Supporting Information:

Quantitative prediction of charge regulation in oligopeptides

Raju Lunkad, Anastasiia Murmiliuk, Pascal Hebbeker, Milan Boublík, Zdeněk Tošner, Miroslav Štěpánek, and Peter Košovan[*]

Department of Physisal and Macromolecular Chemistry, Charles University, Hlavova 8, 128 43 Prague, Czech Republic

E-mail: peter.kosovan@natur.cuni.cz

November 21, 2020

Contents

1 Simulations

1.1 Coarse Grained (CG) Simulations

1.1.1 Simulation Protocol and data analysis

All Simulations were performed using time step $\delta t = 0.01\tau$, where $\tau = \sigma\sqrt{m/\epsilon}$. The particle mass m is arbitrary and has no effect on the results. The system was kept at a constant temperature $T = 300\,\mathrm{K}$ via a Langevin thermostat with a damping constant $\gamma = 1.0\tau^{-1}$. The duration of each simulation was 10^5 cycles, and each cycle consisted of 10 reaction moves followed by 100 integration steps of the Langevin dynamics. First 20% of all cycles were discarded as the equilibration. The remaining part was treated as production run and used for analysis. The productive run typically produced approximately 10^3 uncorrelated samples of peptide conformations measured by the autocorrelation time of the radius of gyration, and twice the number of uncorrelated samples of the degree of ionisation. We used the correlation-corrected error estimates to assess the statistical accuracy of our data.[S1]

1.1.2 Peptide conformations in CG Simulations

In this section we discuss only the global conformational characteristics of the peptides on the level of the whole chain. Additional discussion of local conformational characteristics on the level of individual amino acid side-chains is provided in Section 1.3.

Fig.S1 shows the average distances between the first and last central bead in the sequence (end-to-end distance) as a function of pH for each of the two studied peptides. The maximum extension of the peptides is attained at extreme pH values, when one of the blocks is almost fully ionized should be fully stretched, while the other one is neutral and should be coiled. Accordingly, the determined end-to-end distances correspond to about one half of the contour length of the backbone. Around isoelectric point the end-to-end distances attain a minimum value, demonstrating a significant coupling between the conformation and ionization of the peptide. The measured values of end-to-end distances at isoelectric point correspond to
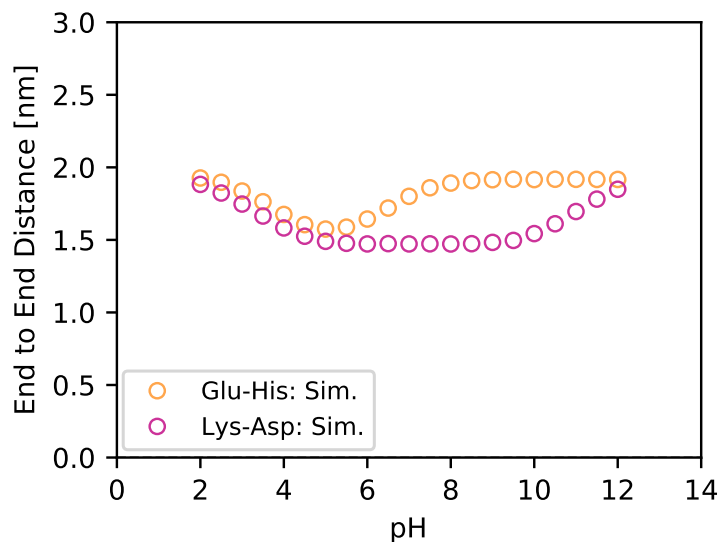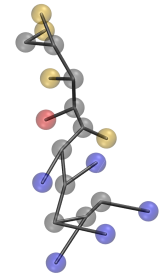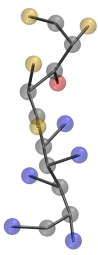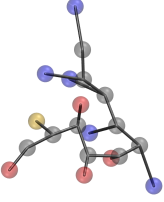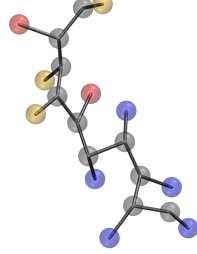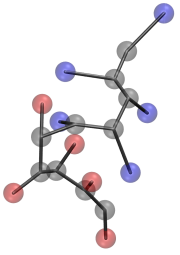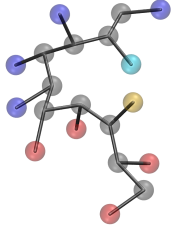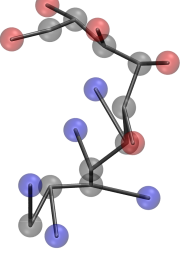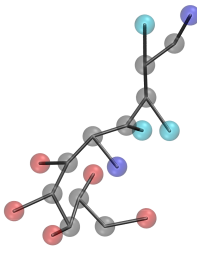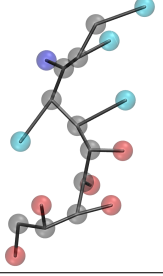
Figure S1: Average end-to-end distances of the studied peptides as a function of pH.

about 80% of the maximum, or 40% of the contour length of the peptide backbone. Thus, it is unlikely that these peptides form hairpin-like conformations at the isoelectric point, even though they consist of two oppositely charged blocks. This claim is further supported by representative simulation snapshots at the extreme pH values and near the isoelectric point, shown in Fig. S1. The snapshots confirm that even though the conformation near isoelectric point is less extended than at the extreme pH values, it is certainly not hairpin-like. This is presumably because our peptides are rather short, such that the electrostatic energy gain upon compaction is not sufficient to overcome the entropy loss. We anticipate that this situation would be different if we had used longer blocks, resulting in the formation of hairpins or aggregation. The role of sequence and chain length in the tendency for hairpin formation might be an interesting topic for further investigations. In the current context, what we observe is the desired behaviour that keeps the peptides freely dissolved as single chains in solution.

Table S1: Simulation snapshots of the studied peptides at selected values of pH. Colour code: grey = backbone, red = ionized acid group, yellow = non-ionized acid group, blue = ionized base group, cyan = non-ionized base group.

pH	$\mathrm{Lys_5 - Asp_5}$	$\mathrm{pH - pI}$	$\mathrm{Glu_5 - His_5}$	pH
2		$\ll 0$		2
4		$\lesssim 0$		3
7		≈ 0		5
9		$\gtrsim 0$		7
12		$\gg 0$		12

1.2 All Atom (AA) Simulations

1.2.1 Simulation Model and Setup

We simulated the tetramers solvated with 4028 water molecules, neutralised by adding Cl^- and Na^+ ions, with additional Na^+ and Cl^- ions to represent the added salt, which determined the ionic strength. In total, $11Cl^-$ and $7Na^+$ ions were present for the Glu_4 and Asp_4, while $7Cl^-$ and $11Na^+$ ions were present for the His_4 and Lys_4. The system was simulated in a cubic box with an edge length of $L = 6.00\,nm$, yielding the salt concentration of $0.05\,M$. Gromacs 2018.6 package was used for AA MD simulations.[S2,S3]

1.2.2 Interaction potentials

We used AMBER99sb-ILDN force filed for the peptide and TIP3P force field for the water molecules. The LINCS algorithm was used to impose the constraints on the bond lengths. The Particle Mesh Ewald (PME) method was used for long-range electrostatic interactions. The Van der Waals interactions were truncated at $1.2\,nm$.

1.2.3 Simulation protocol and data analysis

We performed 5×10^4 energy minimisation steps using the steepest descent method to remove high-energy contacts. After energy minimisation, we performed an $[NVT]$ run of $500\,ps$ using the velocity re-scaling algorithm at a temperature of $T = 300\,K$ with the thermostat coupling constant $\tau = 0.1\,ps$. Last, we performed an $[NPT]$ run of $100\,ns$ using Parrinello-Rahman algorithm at pressure $1\,bar$ with a pressure coupling constant $\tau = 2.0\,ps$. The integration time step used was $2\,fs$ for all simulations. The last $90\,ns$ of the $[NPT]$ run were used for production and data analysis.

(a) Glu$_4$

(b) His$_4$
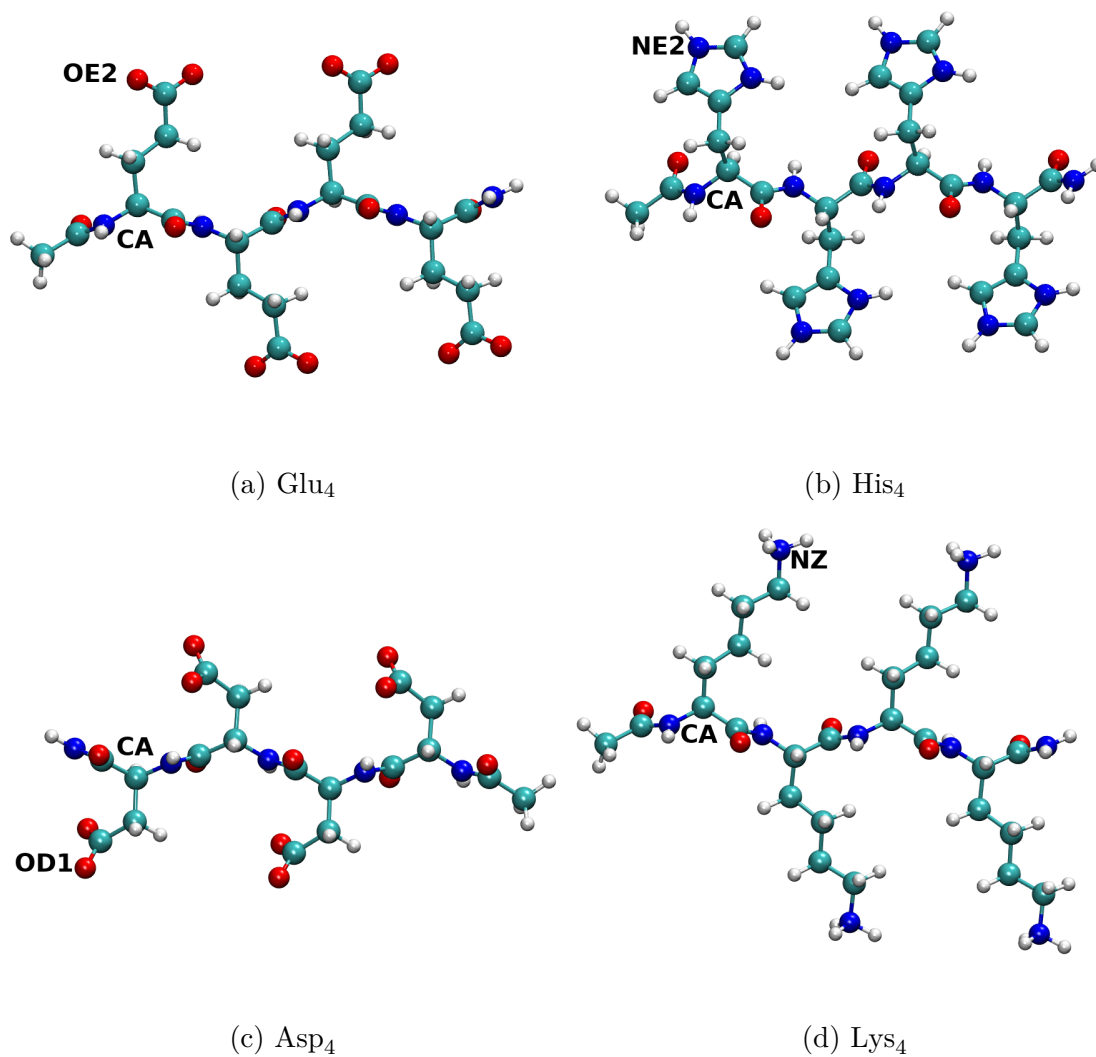
(c) Asp$_4$

(d) Lys$_4$

Figure S2: Initial configurations of the peptides used in AA simulations. Labels mark the atoms which we used to measure various intra-molecular distances.

