ~70% of Protein Data Bank structures with ~10 million pK_a and ~120,000 isoelectric points. In comparison, Proteome-*pI* 2.0 contains 5.38 billion pK_a dissociation constant predictions for proteins from 20,115 proteomes divided into major kingdoms (Viruses, *Archaea*, Bacteria, and *Eukaryota*). Therefore, the researcher is not limited to the analysis of single proteins and can ask more interesting questions than it was possible in the past. Individual proteome statistics allow the comparison of the distribution of isoelectric points, which can be significantly biased in some organisms (**Figure 6**). For example, *Archaea* have the smallest proteins (except for viruses), but the isoelectric point of the proteome can differ greatly among individual species. This may be because *Archaea* are known for living in extreme environments (e.g. low or high pH), which affects the range of isoelectric points in their proteomes (**Figures 6** and **7**). Similarly, high-throughput predictions allow the investigation of the distribution of pK_a dissociation constants for specific charged groups (**Figure 8**). Here, it is worth mentioning that only histidine and, to a smaller extent, N-terminal pK_a predictions are normally distributed, whereas glutamate, aspartate, and C- terminal pK_a predictions have a skewed distribution with a long tail towards the high pH values. The opposite is true for tyrosine and lysine pK_a predictions (the skew is towards low pH values). This may indicate that, although most of the pK_a predictions are focused on model pK_a values (such as presented in **Table 1**), frequently this may not be the case, and they may change significantly, depending on the surrounding amino acids' charges.

Additionally, Proteome-*pI* 2.0 contains the prediction of isoelectric points for *in silico* digests of proteomes (9.58 billion peptides) with the five most commonly used proteases (trypsin, chymotrypsin, trypsin + LysC, LysN, ArgC), facilitating bottom-up proteomics analyses.

Finally, both databases, Proteome-*pI* and Proteome-*pI* 2.0, allow downloading predictions of isoelectric point done with multiple methods for all major protein sequence databases (the last update contains the predictions for the following databases: Swiss-Prot: ~561,000 (35, 36), PDB: ~601,000 (37), UniProtKB/TrEMBL: ~219 million (33), *nr*: 409 million (38) sequences).

User statistics of the Isoelectric Point Calculator (IPC), IPC 2.0, Proteome-*pI*, and Proteome-*pI* 2.0

The number of users that benefit from the work presented in the cycle of publications is hard to estimate but can be considered to be substantial. First, IPC and IPC 2.0 are computer programs that can be run both as standalone programs and via a web interface. In the latter case, some statistics are available (**Table 12**). The oldest version (IPC 1.0) was released in December 2015, and the webpage has since been visited by over 225,000 users. The upgraded version, IPC 2.0, is a relatively young web service and has had only ~6,500 visitors to date. It is impossible to estimate the number of standalone version users. The programs can be downloaded as publication supplements, directly from web server pages, Python Package Index (PyPI), or RePOD repository. Moreover, IPC has been integrated into numerous programs (e.g. Rapid Peptides Generator (39), *idpr* Bioconductor package , Pep-Calc (40)). In part, the popularity of the package may be due to its liberal licensing (in the public domain). This gives other researchers great freedom, and, although giving credit is not required, the program has been cited well in publications, books, websites, and even patents (e.g. TW202019466A). In contrast, the databases are available mainly from websites, and, although the number of visitors or publication readers is much smaller, this does not mean that the Proteome-*pI* database is not used or cited.

The IPC and IPC 2.0 tools can be considered to be useful for many scientists (the isoelectric point is a fairly standard, yet useful, feature of the protein), and the Proteome-*pI* database is a highly specialized resource used extensively by experts. Although Proteome-*pI* was designed for biologists focused on a single model organism, quite unexpectedly some high-throughput follow-up studies that use a huge proportion of or even the entire Proteome-*pI* data for specialized analyses can also be noted (e.g. high-throughput analysis of specific taxons, such as plants (41), fungi (42), and groups of interacting proteins (43)).

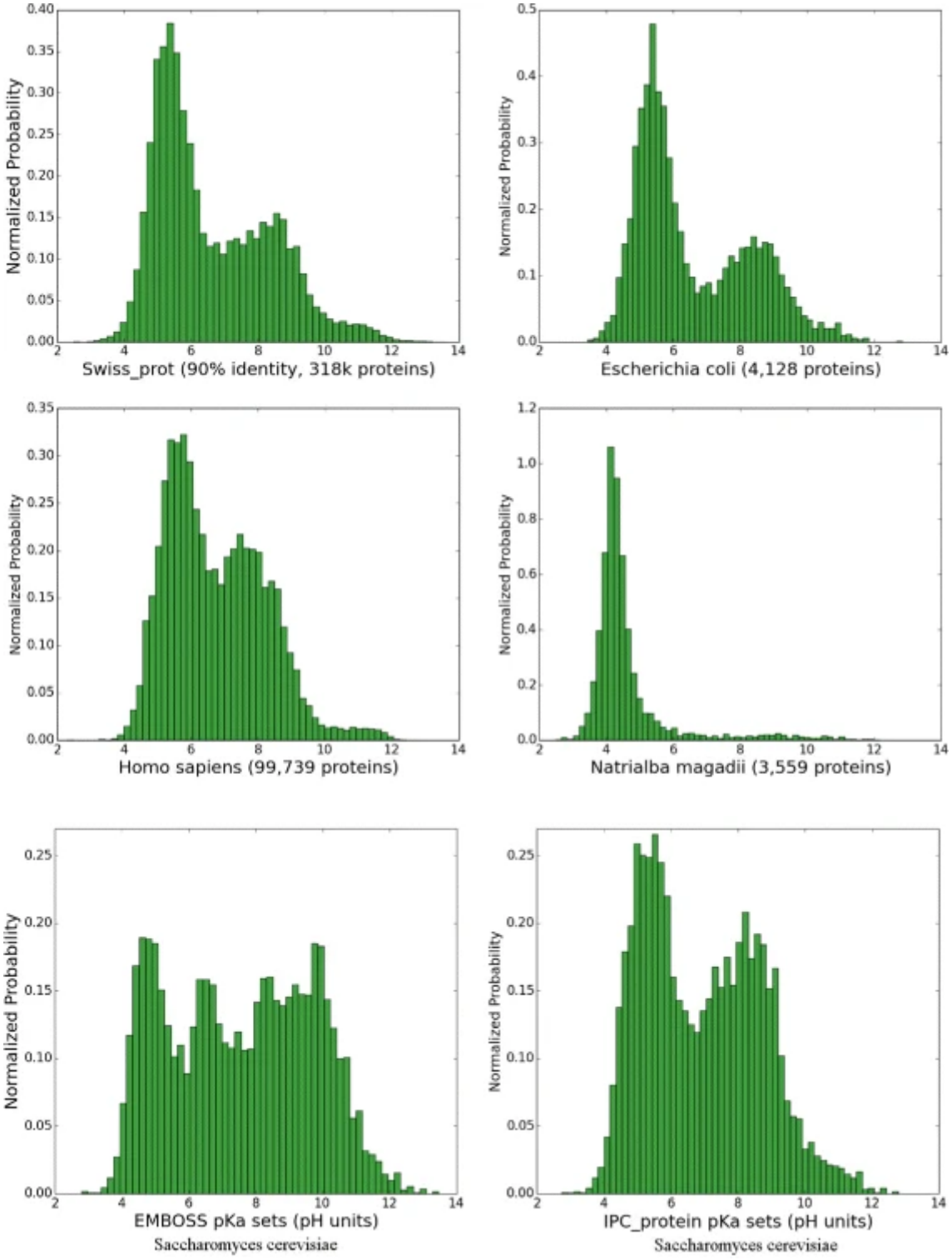

Figure 1. Histograms of the isoelectric points of proteins. The top and middle panels were calculated using the IPC_protein pK_a set and represents the *pI* distribution (in 0.25-pH-unit intervals) in the SwissProt database, human proteome, *Escherichia coli*, and extreme halophilic archaeon *Natrialba magadii*. The bottom two panels present the isoelectric points of the yeast proteome (6,721 proteins) calculated using the EMBOSS pK_a set (as presented in the *Saccharomyces* Genome Database (44)) (left) and the IPC_protein pK_a set (right) for comparison. Reproduced from Kozlowski 2016.

Figure 2. Example output of the Isoelectric Point Calculator for the *Mycoplasma genitalium* G37 proteome (476 proteins). The scatter plot with the predicted isoelectric points versus molecular weight for all proteins is presented at the top. Below this, for individual proteins, *pI* predictions based on different pK_a sets are presented alongside the molecular weight and amino acid composition. Reproduced from Kozlowski 2016.

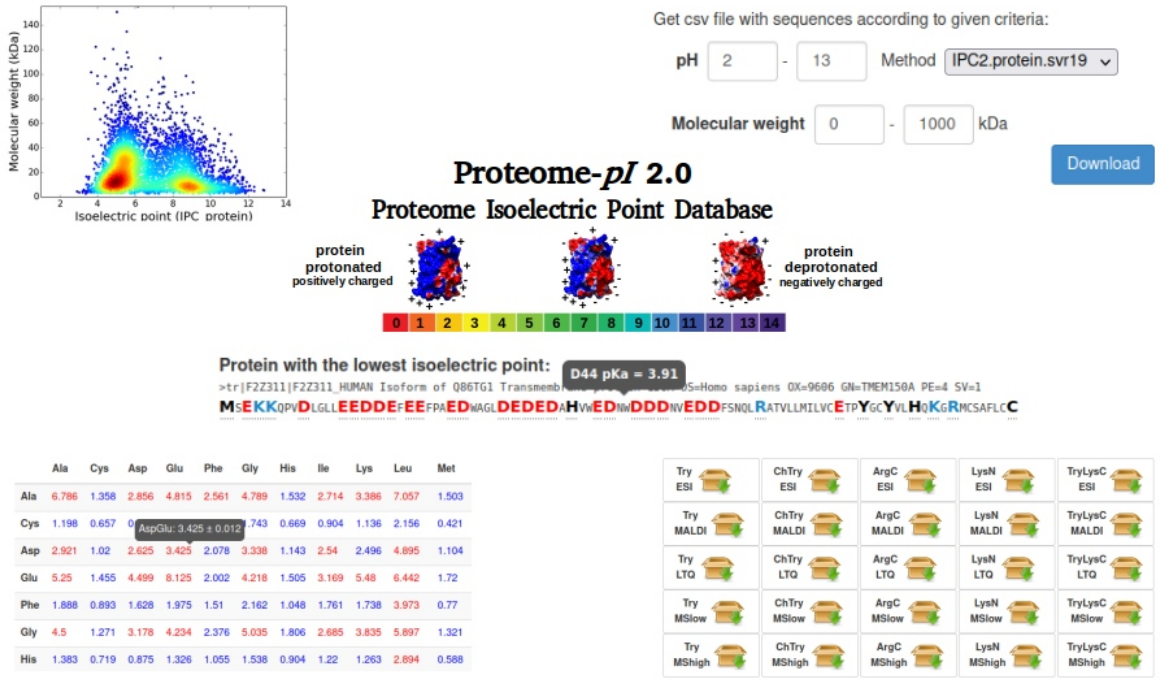

Figure 3. An overview of the Proteome-*pI* 2.0 database. Isoelectric points and molecular weights for individual proteins from 20,115 proteomes are visualized on virtual 2D-PAGE plots (top left) and can be retrieved according to the predictions of one of 21 algorithms (top right). The data for individual proteins are accompanied by dissociation constant (pK_a) predictions (middle). The proteomes are digested *in silico* by one of the five most commonly used proteases (trypsin, chymotrypsin, trypsin+LysC, LysN, ArgC) (bottom right). Additionally, auxiliary statistics are provided (e.g. di-amino acid frequencies) (bottom left). Reproduced from Kozlowski 2021b.

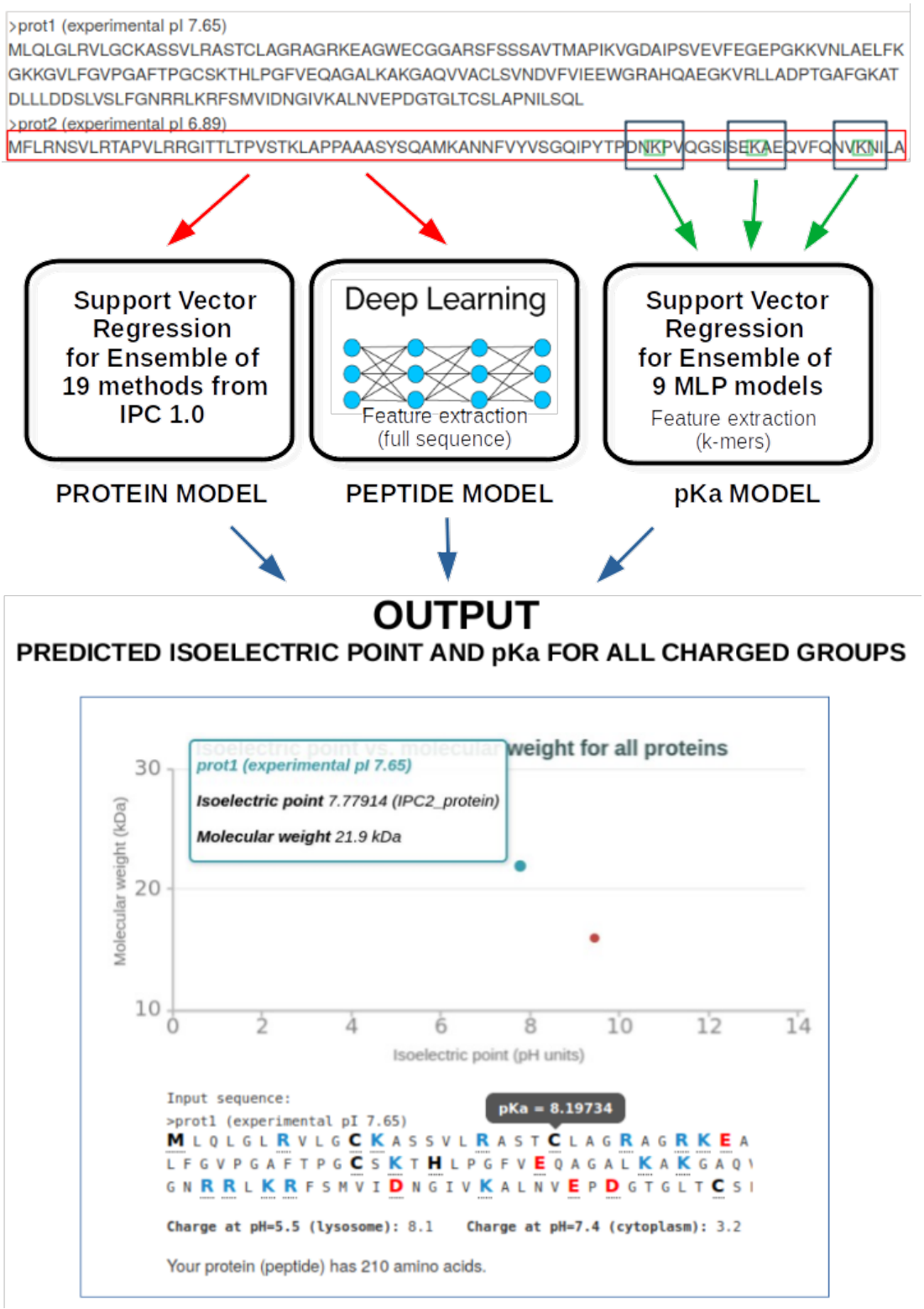

Figure 4. A schematic workflow of the Isoelectric Point Calculator 2.0. The input (single sequence or multiple FASTA query) is processed by three independent models. For proteins, support vector regression is used. For peptides, a deep learning model is used (see also Figure 5). For pK_a values, an ensemble of nine multilayer perceptron models based on kmers of different lengths is used.