1.3 Determination and validation of parameters for the CG model

To determine the bond lengths for the CG model of the peptides, we calculated the distributions of distances between the atoms of each tetramer from AA simulations. Furthermore, we calculated additional distributions from AA simulations, which we compared to the corresponding distributions from CG simulations to verify the validity of our model. Lastly, we determined the same set of distances from the peptide structures after simple energy minimization using the Avogadro software.[S4]

Specifically, we calculated the distance distributions between CA atoms in the peptide backbones from the AA simulations, which we then used to set the equilibrium bond length, r_{CC} between the C beads in the CG simulations. Subsequently, we calculated the distance distributions between the CA atom in the peptide backbone and the charged atom in the side chain. The charged atom was OE2 for Glu_4, NE2 for His_4, OD1 for Asp_4 and NZ for Lys_4, as indicated in Fig.S2. We used these distances to set the CG equilibrium bond lengths between the C and A beads of the acidic side chains r_{AC}, and between the C and B beads of the basic side-chains, r_{BC}. Furthermore, to verify the charge-charge distance predicted from the CG simulations, we also calculated the distribution between the charged groups on the nearest-neighbour and next-nearest-neighbour amino acid side chains, and compared them between the AA and CG simulations. To verify whether any of the above distances depend on pH, we used CG simulations at two extreme pH values: pH = 1 and pH = 13. At pH = 1 the basic groups are fully charged, while the acidic groups are uncharged. These results should match the AA simulations of the fully charged basic peptides. At pH = 13 this situation is reversed, and these results should match the AA simulations of the fully charged acidic peptides.

1.3.1 Distances between the central beads.

Table S2 shows that the distances between CA atoms on the peptide backbone from AA simulations were very well reproduced by the distances between the C beads in the CG
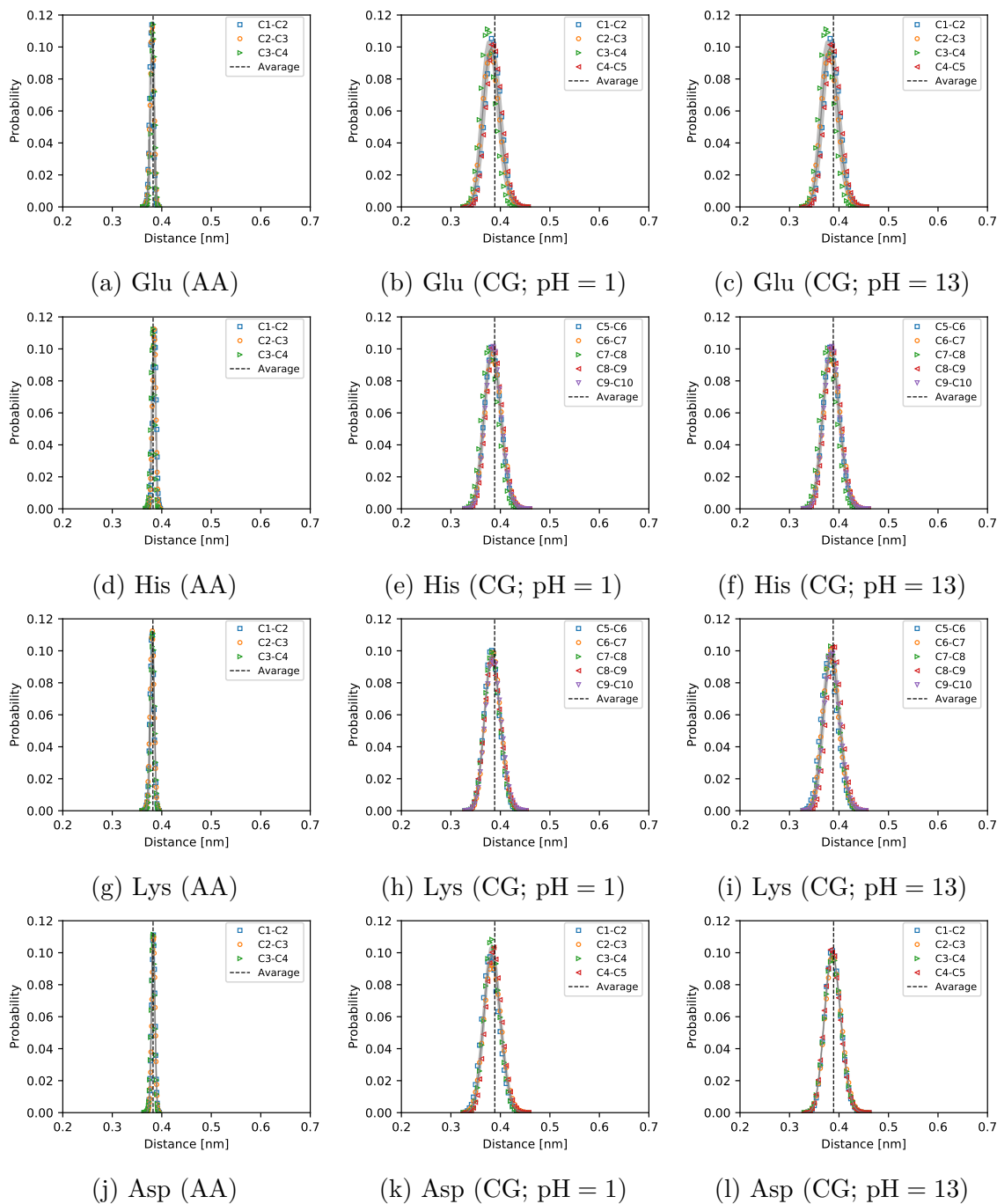
Figure S3: Distribution of distances between CA atoms on neighbouring amino acids from AA and CG simulations (the latter at the extreme pH values). The vertical line shows the average distance after averaging over individual pairs. These average values are listed in Table S2.

Table S2: Average distance between the CA atoms on neighbouring amino acids in the AA simulations; Distance between the same atoms determined using energy minimisation in the Avogadro software; Average distance between the central beads in the CG simulations, r_{CC}, at the extreme pH values.

Peptide	Acid Amino acid	distance [nm]	Base Amino acid	distance [nm]
$\mathrm{Glu}_5 - \mathrm{His}_5$	Glu (AA)	0.382 ± 0.004	His (AA)	0.382 ± 0.004
	Glu (CG; pH = 1)	0.389 ± 0.016	His (CG; pH = 1)	0.388 ± 0.016
	Glu (CG; pH = 13)	0.389 ± 0.016	His (CG; pH = 13)	0.388 ± 0.016
	Glu (Avogadro)	0.388	His (Avogadro)	0.389
$\mathrm{Lys}_5 - \mathrm{Asp}_5$	Asp (AA)	0.382 ± 0.006	Lys (AA)	0.382 ± 0.006
	Asp (CG; pH = 1)	0.389 ± 0.016	Lys (CG; pH = 1)	0.388 ± 0.016
	Asp (CG; pH = 13)	0.389 ± 0.016	Lys (CG; pH = 13)	0.388 ± 0.016
	Asp (Avogadro)	0.391	Lys (Avogadro)	0.389

models. The differences between AA and CG models within approx. 1% are well below the statistical uncertainty. The distances measured using the Avogadro software agree very well with all simulations. The distributions of distances in Fig. S3 reveal that all distributions consist of a single peak. The AA distributions are very narrow, while the CG distributions are slightly broader. However, the average distances from these distributions show no visible dependence on the type of the amino acid, or on the pH.

1.3.2 Distances between central beads C and side-chain beads A or B.

Distances between the CA atoms on peptide backbone and the charged atoms on the respective side chains show a similar trend to the distances between CA atoms. Table S3 reveals that the differences between the AA and CG simulations are slightly larger in Asp and Lys, but they remain within the estimated statistical error; thus, we consider them insignificant. Also the distances measured using the Avogadro software agree with all simulations, although the differences are slightly beyond the estimated statistical error of the simulation data. Fig. S4 reveals that these differences could be attributed to a more complex
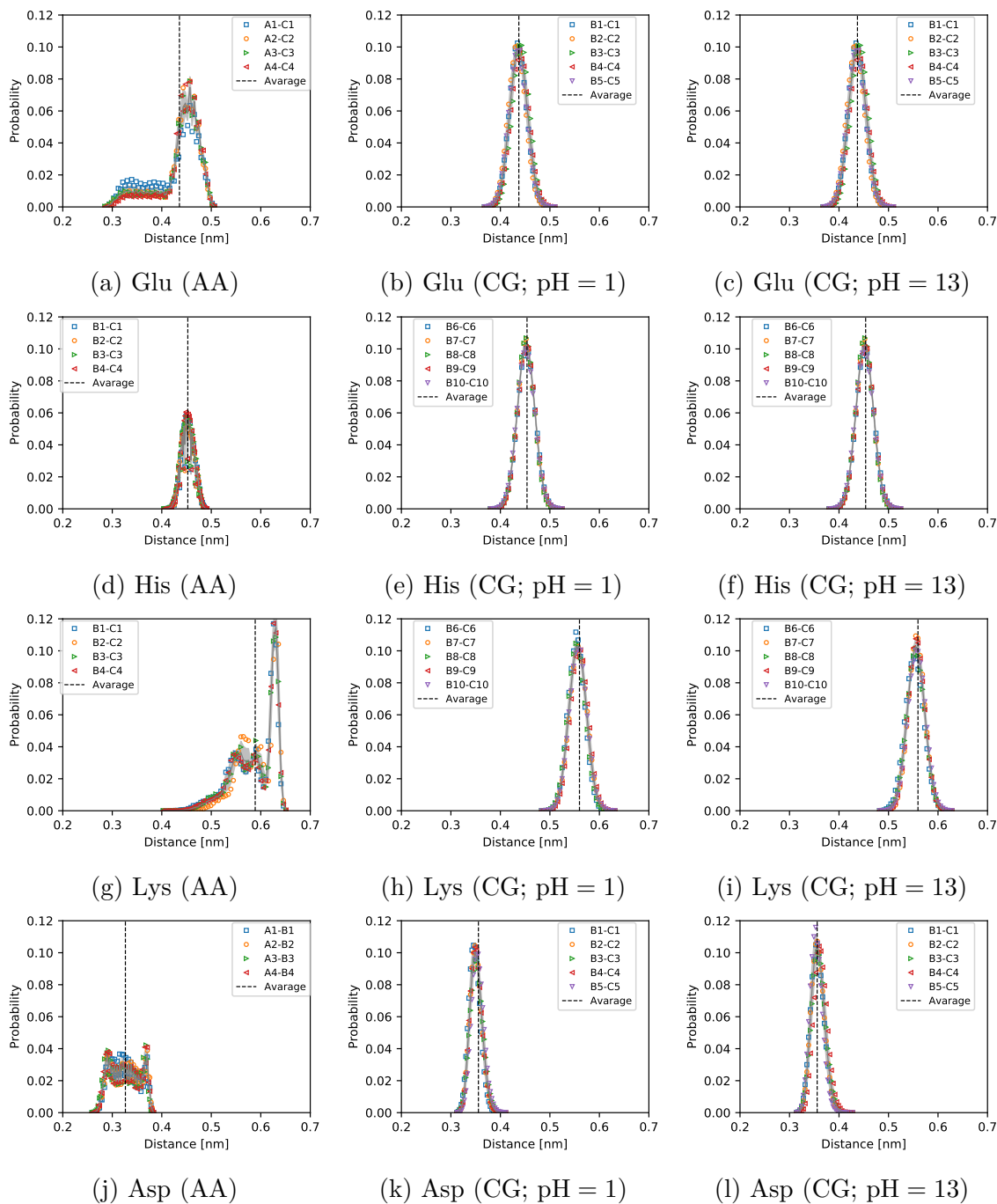
Figure S4: Distribution of distances between the CA atoms and the charged group on the amino acids from the AA and CG simulations (the latter at the extreme pH values). The vertical line shows the average distance after averaging over individual pairs. These average values are listed in Table S3.