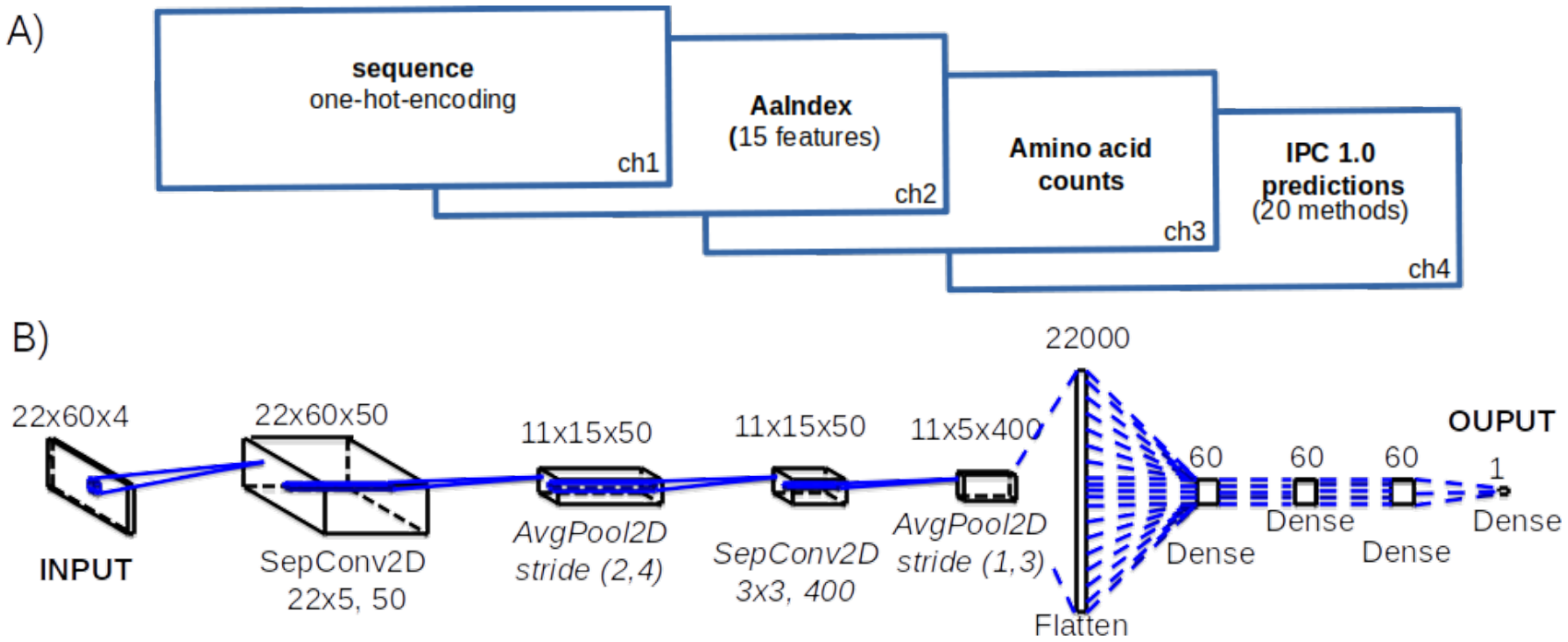

Figure 5. Deep learning architecture for peptide isoelectric point prediction. A) the input is integrated as a four-channel 'image.' The length of the peptide is up to 60 amino acids (with padding if needed; the width of 'image'). There are 20 amino acids plus 'X' for unknown and '0' for padding, giving 22 in total (the height of 'image'). In the first channel, the sequence is stored in one-hot-encoding format. In the second channel, there are the AAindex features (for details, see Supplementary Table S2; 15 rows, seven remaining padded). Both channels store the information by individual amino acids. In the third and fourth channels, the amino acid counts and Isoelectric Point Calculator (IPC) 1.0 predictions are stored. The scalars in these two channels are extended into vectors (an individual row corresponds to a single prediction; e.g. prediction by Dawson*60). This makes it possible to share the information about IPC 1.0 predictions across many filters during the convolution. B) The input is processed by separable convolution filters (the use of separable filters is crucial) followed by average pooling (here, better than *maxpooling*). The initial filters have a size of 22×5 to allow efficient amino-acid-related motif discovery in the first and second channels. After two rounds of convolution and pooling, everything is flattened and processed by three dense layers. In all layers, *selu* activation is used. Reproduced from Kozlowski 2021b.

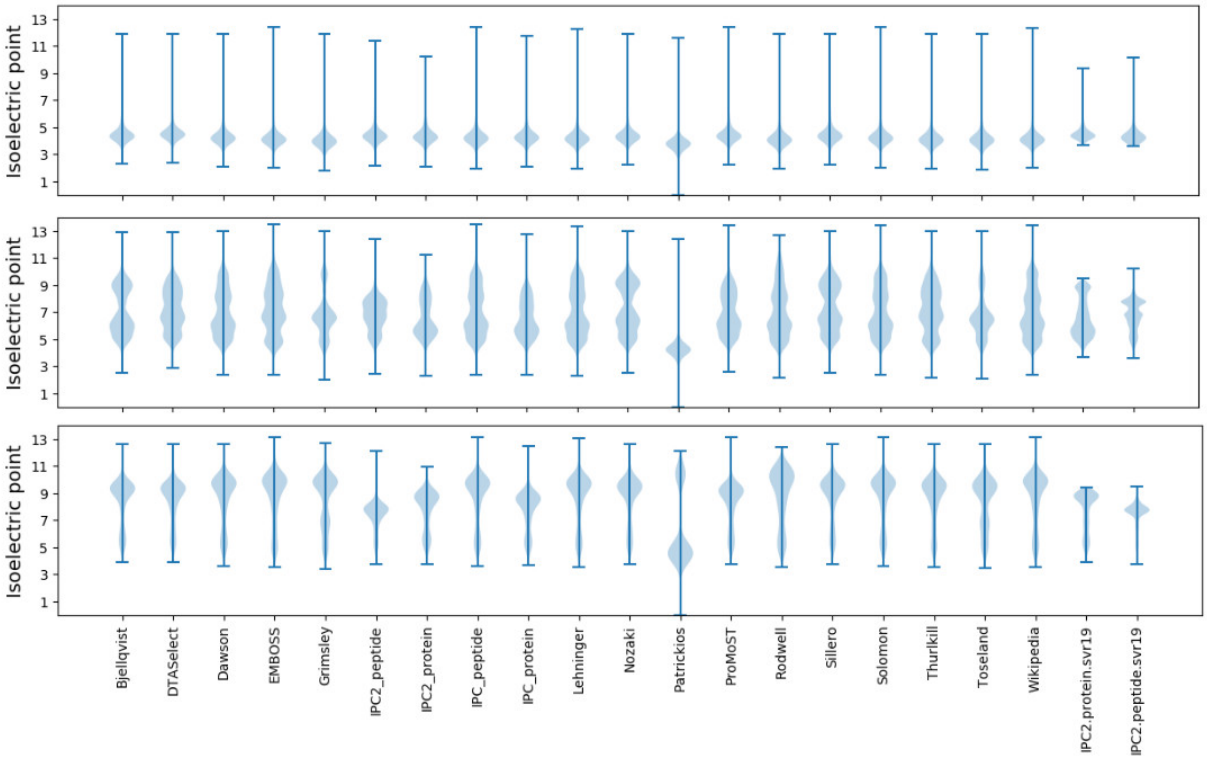

Figure 6. Isoelectric point predictions according to different methods. (Top) *Natronolimnobius baerhuensis*: an archaeon living in soda lakes; (middle) *Danio rerio* (expected, bimodal distribution of isoelectric points); (bottom) *Methanothermus fervidus*: a thermophilic methanogen. Reproduced from Kozlowski 2021b.