Table S3: Average distance between the CA atoms and the charged group on the amino acids in the AA simulations; Distance between the same atoms determined using energy minimisation in the Avogadro software; Average distance between the central bead and the A or B bead in the CG simulations, r_{AC} and r_{BC}, at the extreme pH values;

Peptide	Acid Amino acid	distance [nm]	Base Amino acid	distance [nm]
$Glu_5 - His_5$	Glu (AA)	0.436 ± 0.044	His (AA)	0.453 ± 0.013
	Glu (CG; pH = 1)	0.437 ± 0.018	His (CG; pH = 1)	0.454 ± 0.018
	Glu (CG; pH = 13)	0.437 ± 0.018	His (CG; pH = 13)	0.454 ± 0.018
	Glu (Avogadro)	0.453	His (Avogadro)	0.462
$Lys_5 - Asp_5$	Asp (AA)	0.327 ± 0.029	Lys (AA)	0.589 ± 0.042
	Asp (CG; pH = 1)	0.356 ± 0.012	Lys (CG; pH = 1)	0.560 ± 0.018
	Asp (CG; pH = 13)	0.356 ± 0.012	Lys (CG; pH = 13)	0.560 ± 0.018
	Asp (Avogadro)	0.385	Lys (Avogadro)	0.639

shape of the AA distributions, which was not fully reproduced by the CG simulations. This could be attributed to cis-trans conformational transitions, hydrogen bonds, or other specific interactions, which were not explicitly included in the CG model. Nevertheless, the average values of the distances are reproduced within the statistical uncertainty of approx. 5%.

1.3.3 Distances between the nearest-neighbour side-chain beads.

All previously discussed distances were used as inputs in constructing the CG model; therefore, as expected, the CG model reproduces well their values calculated from the AA simulations. To verify the validity of the CG model, we compared the distributions of distances between neighbouring charged groups on the peptides. The ability of the CG model to reproduce these distances is crucial to quantitatively account for the effect of electrostatic interactions within the peptide. Table S4 reveals that the average distances between the nearest-neighbour charged groups are well reproduced, but the statistical uncertainty has increased to approximately 10%. Interestingly, the distances determined from Avogadro rea-

(a) Glu (AA) (b) Glu (CG; pH = 1) (c) Glu (CG; pH = 13)

(d) His (AA) (e) His (CG; pH = 1) (f) His (CG; pH = 13)

(g) Lys (AA) (h) Lys (CG; pH = 1) (i) Lys (CG; pH = 13)

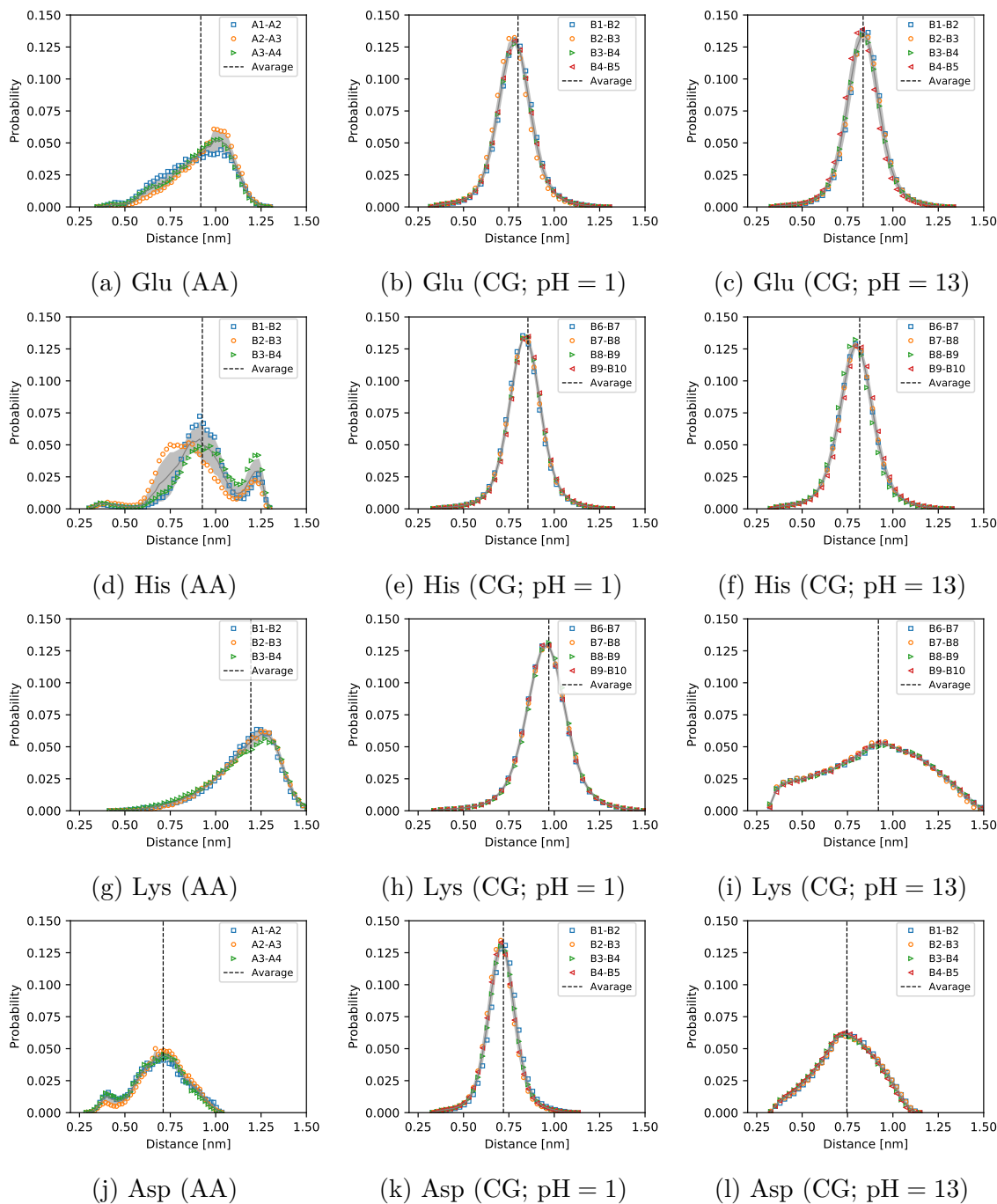(j) Asp (AA) (k) Asp (CG; pH = 1) (l) Asp (CG; pH = 13)

Figure S5: Distribution of distances between charged group on neighbouring amino acids from the AA and CG simulations (the latter at the extreme pH values). The vertical line shows the average distance after averaging over individual pairs. These average values are listed in Table S4.

Table S4: Average distance between the charged group on the nearest-neighbour amino acids in the AA simulations; Distance between the same atoms determined using energy minimisation in the Avogadro software; Average distance between the A and the nearest-neighbour A bead, or B and the nearest-neighbour B bead, in the CG simulations at the extreme pH values.

Peptide	Acid Amino acid	distance [nm]	Base Amino acid	distance [nm]
$Glu_5 - His_5$	Glu (AA)	0.92 ± 0.16	His (AA)	0.93 ± 0.18
	Glu (CG; pH = 1)	0.80 ± 0.11	His (CG; pH = 1)	0.86 ± 0.11
	Glu (CG; pH = 13)	0.84 ± 0.11	His (CG; pH = 13)	0.82 ± 0.12
	Glu (Avogadro)	1.04	His (Avogadro)	0.91
$Lys_5 - Asp_5$	Asp (AA)	0.71 ± 0.14	Lys (AA)	1.19 ± 0.18
	Asp (CG; pH = 1)	0.72 ± 0.09	Lys (CG; pH = 1)	0.97 ± 0.14
	Asp (CG; pH = 13)	0.75 ± 0.16	Lys (CG; pH = 13)	0.92 ± 0.27
	Asp (Avogadro)	0.93	Lys (Avogadro)	1.14

sonably agree with the simulations. The average distances are weakly affected by the pH, albeit still below the estimated statistical error. In contrast, the shape of the distributions in Fig. S5 clearly depends on pH and on the type of the amino acid. For Glu and Lys, the AA simulations yield a skewed distribution with a single peak; inc contrast, for His and Asp, they yield a double-peak distribution. The CG simulations yield a single-peaked distribution in all cases, but the width of this peak depends on the pH and on the type of the amino acid. This difference between the CG and AA distributions can be again attributed to specific details of the atomistic structure, which were not fully included in the CG simulations.

1.3.4 Distances between the next-nearest-neighbour side-chain beads.

Finally, we compared the distances between charged groups on next-nearest-neighbour amino acids, measured in the AA and CG simulations. Table S5 reveals an even greater statistical uncertainty of the average value, approx. 10–20%. Within the given uncertainty, the CG and AA simulations still agree with each other. However, the distances depend on pH: the
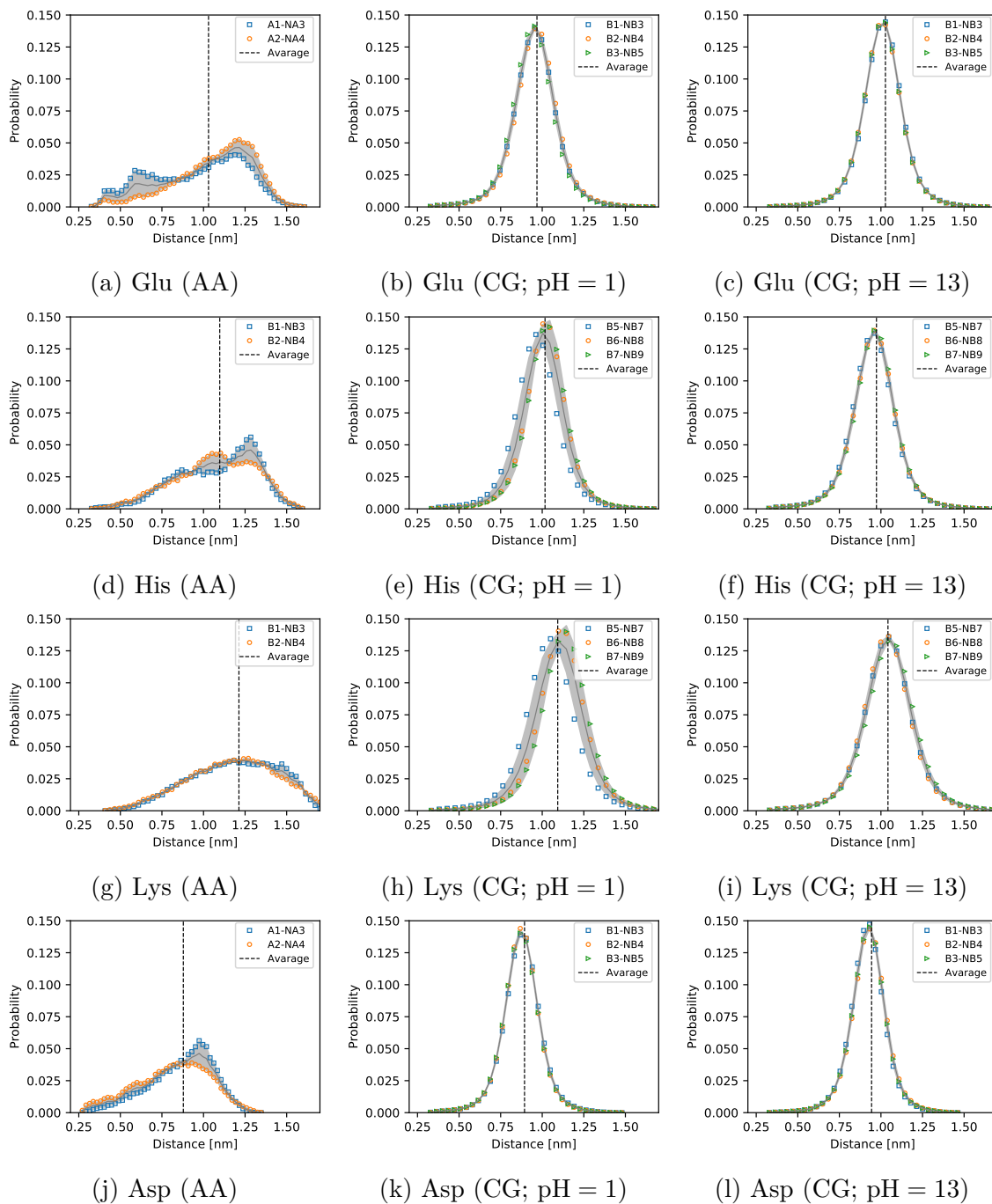
Figure S6: Distribution of distances between charged group on next-nearest-neighbour amino acids from the AA and CG simulations (the latter at the extreme pH values). The vertical line shows the average distance after averaging over individual pairs. These average values are listed in Table S5.

Table S5: Average distance between the charged group on next-nearest-neighbour amino acids in the AA simulations; Distance between the same atoms determined using energy minimisation in the Avogadro software; Average distance between the A and next-nearest-neighbour A beads, or B and next-nearest-neighbour B bead, in the CG simulations at the extreme pH values.

Peptide	Acid Amino acid	distance [nm]	Base Amino acid	distance [nm]
$Glu_5 - His_5$	Glu (AA)	1.03 ± 0.26	His (AA)	1.10 ± 0.24
	Glu (CG; pH = 1)	0.97 ± 0.14	His (CG; pH = 1)	1.02 ± 0.14
	Glu (CG; pH = 13)	1.03 ± 0.13	His (CG; pH = 13)	$0.97 \pm 0..14$
	Glu (Avogadro)	0.74	His (Avogadro)	0.73
$Lys_5 - Asp_5$	Asp (AA)	0.88 ± 0.21	Lys (AA)	1.21 ± 0.27
	Asp (CG; pH = 1)	0.89 ± 0.12	Lys (CG; pH = 1)	1.09 ± 0.16
	Asp (CG; pH = 13)	0.94 ± 0.12	Lys (CG; pH = 13)	1.04 ± 0.16
	Asp (Avogadro)	0.74	Lys (Avogadro)	0.73

distances measured in the uncharged state are systematically lower than those measure in the charged state, although this difference is on the verge of the estimated statistical error. The distances between from AA simulations are better matched by CG results at a pH which corresponds to the charged state of the respective group.

This difference can be explained by the electrostatic repulsion between charged groups, which is absent in the neutral state. Furthermore, we observe that the structures obtained by energy minimisation using the Avogadro software yield significantly lower average distances than any of the simulation models. The distributions from AA simulations in Fig. S6 are all single-peaked and rather broad, while the distributions from CG simulations are narrower and more symmetric. Moreover, the next-nearest-neighbour distances are only slightly greater than the nearest-neighbour distances, demonstrating that the side chains prefer the trans conformation. In turn, the role of chain flexibility is demonstrated by the next-nearest-neighbour distances, which are significantly greater than the distances in all-trans conformations determined from Avogadro. Earlier on, we tested a simpler model in

which each amino acid was represented by just one bead. This one-bead model was able to reproduce the nearest-neighbour distances, while yielding next-nearest-neighbour distances approximately twice as large as the nearest-neighbour distances. The two-bead model can approximate well these two distances. They are crucial to account for the electrostatic effect on the ionization of peptides, which suggests that the two-bead model is suited for quantitative predictions of this effect.

1.4   Computational demands and costs

Various simulations described above dramatically differ in computational demands. Our CG simulations typically required approximately 5 hours of computer run time on a single CPU core for each data point of the charge(pH) curve. Thus, one such curve with 20 data points could be obtained within approximately 100 CPU core-hours. Because all these simulations were independent, they could be run simultaneously on different CPUs, thereby making it possible to obtain the full charge(pH) curve within a few hours when run on a small computer cluster.

In contrast, the AA simulations of 100 ns required approximately 5 days on 8 CPU cores, equivalent to approximately 1000 CPU core-hours. However, running one simulation per amino acid was enough to obtain the desired parameters of the CG model. When considering more complicated peptide sequences, it may be necessary to perform such an AA simulation for each pair of amino acids. Thus, running the AA simulations to parametrise the CG model is much more demanding than obtaining the whole charge(pH) curve from CG simulations, although the former can be considered a moderate computational demand. In addition, energy minimisation provides a much cheaper and faster way of estimating bond lengths for the CG model. This process is completed within several seconds on a desktop PC, providing parameter values for the amino acids studied here similar to those determined by expensive simulations. Because the energy minimisation could fail for other structures, its predictions should always be verified using an all-atom simulation.

Thus, parametrisation of the model can be completed within one week and, once the model parameters are available, the charge(pH) curves of various peptide sequences can be obtained within hours. The economic cost of the simulations could be estimated by noting the commercial prices, approximately 0.05 EUR per CPU core-hour. Hence, we estimate the cost of parameterising one amino acid for the CG model as 50 EUR, and the cost of predicting one charge(pH) curve from the CG model as 5 EUR. The above costs include direct and indirect costs of computer time, but they do not include the cost of several person-hours needed for running the simulations and analysing the results.

The cost of performing the simulations should be compared to the cost of purchasing 100 mg of one custom-synthesized peptide, approximately 400 EUR. Thus, merely purchasing the sample to start the experiments is more costly than parametrising the model and obtaining the CG simulation results. Thus, the experimental quantification of the ionisation response is much more expensive than the simulations due to additional instrument time and personnel costs required to perform the experiments.

Notably, the above cost estimation assumes that the required protocols for running the simulations, performing the experiments and analysing the results are readily available and that all steps can be performed routinely. It does not include the costs and effort needed to establish these protocols.

2 Experiments

2.1 Materials

The following peptides with acetyl and amide terminal groups and trifluoroacetate (TFA) as counterion were purchased from Biomatik LLC, Wilmington, Delaware, USA: Ac-E5-H5-NH2 ($Glu_5 - His_5$; $M = 1390.33$ g/mol; lot number P180808-DG671108 97.16% HPLC purity for CZE and lot number P190902-LL671108 97.54% HPLC purity for potentiometric titrations and NMR); Ac-K5-D5-NH2 ($Lys_5 - Asp_5$; $M = 1275.36$ g/mol; lot number P180711-

JQ665893 95.83% HPLC purity for CZE and lot number P190816-LC665893 96.99% HPLC purity for titrations and NMR). All peptides were purified, and HPLC and MS spectra were measured for all peptide sequences.

The buffers used for CZE experiments were prepared by mixing of weak acid with strong base (pH less than 7) and by mixing of strong acid with weak base (pH more than 7). Compositions of the used buffers are specified in Table S6.

Standardised solutions of HCl and NaOH from Carl Roth GmbH (Karslruhe, Germany) were used to prepare $0.1\,M$ stock solutions, which were subsequently diluted to $0.01\,M$. To prevent contamination by CO_2, the standardised solutions were kept under soda lime at least 24 hours before the measurements.

Deuterium oxide 99.8 % purity with a trace of 3-(trimethylsilyl)-1-propanesulfonic acid sodium salt (DSS) of 97 % purity from Sigma-Aldrich was used for field-frequency lock.

2.2 Capillary Zone Electrophoresis experiments (CZE)

### 2.2.1 Instrumentation and experimental protocol

All CZE experiments were performed using Agilent 7100 capillary electrophoresis equipment operated under ChemStation software (Agilent Technologies, Waldbronn, Germany). Detection was performed with the built-in diode array detector (DAD). Fused fluorocar-

Table S6: The buffers prepared for CZE measurements (ionic strength is always 10 mM).