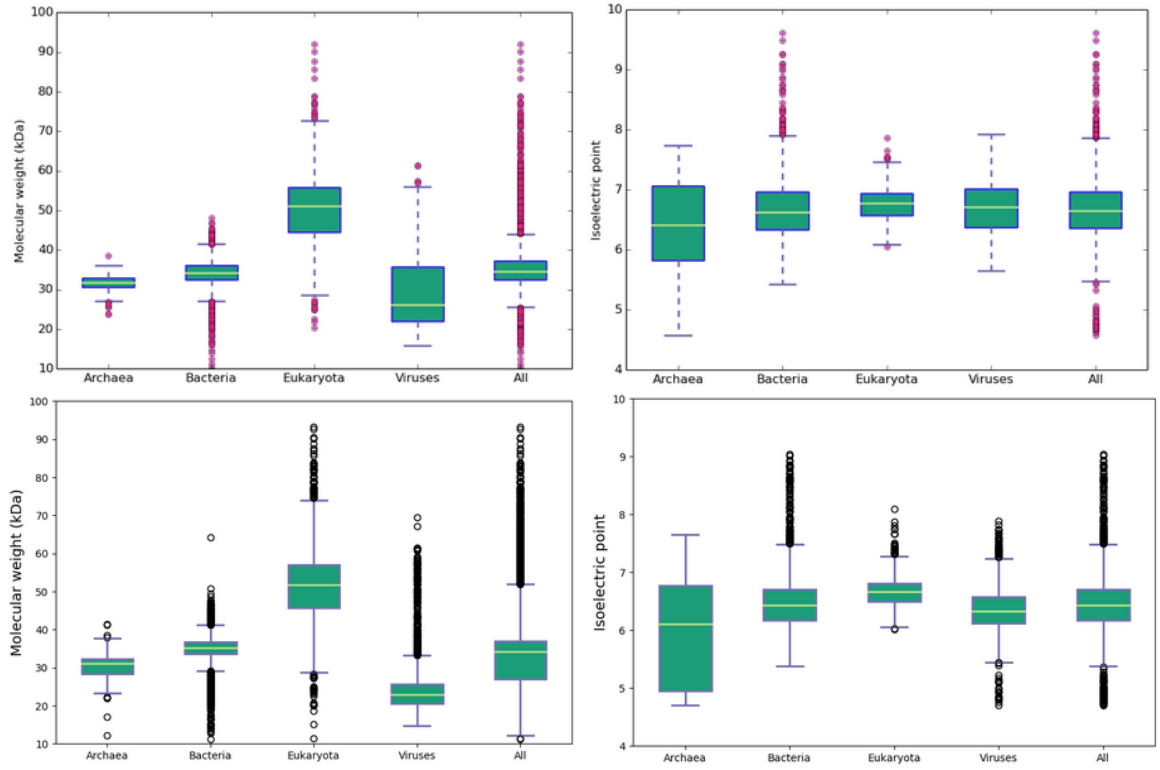

Figure 7. Isoelectric points and molecular weights across the kingdoms of life. (Top) Data from 135 archaea, 127 viruses, >50 proteins, 3,775 bacteria, and 614 eukaryote proteomes; the plots are reproduced from Proteome-*pI* (Kozlowski 2017). (Bottom) Data from the proteomes of 331 archaea, 4,046 viruses, 8,105 bacteria, and 1,612 eukaryotes with ≥50 proteins; the plots are reproduced from Proteome-*pI* 2.0 (Kozlowski 2021b).

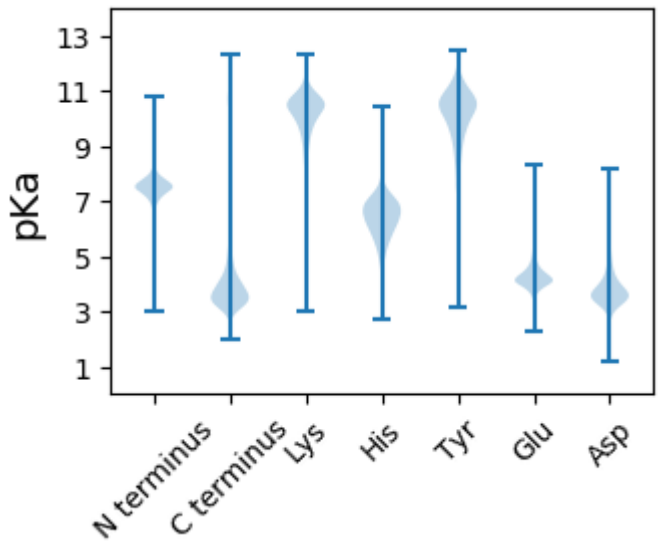

Figure 8. Dissociation constant (pK_a) predictions according to charge location. The plot is based on a random selection of 400 proteomes (100 viruses, 100 archaea, 100 bacteria, and 100 eukaryotes). Reproduced from Kozlowski 2021b.

Table 1. The most commonly used pK_a values for ionizable groups of proteins (a compilation of data from Kozlowski 2016 and Kozlowski 2021a).

	NH^+	COO^-	Asp	Glu	Cys	Tyr	His	Lys	Arg
EMBOSS	*8.6*	3.6	3.9	4.1	8.5	10.1	6.5	10.8	12.5
DTASelect	*8.0*	3.1	4.4	4.4	8.5	10.0	6.5	10.0	12.0
Solomon	*9.6*	2.4	3.9	4.3	8.3	10.1	6.0	10.5	12.5
Sillero	*8.2*	3.2	4.0	4.5	9.0	10.0	6.4	10.4	12.0
Rodwell	*8.0*	3.1	3.68	4.25	8.33	10.07	6.0	11.5	11.5
Patrickios	*11.2*	4.2	4.2	4.2	-	-	-	11.2	11.2
Wikipedia	*8.2*	3.65	3.9	4.07	8.18	10.46	6.04	10.54	12.48
Lehninger	*9.69*	2.34	3.86	4.25	8.33	10.0	6.0	10.5	12.4
Grimsley*	*7.7*	3.3	3.5	4.2	6.8	10.3	6.6	10.5	12.04
Toseland	*8.71*	3.19	3.6	4.29	6.87	9.61	6.33	10.45	12.0
Thurlkill	*8.0*	3.67	3.67	4.25	8.55	9.84	6.54	10.4	12.0
Nozaki	*7.5*	3.8	4.0	4.4	9.5	9.6	6.3	10.4	12.0
Dawson**	*8.2*	3.2	3.9	4.3	8.3	10.1	6.0	10.5	12.0
Bjellqvist	*7.5*	3.55	4.05	4.45	9.0	10.0	5.98	10.0	12.0
ProMoST	*7.26*	3.57	4.07	4.45	8.28	9.84	6.08	9.8	12.5
IPC_protein	*9.094*	2.869	3.872	4.412	7.555	10.85	5.637	9.052	11.84
IPC2_protein	*5.779*	6.065	3.766	4.497	7.890	11.491	5.492	9.247	10.223
IPC_peptide	*9.564*	2.383	3.887	4.317	8.297	10.071	6.018	10.517	12.503
IPC2_peptide	*7.947*	*2.977*	3.969	4.507	9.454	9.153	6.439	8.165	11.493

*Arg was not included in the study, and the average pK_a from all other pK_a sets was taken.

** NH2 and COOH were not included in the study, and they were taken from Sillero.

Note that Bjellqvist and ProMoST use different amounts of additional pK_a values (not shown), which take into account the relative position of the ionized group (whether it is located on the N- or C-terminus or in the middle).

Table 2. Detailed statistics for the available datasets used in the Isoelectric Point Calculator (IPC) 1.0. Reproduced from Kozlowski 2016, enriched with hyperlinks to the datasets.

Dataset	Initial no. of entries	No. of entries with sequence and *pI*	No. of entries after removing outliers	No. of entries after removing duplicates
Gauci et al.	5,758	5,758	NA	NA
PHENYX	7,582	7,582	NA	NA
SEQUEST	7,629	7,629	NA	NA
IPC_peptide	-	20,969	20,969	16,882 [25] [75]
SWISS-2DPAGE	2,530	1,054	1,029	982
PIP-DB	4,947	2,427	2,254	1,307
IPC_protein	-	3,481	3,283	2,324 [25] [75]

NA, not available; refers to the situation where the given dataset was not created because a merged version was used.

Note: all datasets presented in the table are available via hyperlinks; the final datasets were divided randomly into 75% training and 25% testing subsets (denoted as [75] and [25], respectively).

Table 3. Detailed statistics for the available datasets used in the Isoelectric Point Calculator (IPC) 2.0. Reproduced from Kozlowski 2021a.