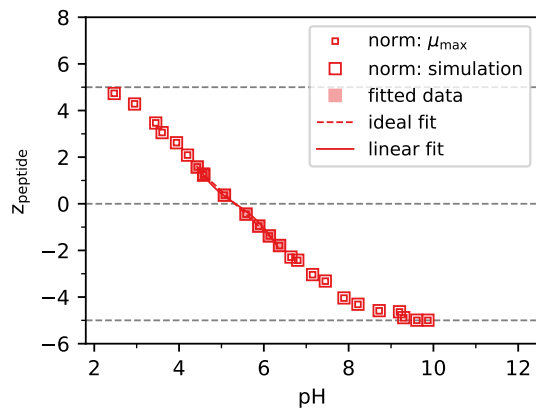| Acid	Base	pH range
Formic	Lithium hydroxide	2.5 - 3.5
Acetic	Lithium hydroxide	3.9 - 4.6
Cacodylic	Lithium hydroxide	4.6 - 7.2
Hydrochloric	Tris(hydroxymethyl)aminomethane	7.5 - 9.2
Hydrochloric	Ammonium hydroxide	9.3 - 10.7
Hydrochloric	Triethylamine	10.3 - 12.0
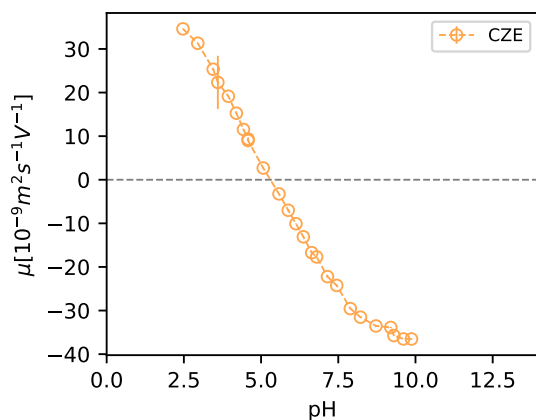
(a) $Glu_5 - His_5$
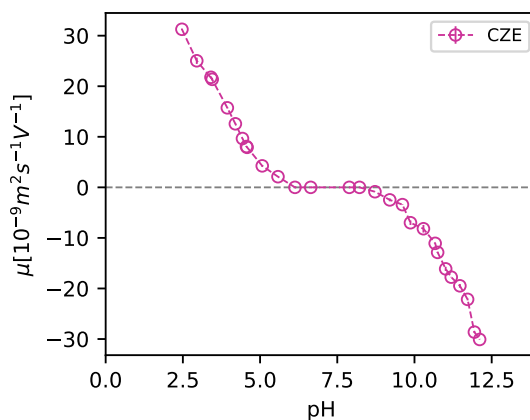
(b) $Glu_5 - His_5$

Figure S7: Electrophoretic mobilities of the peptides determined from CZE experiments.

(a) $Glu_5 - His_5$
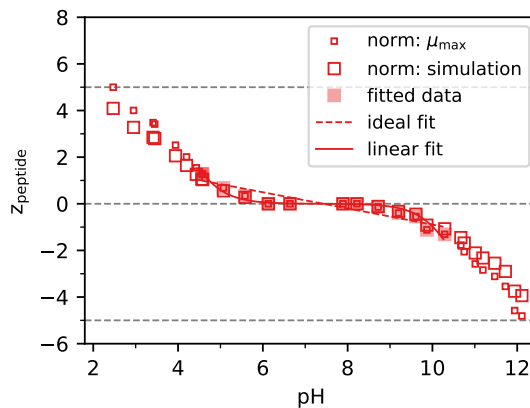
(b) $Lys_5 - Asp_5$

Figure S8: Total charge on the peptides calculated from electrophoretic mobilities determined by CZE. Small squares represent the data re-normalised by $|\mu^{max}(pH)|$. Large squares represent the data re-normalised using $\alpha(pH)$ obtained from simulations at $|\mu^{max}(pH)|$. Filled squares show the data points which were fitted by the Henderson-Hasselbalch equation (solid line), and by a line (dashed line). Isoelectric points determined by both fits coincide at least within two significant figures.

bon capillaries (50 μm i.d., 375 μm o.d.) by Agilent Technologies with a total length of 50 cm and effective length to the DAD detector of 41.5 cm were used to perform the experiments. Before the first use, each new capillary was flushed with 0.1 M sodium hydroxide from Agilent Technologies and then with deionised water for 10 min. All CZE measurements were performed in the running background electrolyte (BGE) with an ionic strength of 10 mM. The samples of peptide solutions with a concentration of monomeric units [Glu] = [His] = [Asp] = [Lys] =0.5 mM were prepared by dissolving the relevant amount of peptide directly in the running buffer. The PeakMaster 5.3 was used to calculate the properties of all buffers used in CZE experiments.[S5,S6] These buffers are listed in Table S6. We used pure 10 mM HCl and LiOH only at the lowest and highest pH (pH = 2.0 and pH = 12.1, respectively). All running buffers were filtered with 0.45 μm PVDF membranes. The samples were injected hydrodynamically using the pressure of 30 mbar for 5 s; the applied voltage was always $\pm$ 10 kV. The solutions were thermostated at 25°C. Capillary was flushed by running buffer for 3 min before each measurement, and each run was repeated three times. DAD detection was performed at a wavelength of 200 nm. The CEVal software[S7] was used to analyse the raw data and to determine the effective mobility.

2.2.2 Determination of charge on the peptide

From CZE, we obtained the absolute electrophoretic mobilities, as shown in Fig. S7. To obtain the charge on the peptide from the mobilities, we first renormalized the absolute mobilities by their maximum values for the given peptide, $\mu^{\mathrm{max}} = \max(|\mu(\mathrm{pH})|)$.

$$z_{\mathrm{cze}}(\mathrm{pH}) = \frac{\mu(\mathrm{pH} - \mathrm{pI}_{\mathrm{cze}})}{|\mu^{\mathrm{max}}|} \tag{1}$$

This renormalisation procedure is based on the assumption that diffusion coefficients of the peptides do not significantly change with pH, as confirmed independently by DOSY NMR (Fig. S22). Ideally, one should observe a plateau in the mobility at a high or low pH value,

indicating that the peptide is fully ionised. However, we were not always able to measure the maximum mobility at a pH value, which would ensure that the peptide was fully ionised. Therefore, such a normalisation could not yield a reliable value of the charge on the peptide. To correct this deficiency, we renormalised the $\mu^{\mathrm{max}}(\mathrm{pH}^{\mathrm{max}})$ by the peptide charge determined from the simulations, $z_{\mathrm{sim}}(\mathrm{pH}^{\mathrm{max}} + \Delta\mathrm{pI})$

$$z_{\mathrm{cze}}(\mathrm{pH}) = \frac{\mu(\mathrm{pH} - \mathrm{pI}_{\mathrm{cze}})}{\mu^{\mathrm{max}}(\mathrm{pH}^{\mathrm{max}})} z_{\mathrm{sim}}(\mathrm{pH}^{\mathrm{max}} + \Delta\mathrm{pI}) \tag{2}$$

where $\Delta\mathrm{pI} = \mathrm{pI}_{\mathrm{cze}} - \mathrm{pI}_{\mathrm{sim}}$. The effect of different renormalisations is shown in Fig. S8. The value of $z_{\mathrm{cze}}(\mathrm{pH})$ of $\mathrm{Glu}_5 - \mathrm{His}_5$ is barely affected by the different normalisation because the $\mu^{\mathrm{max}}(\mathrm{pH}^{\mathrm{max}} = 10)$ is well in the plateau region where the peptide is fully ionised. The value of $z_{\mathrm{cze}}(\mathrm{pH})$ of $\mathrm{Lys}_5 - \mathrm{Asp}_5$ is visibly affected by the different normalisation because the $\mu^{\mathrm{max}}(\mathrm{pH}^{\mathrm{max}} \approx 2.5)$ is still in the region where the peptide should not be fully ionized.

We note that we also attempted to compute the charge on the peptides without renormalisation, using the diffusion coefficients determined from NMR (Fig. S22). However, this attempt yielded a significantly lower peptide charge, which was not consistent with other methods (simulations, NMR, and titrations). This observation is in line with the notion that electrophoretic mobility yields the effective charge, rather than the bare charge of the analyte.[S8–S10] The difference between the effective and bare charge increases with the increase in the charge on the peptide, particularly affecting the result at high and low pH values.

### 2.2.3 Determination of the isoelectric point

To determine the isoelectric point from the CZE data, we determined the intersection of $z_{\mathrm{cze}}(\mathrm{pH})$ with $z = 0$ by fitting the data in the range $z \in \{-2, 2\}$ for he $\mathrm{Glu}_5 - \mathrm{His}_5$ peptide, and $z \in \{-1.5, 1.5\}$ for he $\mathrm{Lys}_5 - \mathrm{Asp}_5$ peptide. To ensure that our result was not affected by the arbitrary choice of the fitting range, we tested various ranges and fitted the date using the Henderson-Hasselbalch equation and a straight line. Because both fit functions were

symmetric around the pI, and the data was approximately symmetric as well, all fits yielded consistent values of pI within two significant figures. Additionally, the use of symmetric fit functions ensures that the determined isoelectric point is not affected by the renormalisation of the effective mobility. The curves obtained from the fits and the data points used in the fits are shown in Fig. S8. We avoided extending the range to higher values of z because the fit functions were not appropriate approximations, and the fit results were significantly affected by a few data points that were farther from the target value $z = 0$. The precise determination of pI of $\text{Lys}_5 - \text{Asp}_5$ was particularly tricky because its charge is almost zero in a broad pH range.

2.3 Potentiometric Titration

2.3.1 Instrumentation and experimental protocol

Potentiometric titrations were performed using a Metrohm 888 Titrando Compact titrator equipped with a Metrohm LL Biotrode 3 mm glass electrode, a Pt1000 temperature sensor, a titration vessel for 1 ml, magnetic stirrer and Titrando Software. In order to prevent the absorption of carbon dioxide from the air, standardised solutions of HCl and NaOH from Carl Roth GmbH (Karslruhe, Germany) were used to prepare 0.1 M stock solutions, which were subsequently diluted to 0.01 M. Moreover, the standardised solutions were kept under soda lime at least 24 hours before the measurements. The solutions of the peptides were prepared at concentrations of monomeric units $[\text{Glu}] = [\text{His}] = [\text{Asp}] = [\text{Lys}] = 5\,\text{mM}$ dissolved in 0.01 M standardised HCl, yielding the concentration concentration of peptide chains $c_{\text{peptide}} = 1\,\text{mM}$. Sample volumes of approximately 2 ml were weighed to determine the precise amount and then titrated by standardised 0.01 M NaOH using an automated dynamic pH titration method with signal drift 1 mV and waiting time $10 - 50\,\text{s}$. To prevent excessive contamination by CO_2, the stock solutions were kept under soda lime, and the titration vessels were sealed during the titration, for 15–90 min, depending on the sample. Blank titrations were performed under the same conditions, before and after each peptide
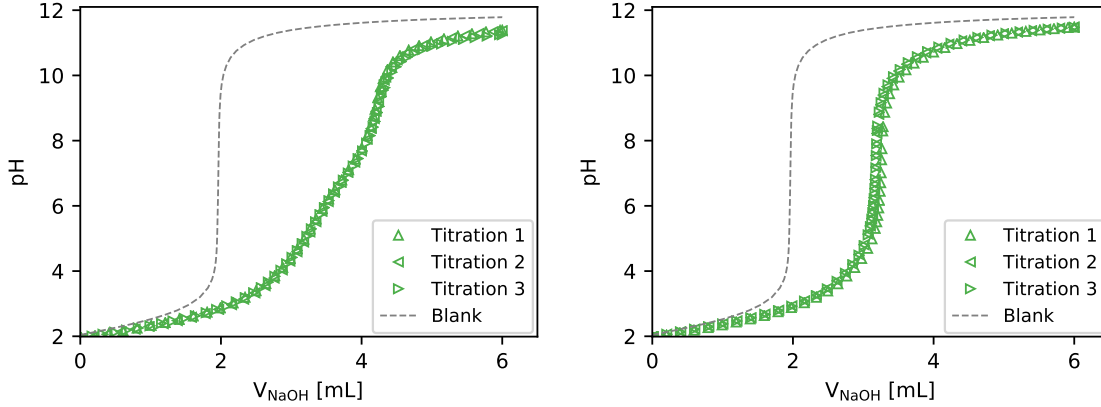
(a) $\mathrm{Glu_5 - His_5}$ (b) $\mathrm{Lys_5 - Asp_5}$

Figure S9: Potentiometric titration of the peptides.

titration, to estimate the reproducibility and reliability of the procedure and to estimate the concentration of CO_2 in the stock solution.