Dataset	Entries	Details
IPC2_protein	2,324	The dataset consists of proteins derived from two databases: PIP-DB and SWISS-2DPAGE (13, 14). The outliers are defined at 0.5-pH-unit difference between the predicted and experimental isoelectric point threshold. The same protein dataset is used in IPC and IPC 2.0. Average protein size: 387 amino acids.
- **IPC_protein_25** (25% test set)	581	
- **IPC_protein_75** (75% training set)	1,743	
IPC2_peptide	119,092	The dataset consists of the peptides from HiRIEF high-resolution isoelectric focusing experiments from Branca et al. 2014 (12) and Johansson et al. 2019 (12). Merged dataset from seven independent experiments: 3.7–4.9 (8,713 peptides), 3.7–4.9 (7,361 peptides), 3.7–4.9 (35,595 peptides), 3–10 (23,975), 3–10 (15,000 peptides), 6–11 (36,827 peptides), 6–9 (38,057 peptides). Average peptide size: 14.6 amino acids.
- **IPC2_peptide_25** (25% test set)	29,774	
- **IPC2_peptide_75** (75% training set)	89,318	
IPC2_pKa	1,337	pK_a values from PKAD database (157 proteins). Due to the small number of samples, the test set and training set were built as follows: 260 pK_a values from 34 proteins used in the pK_a Rosetta method (15) were selected as a test set; the remaining samples from the PKAD database were used as the training set.
- **IPC2_pKa_25** (test set)	260	
- **IPC2_pKa_75** (training set)	1,079	

The full datasets were not used directly. First, the sequences were clustered (to remove duplicates and to average isoelectric points if multiple experimental data existed), then they were split randomly into 25% and 75% sets (test and training datasets, respectively). The training sets were used for the training and (hyper)parameter optimization, and the test sets were used only once to assess the final performance of the models. For individual datasets' sequences and experimental isoelectric points, see Supplementary Data 1 in Kozlowski 2021a.

Table 4. Prediction of isoelectric points using the 25% testing datasets in the Isoelectric Point Calculator (IPC) 1.0. Reproduced from Kozlowski 2016. A similar table for the 75% training datasets is available in Kozlowski 2016.

	Protein dataset				Peptide dataset		
Method	**RMSD**	**%**	**Outliers**	**Method**	**RMSD**	**%**	**Outliers**
IPC_protein	0.874	0	46	**IPC_peptide**	0.251	0	232
Toseland	0.934	14.9	52	Solomon	0.255	0.9	235
Bjellqvist	0.944	17.7	47	Lehninger	0.262	2.5	236
Dawson	0.945	17.8	56	EMBOSS	0.325	18.5	372
Wikipedia	0.955	20.5	55	Wikipedia	0.421	47.9	1467
Rodwell	0.963	22.8	58	Toseland	0.425	49.1	990
ProMoST	0.966	23.6	52	Sillero	0.428	50.3	1223
Grimsley	0.968	24.2	60	Dawson	0.435	52.9	1432
Solomon	0.970	24.8	58	Thurlkill	0.481	69.7	1361
Lehninger	0.970	25.0	59	Rodwell	0.502	78.4	1359
pIR	1.013	38.0	58	DTASelect	0.550	99.1	1714
Nozaki	1.024	41.3	56	Nozaki	0.602	124.3	1368
Thurlkill	1.030	43.4	61	Grimsley	0.616	131.4	1550
DTASelect	1.032	44.1	58	Bjellqvist	0.669	161.5	1583
pIPredict	1.048	49.4	56	pIPredict	1.024	493.6	2720
EMBOSS	1.056	52.3	69	ProMoST	1.239	873.4	2649
Sillero	1.059	53.2	63	pIR	1.881	4159.7	3358
Patrickios	2.392	3201.8	227	Patrickios	1.998	5479.1	2739
Avg_pI*	0.960	22.1	53	Avg_pI	0.454	59.6	1571

* Average from all pK_a sets without Patrickios (highly simplified pK_a set) and IPC sets. Note that the average *pI* is calculated on the level of individual protein or peptide; therefore, it does not represent the average of values presented in the table for the individual methods.

%: note that the pH scale is logarithmic with base 10; therefore, the percentage difference corresponds to pow(10, x), where x is equal to the delta of the RMSDs of two error estimates represented in pH units; for example, the % difference between Toseland and IPC_protein is pow(10, (0.934 – 0.874)).

Protein dataset: IPC_protein was trained on 1,743 proteins with 10-fold cross-validation (data in Table 2 in Kozlowski 2016); tested on 581 proteins not used for training (data in this Table 4).

Peptide dataset: IPC trained on 12,662 peptides with 10-fold cross-validation (data in Table 2 in Kozlowski 2016); tested on 4,220 peptides not used for training (data in this Table 4). Outliers correspond to the number of predictions for which the difference between the experimental *pI* and predicted *pI* was greater than the threshold of the mean standard error (MSE) of 3 for the protein dataset and MSE of 0.25 for the peptide dataset.

Table 5. Isoelectric point prediction accuracy on leave-out 25% datasets. Reproduced from Kozlowski 2021a.

Method	Protein dataset[a]				Method	Peptide dataset[b]			
	RMSD	MAE	R^2	Outliers[c]		RMSD	MAE	R^2	Outliers[c]
IPC2.protein.svr.19	0.8479	0.5906	0.5934	247	**IPC2.peptide.Conv2D**	0.2216	0.1216	0.9761	2691
IPC2_protein	0.8608	0.6052	0.5748	251	**IPC2.peptide.svr.19**	0.2299	0.1155	0.9743	2490
IPC_protein	0.8677	0.6109	0.5760	250	**IPC2_peptide**	0.2482	0.1394	0.9700	3179
ProMoST	0.9113	0.6444	0.5183	263	Bjellqvist	0.4051	0.2836	0.9204	11639
Toseland	0.9278	0.6537	0.5095	250	Nozaki	0.4083	0.2673	0.9191	9837
Dawson	0.9365	0.6586	0.4977	263	DTASelect	0.4235	0.2796	0.9130	10606
Bjellqvist	0.9369	0.6536	0.5005	260	Thurlkill	0.4466	0.2535	0.9033	7182
Wikipedia	0.9484	0.6795	0.4860	262	Sillero	0.4747	0.2696	0.8907	7607
Rodwell	0.9579	0.6762	0.4706	262	Dawson	0.4910	0.2642	0.8831	6698
Grimsley	0.9588	0.6953	0.4779	265	Wikipedia	0.5178	0.2974	0.8700	8326
Lehninger	0.9617	0.6783	0.4607	266	Grimsley	0.5264	0.3796	0.8656	15956
Solomon	0.9631	0.6746	0.4606	272	Rodwell	0.5855	0.3429	0.8337	9857
pIR	1.0148	0.7556	0.4161	315	Toseland	0.5860	0.3896	0.8335	13152
Nozaki	1.0164	0.7219	0.3980	288	EMBOSS	0.5971	0.3557	0.8271	11022
Thurlkill	1.0250	0.7573	0.3948	302	PredpI-iTRAQ8	0.6302	0.3503	0.8027	12059
DTASelect	1.0278	0.7798	0.3947	319	PredpI-TMT6	0.6365	0.3518	0.7988	12135
EMBOSS	1.0498	0.7757	0.3734	308	PredpI-plain	0.6480	0.3710	0.7913	12813
Sillero	1.0519	0.7694	0.3461	308	IPC_peptide	0.7459	0.4860	0.7302	13599
Patrickios	2.3764	1.8414	<0	517	Solomon	0.7518	0.4929	0.7259	13777
PredpI-TMT6	NA	NA	NA	NA	Lehninger	0.7697	0.5209	0.7127	15200
PredpI-plain	NA	NA	NA	NA	pIR	0.8529	0.7303	0.6387	27158
PredpI-iTRAQ8	NA	NA	NA	NA	ProMoST	1.1026	0.7562	0.4104	18513
					Patrickios	2.0172	1.3927	<0	22818

[a]Protein dataset consisting of 581 proteins (25% randomly chosen proteins, not used for the training or optimization); the same as for the Isoelectric Point Calculator (IPC) 1.0.

[b]Peptide dataset consisting of 29,774 peptides (25% randomly chosen peptides, not used for the training or optimization).

[c]The outliers were defined at 0.5- and 0.25-pH-unit differences between the predicted and experimental *pI* thresholds for the protein and peptide datasets.

NA: the PredpI program was designed only for peptides within the 3.7–4.9 pH range; therefore, for proteins, it returned 0 and could not be evaluated on the protein dataset.

New machine learning models developed in the IPC 2.0 study are shown in **bold**, and the first version of IPC (Kozlowski 2016) is underscored. Scores were calculated after 10-fold cross-validation. The table is sorted by RMSD. For individual methods' predictions, see Supplementary Data 2 in Kozlowski 2021a. For more details about the datasets, see Table 3.

Table 6. The effect of model architecture on the performance of isoelectric point prediction. Reproduced from Kozlowski 2021a.

Method	Peptide test dataset (30,279)				Method	Protein test dataset (581)			
	RMSD	MAE	R^2	Outliers		RMSD	MAE	R^2	Outliers
IPC2.peptide.Conv2D	0.2216	0.1216	0.9761	2691	IPC2.protein.svr.19	0.8466	0.5907	0.5965	247
IPC2.peptide.svr.19	0.2298	0.1155	0.9743	2490	IPC2_protein	0.8590	0.6052	0.5835	251
IPC2.peptide19	0.2376	0.1271	0.9726	2980	IPC_protein	0.8679	0.6109	0.5779	250
IPC2.peptide1320	0.2394	0.1245	0.9721	3055	ProMoST	0.9116	0.6443	0.5219	263
IPC2_peptide	0.2483	0.1394	0.9700	3179					
Bjellqvist	0.4051	0.2836	0.9204	11639					
IPC_peptide	0.7458	0.4860	0.7302	13599					

IPC2.peptide.Conv2D: an input layer is an 'image' (60 × 22 × 4). The rows correspond to amino acid register (up to 60, padded if necessary). The columns in the first channel correspond to amino acid sequence (60 × 22: 20 standard, 'X' for unknown amino acid, and '0' for padding). The second channel corresponds to the most informative features from AAindex (60 × 15). The third channel contains charged amino acid counts, and the predicted *pI* from other simple methods is stored (60 × 20) in the fourth channel. The scalars in the third and fourth channels are duplicated to form a 60-long vector. Then, the SeparableConvolution2D, AveragePooling2D, and Dense layers follow. For details of the model architecture, see Figure 5.

IPC2.peptide.svr.19: a support vector regression (SVR) model with 19 isoelectric points predicted by simple methods (those that use the Henderson–Hasselbach equation, including the IPC2_peptide model). The input was limited to *pI* values only, as adding other features worsened the SVR convergence and the prediction accuracy. SVR parameters were optimized by GridSearchCV (RBF kernel, C = 1,500, epsilon = 0.1293). Note that this model is better than the optimized version (IPC2_peptide), which means that SVR could learn from *pI*s predicted by other methods better than the optimization and even better than simple multilayer perceptron (MLP) models based on sequence alone (IPC2.peptide1320) – the same input as IPC2.peptide19. Additionally, the SVR model produces the fewest outliers.

IPC2.peptide19: a model that takes 19 isoelectric points predicted by simple methods as input (the same input as IPC2.peptide.svr.19). MLP model: *dense* (760, *selu*), *dense* (760, *softplus*), *dense* (190, *selu*), *dense* (1). The number of neurons and type of activation were optimized by *RandomizedSearchCV*.

IPC2.peptide1320: a model that takes one-hot-encoded sequence (a flat vector of 1320; 60 × 22) as input. MLP model: *dense* (1320, *softplus*), *dropout* (0.7), *dense* (60, *selu*), *dense* (30, *selu*), *dense* (1). The number of neurons and type of activation were optimized by *RandomizedSearchCV*.

IPC2_peptide: a simple model based on the optimization of pK_a values conducted similarly to the Isoelectric Point Calculator (IPC) 1.0 in 2016, but using a larger and more robust dataset (119,093 peptides, split into 75% for training and 25% for validation) and differential evolution instead of basinhopping. For individual pK_a values, see Table 1.

Bjellqvist: the best method for a peptide dataset developed by other researchers (based on a simple algorithm using pK_a values by Bjellqvist and the Henderson–Hasselbach equation; used in the Expasy Compute pI/MW tool).

IPC_peptide: a peptide model based on basinhopping optimization of pK_a values performed in 2016 (IPC 1.0).

IPC2.protein.svr.19: identical to IPC2.peptide.svr.19, but optimized with the protein dataset.

IPC2_protein: identical to IPC2_peptide, but optimized with the protein dataset.

IPC_peptide: a peptide model based on basinhopping optimization of pK_a values conducted in 2016 (IPC 1.0).

ProMoST: the best method for a protein dataset developed by other researcher (based on a simple algorithm using pK_a values and the Henderson–Hasselbach equation; 72-parameter model, including C- and N-terminal corrections for charges).

Table 7. Accuracy of pK_a prediction using the Rosetta pKa dataset. Reproduced from Kozlowski 2021a.

Method	Rosetta pK_a dataset[a]			Method	Rosetta pK_a dataset[a]		
	RMSD	MAE	Outliers[b]		RMSD	MAE	Outliers[b]
D (74; 3.45 ± 0.80)				**Y** (17; 10.89 ± 0.82)			
IPC2_pKa	**0.3883**	**0.2238**	**6**	**Rosseta (Site repack)**	**0.7750**	**0.6177**	**7**
Rosseta (Site repack)	0.8193	0.5824	27	Rosseta (Neighbor repack)	0.8370	0.6647	9
Rosseta (Ensemble average)	0.8413	0.5460	25	Rosetta (Standard)	0.9579	0.8000	9
Rosseta (Neighbor repack)	0.8676	0.6378	34	IPC2_pKa	0.9766	0.8261	10
Rosetta (Standard)	1.0651	0.8554	46	Rosseta (Ensemble average)	1.1892	0.9529	13
H (76; 6.58 ± 0.98)				**K** (22; 10.66 ± 0.52)			
Rosseta (Site repack)	**0.8247**	**0.6408**	31	**IPC2_pKa**	**0.2933**	**0.1909**	**2**
IPC2_pKa	0.8523	0.5105	**27**	Rosseta (Neighbor repack)	0.6216	0.5091	7
Rosseta (Neighbor repack)	0.8559	0.6487	32	Rosetta (Standard)	0.6498	0.5046	8
Rosseta (Ensemble average)	1.0244	0.7566	39	Rosseta (Site repack)	0.6705	0.5227	7
Rosetta (Standard)	1.2303	0.9961	50	Rosseta (Ensemble average)	0.7135	0.5364	6
E (71; 4.16 ± 0.80)				**All** (260*)			
IPC2_pKa	**0.3625**	**0.1951**	**7**	**IPC2_pKa**	**0.5762**	**0.3364**	**54**
Rosseta (Neighbor repack)	0.8744	0.5887	29	Rosseta (Site repack)	0.8262	0.6165	102
Rosetta (Standard)	0.8880	0.7324	38	Rosseta (Neighbor repack)	0.8332	0.6185	111
Rosseta (Site repack)	0.9303	0.6549	30	Rosseta (Ensemble average)	0.9207	0.6746	114
Rosseta (Ensemble average)	0.9317	0.6972	34	Rosetta (Standard)	1.0300	0.8296	151

[a]For the validation of *pK*$_a$, the dataset from Kilambi and Gray 2012 (31) was used (260* residues from 34 proteins). The numbers next to the residue type indicate the number of cases and the average pK_a value with standard deviation.

[b]The outliers are defined at 0.5-pH-unit differences between the predicted and experimental pK_a thresholds.

*The dataset consists of 260 instead of 264 residues due to parsing problems (four missing residues could not be mapped to the protein sequence, due to the wrong residue register). Scores were calculated after 10-fold cross-validation.

Table 8. General statistics of the **Proteome-*pI*** database (5,029 proteomes with 21,721,250 proteins in total). Reproduced from Kozlowski 2017.

	Number of proteomes	Total number of proteins	Mean number of proteins (±SD)	Mean size of proteins (±SD)	Mean mw of proteins (±SD)
Viruses	504	20,920	42 ± 89	297 ± 375	33 ± 42
Archaea	135	318,388	2,358 ± 920	283 ± 212	31 ± 23
Bacteria	3,776	12,082,903	3,200 ± 2,510	311 ± 240	34 ± 26
Eukaryote	614	9,299,039	15,145 ± 11,830	438 ± 429	49 ± 48
Eukaryote (major)	614	8,629,591	14,055 ± 9,899	434 ± 416	48 ± 46
Eukaryote (minor)	448	669,448	1,494 ± 5,130	495 ± 564	55 ± 63

Table 9. General statistics of the **Proteome-*pI* 2.0** database (20,115 proteomes with 61,329,034 proteins in total). Reproduced from Kozlowski 2021b.