2.3.2 Determination of the charge on the peptide

The primary output of titration is the solution pH as a function of the volume of the added NaOH (V_{NaOH}), as shown in Fig. S9. To calculate the charge on the peptide, we used the electroneutrality condition

$$\sum_i z_i n_i = 0 \tag{3}$$

where the summation runs over all ionic species in the system.

$$z_{\mathrm{titration}} = \frac{V_{\mathrm{HCl}} c_{\mathrm{HCl}} - V_{\mathrm{NaOH}} c_{\mathrm{NaOH}} + (c_{\mathrm{OH}} - c_{\mathrm{H}})(V_{\mathrm{HCl}} + V_{\mathrm{NaOH}})}{c_{\mathrm{peptide}} V_{\mathrm{HCl}}} + z_{\mathrm{max}} x_{\mathrm{TFA}} \tag{4}$$

where c stands for concentration, V for volume, $z_{\mathrm{titration}}$ is the charge on the peptide determined from titration, $z_{\mathrm{max}} = 5$ is the maximum charge on the peptide, $x_{\mathrm{TFA}} = n_{\mathrm{TFA}}/n_{\mathrm{base}}$ is the mole fraction of the trifluoroacetate counterions contained in the peptide sample, relative to the number of basic side-chains on the peptide, V_{HCl} is the initial volume of HCl in which the peptide was dissolved. The concentrations c_{H} and c_{OH} were calculated from the

measured pH and from the pK_w, accounting for the variation of both quantities with temperature and assuming that activity coefficients are equal to one. The temperature correction improved the reproducibility of $z_{\text{titration}}(\text{pH})$ calculated from different runs, however, it could not completely remove the numerical instability at very high or low pH, caused by sensitivity of $z_{\text{titration}}(\text{pH})$ to the precision of pH measurement. When assuming that $x_{\text{TFA}} = 1$, this procedure yielded the values of $z_{\text{titration}}(\text{pH})$, which were similar to $z_{\text{cze}}(\text{pH})$, albeit shifted to lower values of z. As shown in Fig. S10, repeated runs of the same titration yielded highly reproducible $z_{\text{titration}}(\text{pH})$, except for pH $\lesssim 3$ and pH $\gtrsim 10$.

We used blank titrations, as shown in Fig. S10c, to assess the reliability and reproducibility of the calculation of $z_{\text{titration}}(\text{pH})$. These blank titrations were performed before and after each set of titrations with a peptide sample. Ideally, they should yield $z_{\text{titration}}(\text{pH}) = 0$ in the whole range. The example of a typical blank run shown in Fig. S10c highlights that the yielded $|z_{\text{titration}}(\text{pH})| \lesssim 0.1$, except for pH $\lesssim 3$ and pH $\gtrsim 11$. In the high- and low-pH range, the calculation of $z_{\text{titration}}(\text{pH})$ is very sensitive to the precision of the pH measurement and to the value of pK_w, as evidenced by the steep increase or decrease in $z_{\text{titration}}(pH)$ in Fig. S10c. The effect of CO_2 is noticeable at pH $\gtrsim 10$, and this effect was stronger in longer titrations. Therefore, this effect was weak in the blank titration, which was quick, and stronger in the titration of peptides, which were slower.

Because the blank titrations did not show the shift observed in the peptide titrations, we attributed this shift to the unknown excess of TFA anions contained in the peptide samples. Indeed, the excess TFA, commonly found in peptide samples after deprotection from the BOC groups during solid-state synthesis, results in $x_{\text{TFA}} > 1$. This assumption was supported by the results from the titration of a different batch of the same peptide, which yielded slightly shifted titration curves (not shown). Furthermore, it was supported by quantitative analysis of the NMR spectra of $Glu_5 - His_5$, shown in Fig. S11. Unfortunately, the relevant batch of $Lys_5 - Asp_5$ was no longer available in sufficient amount to analyse the TFA content by NMR. To correct for the unknown amount of TFA, we first interpolated titration data from different

(a) $\mathrm{Glu}_5 - \mathrm{His}_5$

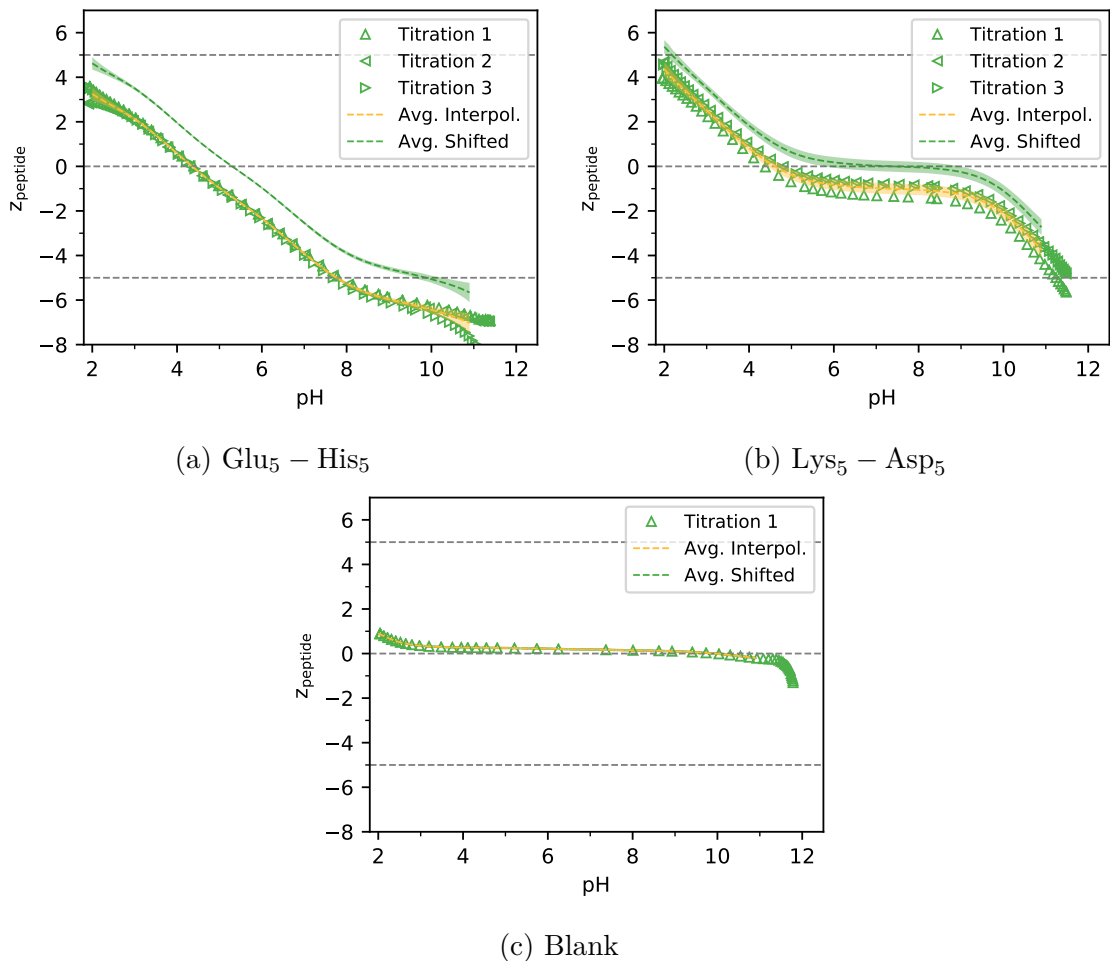(b) $\mathrm{Lys}_5 - \mathrm{Asp}_5$

(c) Blank

Figure S10: Total charge on the peptides and on the blank control, calculated from potentiometric titrations. The Yellow line shows the interpolated average over different runs. The green line show the average shifted by adjusting x_{TFA} so that the curves matched the isoelectric point determined from CZE.

runs and then used the interpolated data to compute the average $z_{\mathrm{titration}}(\mathrm{pH})$ and to estimate the accuracy using the standard deviation of different runs. Then, we used the isoelectric point determined from CZE to adjust the value of x_{TFA}, so that $z_{\mathrm{titration}}(\mathrm{pH} = \mathrm{pI}_{\mathrm{cze}}) = 0$. The results of all intermediate steps of the titration data processing are shown in Fig. S10.

2.4 NMR

2.4.1 NMR measurements and instrumentation

All NMR data were recorded using a Bruker AVANCE III spectrometer operating at the proton Larmor frequency of 600 MHz equipped with a cryogenically cooled probe and stabilising the temperature at 25 °C. The samples were prepared by dissolving the peptides in 10 mM HCl to obtain 15 gL^{-1} peptide concentration. The pH was adjusted by adding NaOH. A capillary insert containing deuterium oxide with a trace of sodium trimethylsilyl-propanesulfonate (DSS) was used for field-frequency lock and chemical shift referencing. ^{1}H spectra were acquired with water suppression using the excitation sculpting method.[S11] Measurements of translational diffusion coefficients were performed with the double stimulated echo experiment with bipolar pulse field gradients described by,[S12] combined with water suppression. The gradients were 1.5 ms long with 24 linearly spaced amplitudes spanning the range $0 - 60$ Gcm^{-1}, and the diffusion time was 300 ms. The calibration was performed using a standard sample of 1% H$_2$O in D$_2$O (doped with GdCl$_3$), for which the value of the HDO diffusion coefficient at 25 °C is 1.9×10^{-9} m^{2}s^{-1}. All data processing and fitting of the diffusion coefficients has been done using the MestReNova and GNAT software.[S13] The chemical shifts were determined by referencing to the signal of DSS for ^{1}H NMR and of TFA for ^{13}C NMR. At a very low pH the signals of TFA were affected by its ionization. Therefore, the chemical shifts of peptides at low pH were referenced to the chemical shifts of CH$_3$ terminal groups of the peptides. The MestReNova Software was used to analyse both 1D and 2D spectra, including the determination centers of mass of multiplets and the ranges of the peaks.

2.4.2 Amount of TFA estimated from the NMR spectra

In Fig. S11 we show the quantitative analysis of NMR spectrum of Glu$_5$ − His$_5$. The peaks corresponding to individual groups on each amino acid suggest that integrated intensity of
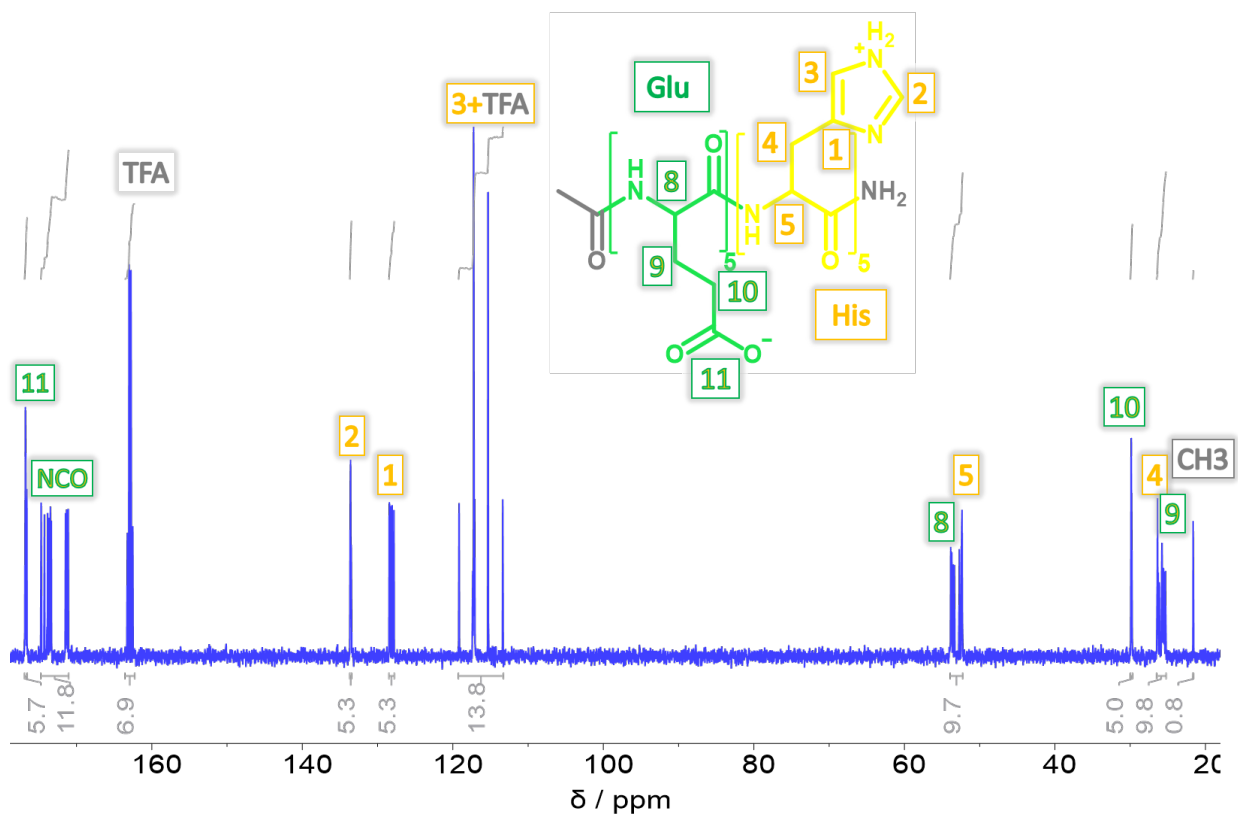
Figure S11: Quantitative analysis of the ^{13}C NMR spectra of Glu$_5$ − His$_5$ to estimate the TFA content.

about 5 to 6 is an equivalent of one amino acid side-chain. If there was a stoichiometric equivalent of TFA, then one would expect that the same holed for peak corresponding to TFA. However, the peak at 165 ppm (COO group), which doesn't overlap with anything and should have integral 5, but it has almost 7. The signal around 120 ppm (CF_3) overlaps with the signal of peptide. Nevertheless, one would expect the integral of both peak to be between 10 and 12, whereas its actual value is almost 14. Thus, we can conclude that the excess of TFA is roughly $2/5 = 0.4$ times the amount of His groups, which is commensurate with shifting the charge on $Glu_5 - His_5$ determined from titrations by approximately 2 charges per chain (Fig.S10a) .

2.4.3 Degree of ionisation from the NMR spectra

Fig. S12 and S15 show ^{1}H and ^{13}C spectra for Glu-His and Lys-Asp, respectively, at pH ranging from 1 to 13, with the chemical structure of the amino acids and the assignment of all peaks. 2D NMR spectra, COSY and ^{1}H-^{13}C HSQC, (Fig. S18 and S19) at pH 2 were used for peak assignment. Specific atoms have been identified in literature as "good reporters" of ionisation, that is, their chemical shifts predominantly reflect ionisation changes on the nearby ionisable group.[S14] Typically, good reporters are located far from the backbone and as close as possible to the ionisable group. Following Ref.,[S14] we used CB for Asp, CD and CG for Glu, CG and CE1 for His, and CD and CE for Lys. We were able to identify two good reporters for each amino acid except for Asp, for which we were able to identify only one good reporter because the signal of CG overlaps with carbonyl signals from the backbone. Details of the spectra, highlighting how the peaks shift with the pH, are shown in Fig.S13, Fig.S14, Fig.S16 and Fig.S17.

The chemical shifts of the good reporters, which could be unambiguously identified in the spectra, were used to calculate the degree of ionisation of each amino acid in the oligopeptide. These peaks typically consisted of multiple sub-peaks, reflecting the fact that same amino acids in different positions in the peptide chain were not equivalent. To determine the
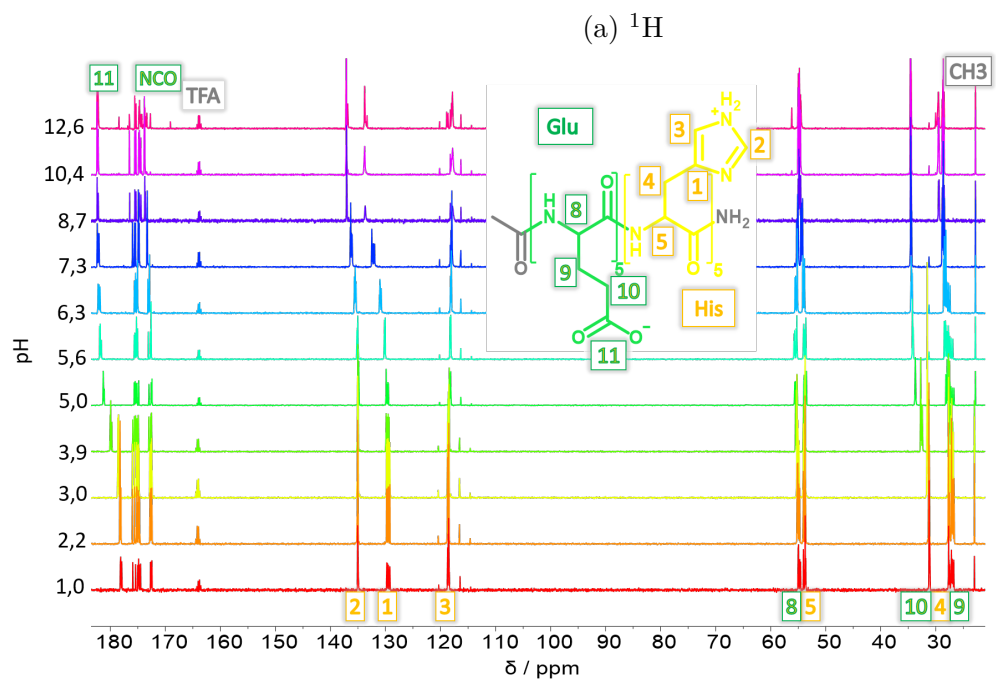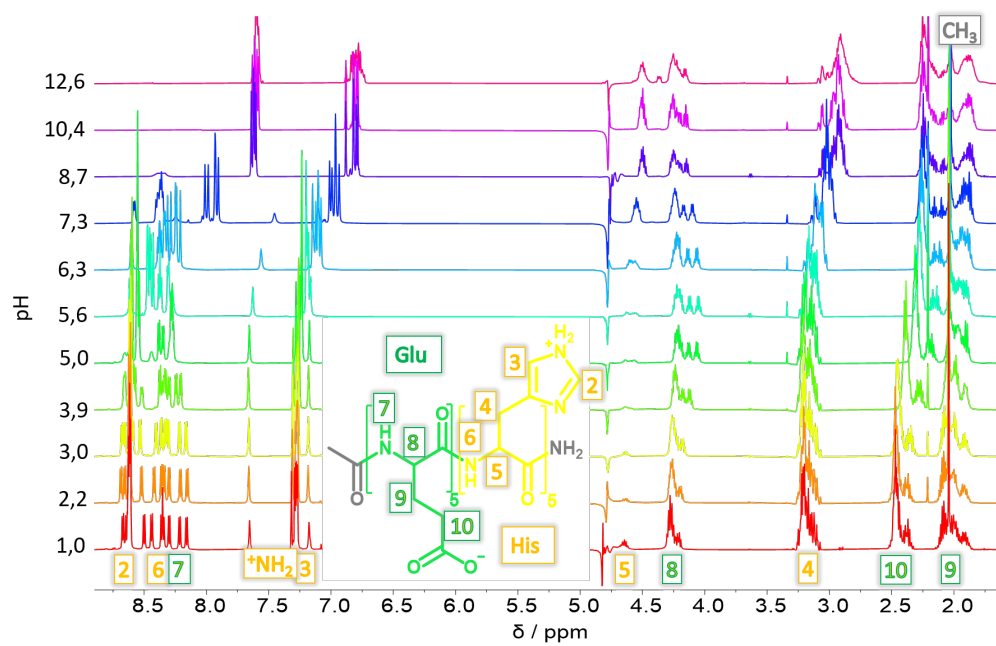
(a) ^{1}H