	Number of proteomes	Total number of proteins	Mean number of proteins (±SD)	Mean size of proteins (±SD)	Mean mw of proteins (±SD)
Viruses	10,064	518,140	51 ± 85	237 ± 300	26.6 ± 33.2
Archaea	331	767,951	2,320 ± 1,263	278 ± 211	30.6 ± 23.1
Bacteria	8,108	30,290,647	3,736 ± 1,785	320 ± 246	35.1 ± 26.8
Eukaryote	1,612	29,752,296	18,457 ± 16,804	467 ± 471	52.1 ± 52.4
Eukaryote (major)	1,612	25,437,198	15,780 ± 11,138	438 ± 420	48.8 ± 46.7
Eukaryote (minor)	637	4,315,098	6,774 ± 14,244	638 ± 676	71.2 ± 75.4

Table 10. Amino acid frequency for the kingdoms of life in the **Proteome-*pI*** database. Reproduced from Kozlowski 2017.

Kingdom	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	Total amino acids
Viruses	6.61	1.76	5.81	6.04	4.25	5.79	2.15	6.53	6.35	8.84	2.46	5.41	4.62	3.39	5.24	7.06	6.06	6.50	1.19	3.94	6,150,189
Archaea	8.20	0.98	6.21	7.69	3.86	7.58	1.77	7.03	5.27	9.31	2.35	3.68	4.26	2.38	5.51	6.17	5.44	7.80	1.03	3.45	89,488,664
Bacteria	10.06	0.94	5.59	6.15	3.89	7.76	2.06	5.89	4.68	10.09	2.38	3.58	4.61	3.58	5.88	5.85	5.52	7.27	1.27	2.94	3,716,982,916
Eukaryota	7.63	1.76	5.40	6.42	3.87	6.33	2.44	5.10	5.64	9.29	2.25	4.28	5.41	4.21	5.71	8.34	5.56	6.20	1.24	2.87	3,743,221,293
All	8.76	1.38	5.49	6.32	3.87	7.03	2.26	5.49	5.19	9.68	2.32	3.93	5.02	3.90	5.78	7.14	5.53	6.73	1.25	2.91	7,555,843,062

Similar statistics for all 5,029 proteomes included in Proteome-*pI* are available online on individual subpages. For di-amino acid frequencies, see Supplementary Table S2 in the Proteome-*pI* publication (Kozlowski 2017) or http://isoelectricpointdb.org/statistics.html

Table 11. Amino acid frequency for the kingdoms of life in the **Proteome-*pI* 2.0** database. Reproduced from Kozlowski 2021b.

Kingdom	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	Total amino acids
Viruses	7.81	1.29	6.20	6.46	3.91	6.72	1.96	6.05	6.24	8.28	2.51	4.99	4.25	3.62	5.31	6.47	6.14	6.66	1.42	3.71	122,870,810
Archaea	8.95	0.90	7.00	7.94	3.65	7.84	1.86	6.03	4.18	9.11	2.14	3.36	4.36	2.48	5.83	6.12	5.84	8.16	1.06	3.18	213,285,886
Bacteria	10.64	0.90	5.67	6.06	3.76	8.01	2.08	5.52	4.22	10.12	2.31	3.35	4.82	3.49	6.18	5.75	5.58	7.42	1.31	2.81	9,693,905,784
Eukaryota	7.38	1.85	5.34	6.55	3.79	6.35	2.50	4.94	5.64	9.38	2.27	4.13	5.56	4.27	5.71	8.45	5.56	6.24	1.24	2.81	13,901,635,566
All	8.72	1.46	5.49	6.36	3.78	7.04	2.32	5.19	5.05	9.67	2.29	3.81	5.24	3.94	5.90	7.33	5.57	6.74	1.27	2.81	23,931,698,046

Similar statistics for the 20,115 individual proteomes included in Proteome-*pI* 2.0 are available online on separate subpages. Additionally, the online version of the table http://isoelectricpointdb2.org/statistics.html is accompanied by an error estimated with 1,000 bootstraps. For di-amino acid frequencies, see Supplementary Table S3 in the Proteome-*pI* 2.0 publication (Kozlowski 2021b).

Table 12. User statistics.

	Citations Google Scholar	Article access	Release date	Number of users	Link to statistics
IPC	245	>26,000	December 2015	>225,000	https://bit.ly/3BnrJBl
Proteome-pI	175	>6,000	September 2016	>19,000	https://bit.ly/3iHhsZd
IPC 2.0	4	>2,000	December 2020	>6,500	https://bit.ly/2Yz2sWr
Proteome-pI 2.0	0	>300	September 2021	>300	https://bit.ly/3aeAVMv

References

1. Kozlowski,L.P. (2016) IPC–isoelectric point calculator. *Biology direct*, **11**, 55.

2. Kozlowski,L.P. (2017) Proteome-pI: proteome isoelectric point database. *Nucleic Acids Res.*, **45**, D1112–D1116.

3. Kozlowski,L.P. (2021) IPC 2.0: prediction of isoelectric point and pKa dissociation constants. *Nucleic Acids Res*, **49**, W285–W292.

4. Kozlowski,L.P. (2021) Proteome-pI 2.0: Proteome Isoelectric Point Database Update. *Nucleic Acids Research*, http://dx.doi.org/10.1093/nar/gkab944.

5. Pace,C.N., Grimsley,G.R. and Scholtz,J.M. (2009) Protein ionizable groups: pK values and their contribution to protein stability and solubility. *Journal of Biological Chemistry*, **284**, 13285–13289.

6. Po,H.N. and Senozan,N.M. (2001) The Henderson-Hasselbalch equation: its history and limitations. *Journal of Chemical Education*, **78**, 1499.

7. Klose,J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. *Humangenetik*, **26**, 231–243.

8. O'Farrell,P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *Journal of biological chemistry*, **250**, 4007–4021.

9. Zhu,M., Rodriguez,R. and Wehr,T. (1991) Optimizing separation parameters in capillary isoelectric focusing. *Journal of Chromatography A*, **559**, 479–488.

10. Kirkwood,J., Hargreaves,D., O'Keefe,S. and Wilson,J. (2015) Using isoelectric point to determine the pH for initial protein crystallization trials. *Bioinformatics*, **31**, 1444–1451.

11. Cologna,S.M., Russell,W.K., Lim,P.J., Vigh,G. and Russell,D.H. (2010) Combining isoelectric point-based fractionation, liquid chromatography and mass spectrometry to improve peptide detection and protein identification. *Journal of the American Society for Mass Spectrometry*, **21**, 1612–1619.

12. Branca,R.M., Orre,L.M., Johansson,H.J., Granholm,V., Huss,M., Pérez-Bercoff,Å., Forshed,J., Käll,L. and Lehtiö,J. (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods*, **11**, 59.

13. Hoogland,C., Mostaguir,K., Sanchez,J.-C., Hochstrasser,D.F. and Appel,R.D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, **4**, 2352–2356.

14. Bunkute,E., Cummins,C., Crofts,F.J., Bunce,G., Nabney,I.T. and Flower,D.R. (2015) PIP-DB: the Protein Isoelectric Point database. *Bioinformatics*, **31**, 295–296.

15. Pahari,S., Sun,L. and Alexov,E. (2019) PKAD: a database of experimentally measured pKa values of ionizable groups in proteins. *Database (Oxford)*, **2019**.

16. Cargile,B.J., Sevinsky,J.R., Essader,A.S., Eu,J.P. and Stephenson,J.L. (2008) Calculation of the isoelectric point of tryptic peptides in the pH 3.5-4.5 range based on adjacent amino acid effects. *Electrophoresis*, **29**, 2768–2778.

17. Skvortsov,V.S., Alekseytchuk,N.N., Khudyakov,D.V. and Reyes,I.R. (2015) pIPredict: a computer tool for prediction of isoelectric points of peptides and proteins. *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*, **9**, 296–303.

18. Perez-Riverol,Y., Audain,E., Millan,A., Ramos,Y., Sanchez,A., Vizcaíno,J.A., Wang,R., Müller,M., Machado,Y.J. and Betancourt,L.H. (2012) Isoelectric point optimization using peptide descriptors and support vector machines. *Journal of proteomics*, **75**, 2269–2274.

19. Gauci,S., Van Breukelen,B., Lemeer,S.M., Krijgsveld,J. and Heck,A.J. (2008) A versatile peptide pI calculator for phosphorylated and N-terminal acetylated peptides experimentally tested using peptide isoelectric focusing. *Proteomics*, **8**, 4898–4906.

20. Heller,M., Ye,M., Michel,P.E., Morier,P., Stalder,D., Jünger,M.A., Aebersold,R., Reymond,F. and Rossier,J.S. (2005) Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.*, **4**, 2273–2282.

21. Wales,D.J. and Doye,J.P.K. (1997) Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A*, **101**, 5111–5116.

22. Byrd,R.H., Lu,P., Nocedal,J. and Zhu,C. (1995) A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.

23. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, **17**, 261–272.

24. Storn,R. and Price,K. (1997) Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, **11**, 341–359.

25. Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1-27:27.

26. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202-205.

27. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

28. Song,Y., Mao,J. and Gunner,M.R. (2009) MCCE2: improving protein pKa calculations with extensive side chain rotamer sampling. *Journal of computational chemistry*, **30**, 2231–2247.

29. Anandakrishnan,R., Aguilar,B. and Onufriev,A.V. (2012) H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research*, **40**, W537–W541.

30. Rostkowski,M., Olsson,M.H.M., Søndergaard,C.R. and Jensen,J.H. (2011) Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. *BMC Struct. Biol.*, **11**, 6.

31. Kilambi,K.P. and Gray,J.J. (2012) Rapid calculation of protein pKa values using Rosetta. *Biophys. J.*, **103**, 587–595.

32. Kiraga,J., Mackiewicz,P., Mackiewicz,D., Kowalczuk,M., Biecek,P., Polak,N., Smolarczyk,K., Dudek,M.R. and Cebrat,S. (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics*, **8**, 163.

33. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, **49**, D480–D489.

34. Reis,P.B.P.S., Clevert,D.-A. and Machuqueiro,M. (2021) pKPDB: a protein data bank extension database of pKa and pI theoretical values. *Bioinformatics*, 10.1093/bioinformatics/btab518.

35. SIB Swiss Institute of Bioinformatics Members (2016) The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Res*, **44**, D27-37.

36. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, **28**, 45–48.

37. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M., *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res*, **49**, D437–D451.

38. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **33**, D501–D504.

39. Maillet,N. (2020) Rapid Peptides Generator: fast and efficient in silico protein digestion. *NAR Genomics and Bioinformatics*, **2**.

40. Lear,S. and Cobb,S.L. (2016) Pep-Calc.com: a set of web utilities for the calculation of peptide and peptoid properties and automatic mass spectral peak assignment. *J Comput Aided Mol Des*, **30**, 271–277.

41. Mohanta,T.K., Khan,A., Hashem,A., Abd_Allah,E.F. and Al-Harrasi,A. (2019) The molecular mass and isoelectric point of plant proteomes. *BMC Genomics*, **20**, 631.

42. Mohanta,T.K., Mishra,A.K., Khan,A., Hashem,A., Abd-Allah,E.F. and Al-Harrasi,A. (2021) Virtual 2-D map of the fungal proteome. *Sci Rep*, **11**, 6676.

43. Chasapis,C.T. and Konstantinoudis,G. (2020) Protein isoelectric point distribution in the interactomes across the domains of life. *Biophysical Chemistry*, **256**, 106269.

44. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*, **40**, D700-705.

5 *Presentation of significant scientific activity carried out at more than one university, scientific institution, especially at foreign institutions*

The scientific activity of the Applicant began during studies at Akademia Swietokrzyska (currently Jan Kochanowski University, Kielce, Poland). His interests were already at the overlap of biology and computer science. First, he simultaneously attended biology (MSc in genetics, 2006) and computer science studies (BSc, 2007). Subsequently, there was a one-year internship (supervisor: Prof. Dr. Hab. Jan Pałyga). In this period, his work focused on linker histone analyses (both theoretical and experimental: phylogenetics, PCR, SDS-PAGE, etc.). Additionally, he taught the classes from genetics and presented some popular science lectures (Science Festival).

In 2007, the Applicant moved to Warsaw to start Ph.D. studies in the Laboratory of Bioinformatics and Protein Engineering headed by Prof. Dr. Hab. Janusz M. Bujnicki at the International Institute of Molecular and Cell Biology. His work has since focused on computational biology. This includes both using the methods already available (e.g. homology modeling, next-generation sequencing [NGS] data analysis) and developing new bioinformatics methods and databases (e.g. a MetaDisorder program for predicting intrinsically disordered proteins, which was the best program in this category during the biannual, blind experiment Critical Assessment of Protein Structure Prediction in 2008 and 2010). When the Applicant worked in the Laboratory of Bioinformatics and Protein Engineering (2007–2015), he was also the lead developer of the GeneSilico Metaserver (https://genesilico.pl/meta2). This webserver was initially published in 2003 and allowed ~10–20 methods to be run. Up to 2015, over 120 bioinformatics methods were integrated. Eventually, the webserver was closed in 2017. Moreover, in this period, the Applicant developed the following resources:

- mRNA3db http://mrna3db.netmark.pl,
- CompaRNA http://genesilico.pl/comparna,
- GDFuzz3D http://iimcb.genesilico.pl/gdserver/GDFuzz3D/.

The Ph.D. defense took place in 2013, and the Applicant stayed in the Laboratory of Bioinformatics and Protein Engineering for the next two years to finish ongoing projects and grants (for instance, NGS sequencing analysis for the Exgenome and Iuventus Plus grants). This resulted in additional publications (e.g. Sierocka et al. 2014, Plotka et al. 2014, Głów et al. 2015, Pietal et al. 2015, Plotka et al. 2015).

In June 2015, the Applicant moved to Dr. Johannes Soding's group at the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany, for postdoctoral research (over two years). During this period, he was responsible for developing a deep learning program for predicting the strength of transcription activation for proteins containing transcription activation domains (tADs). The experimental data, related to over one million random mutants sorted with FACS according to the transcription activation, was obtained by Dr. Ariel Erijman from the Fred Hutchinson Cancer Research Center (USA). The Applicant was responsible for the bioinformatics part of the study. The work was published in 2020 in *Molecular Cell* (Erijman, Kozlowski, et al. 2020).

After returning to Poland, the Applicant was employed at the Institute of Informatics, Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, where he still works today. His current research is focused on proteomics (for instance, the presented cycle of publications) and genomics. Moreover, he continues to work in the structural biology field (bioinformatic analyses of thermophilic proteins, together with Prof. Dr. Hab. Tadeusz Kaczorowski's group from the University of Gdansk). For his ongoing studies, he actively establishes new scientific collaborations (for instance, the Isoelectric Point Calculator 2.0 project required experimental data that were kindly provided by Prof. Janne Lehtiö from the Karolinska Institute in Sweden).

6 *Presentation of teaching and organizational achievements as well as achievements in popularization of science*

The Applicant worked for most of his scientific career in research institutes (IIMCB, Max Planck); therefore, until 2018, his teaching experience was limited: one semester of classes (biochemistry and genetics) at Jan Kochanowski University in 2006/2007; and two short, dedicated tutorials on protein modeling at the University of Warsaw (2012) and Göttingen University (2016). Furthermore, he supervised one Master's thesis at Warsaw University of Technology (2014).
Since October 2018, he has been employed as assistant professor (*pol*. adiunkt) in the Institute of Informatics, Faculty of Mathematics, Informatics, and Mechanics, at the University of Warsaw, where the Applicant has taught the following subjects:

2018/19

- Web applications (classes/laboratory, **30h)**
- Introduction to computer science (classes, **30h)**

- Probability theory and statistics (classes, **30h)**
- Probability theory and statistics (laboratory, **15h)**
- *Statistical data analysis 2 (laboratory, **2x30h)****

2019/20

- *Data analysis and visualization (lecture, **30h**)**
- *Data analysis and visualization (laboratory, **2x30h**)**

2020/21

- *Data analysis and visualization (lecture, **30h**)**
- *Data analysis and visualization (laboratory, **3x30h**)**
- *Architecture of large projects in bioinformatics (lecture, **30h**)**
- *Architecture of large projects in bioinformatics (laboratory, **30h**)**

2021/22

- Introduction to computer science (classes/c*aboratory*, **30h)**

Classes: whiteboard exercises (old-school method); laboratory (in front of the computer).
** In italics the classes were taught in English*

Additionally, the Applicant has reviewed a few Master's and Bachelor's theses. Currently, he is supervising another Bachelor's thesis.

10.12.2021 

……..……..…….
(Applicant's signature)