(b) ^{13}C

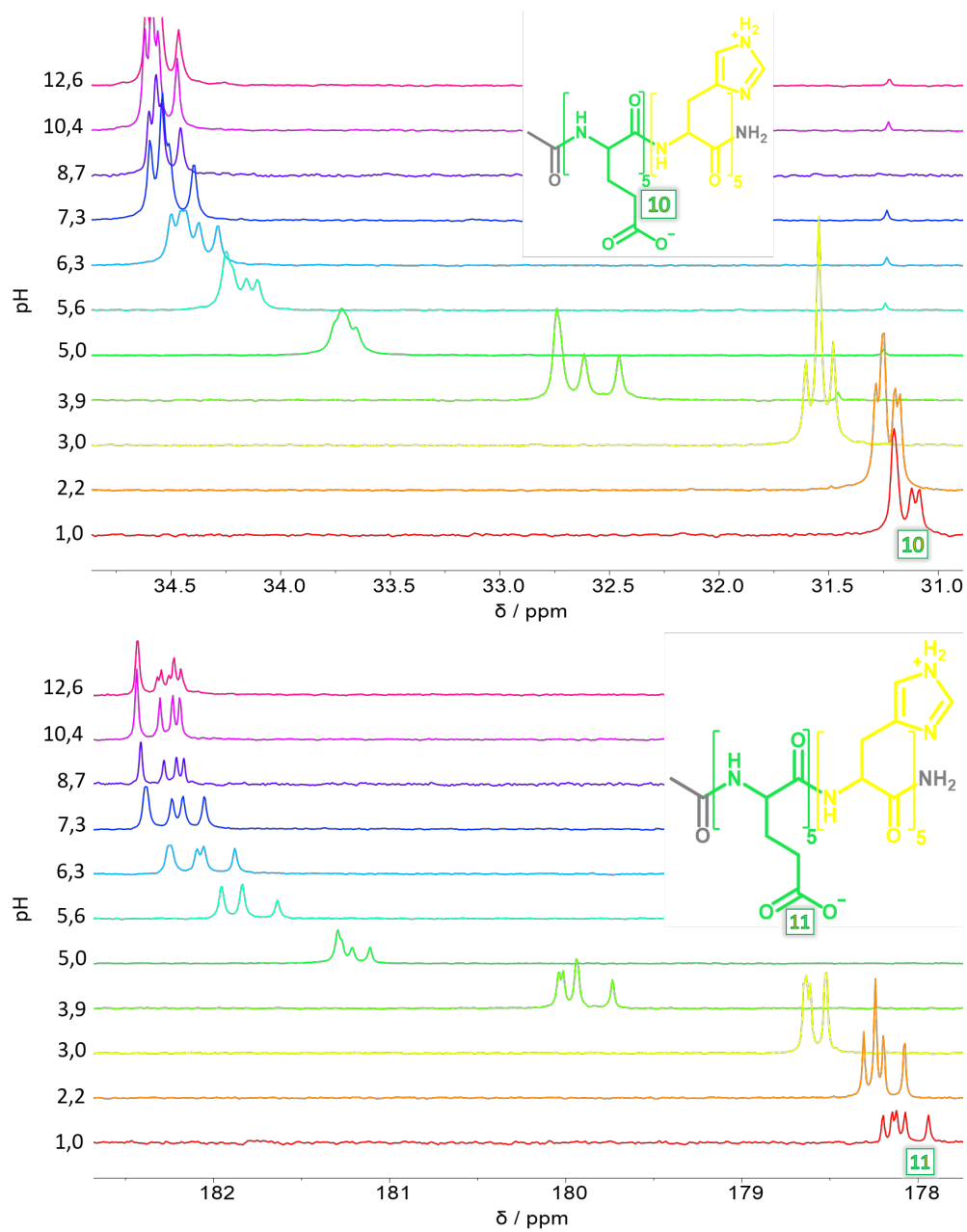Figure S12: NMR spectra of $Glu_5 - His_5$ at various pH.

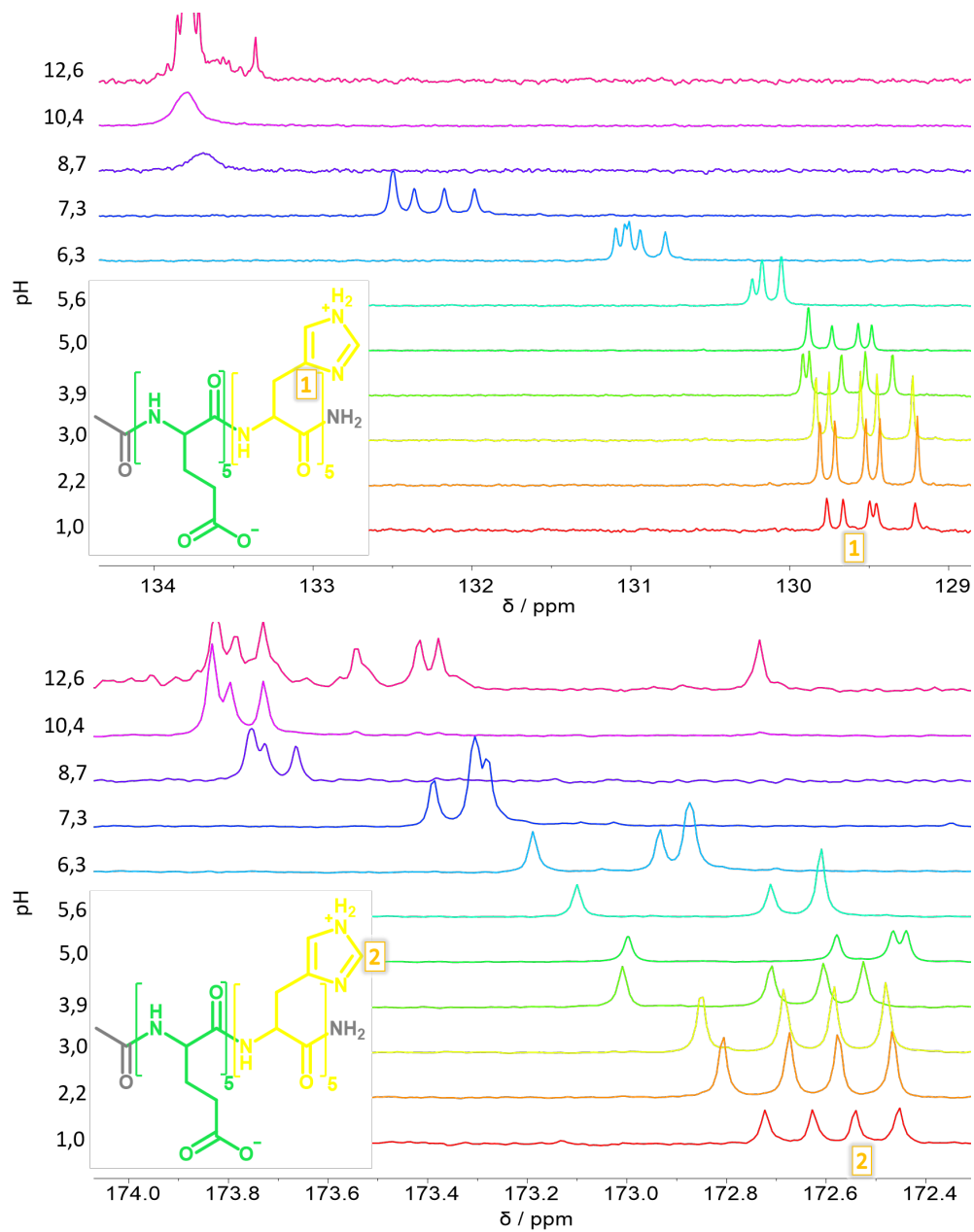Figure S13: Details of NMR spectra of good reporters for Glu in $Glu_5 - His_5$ at various pH.

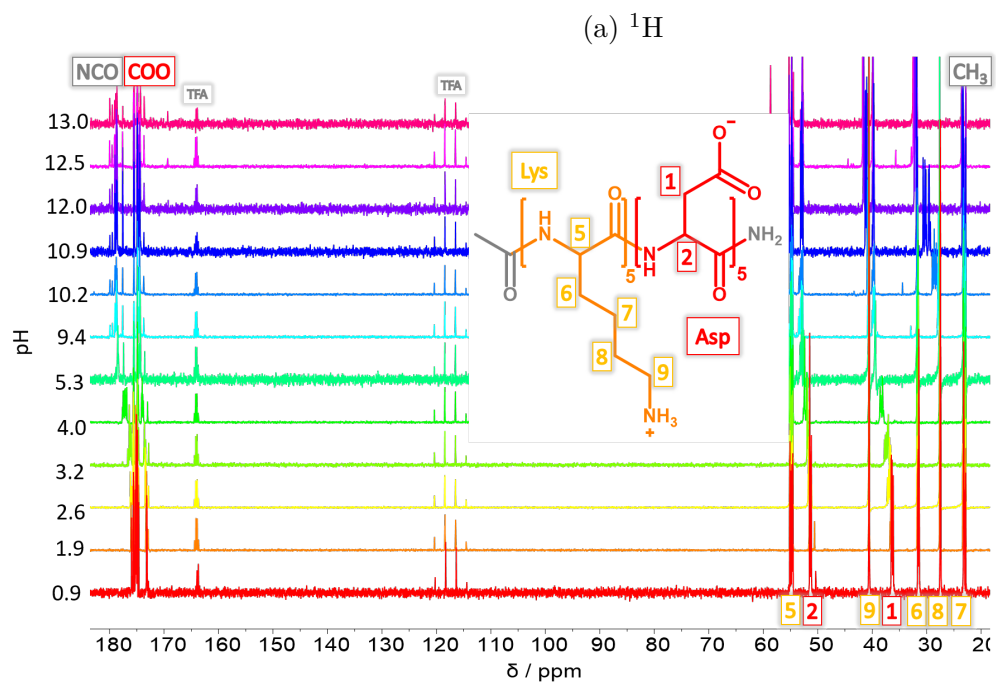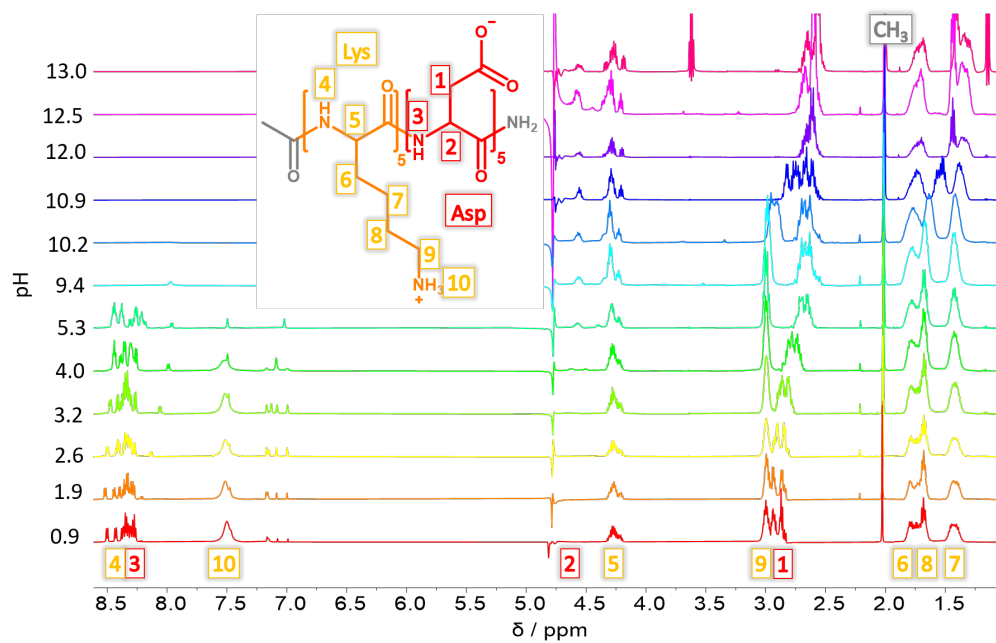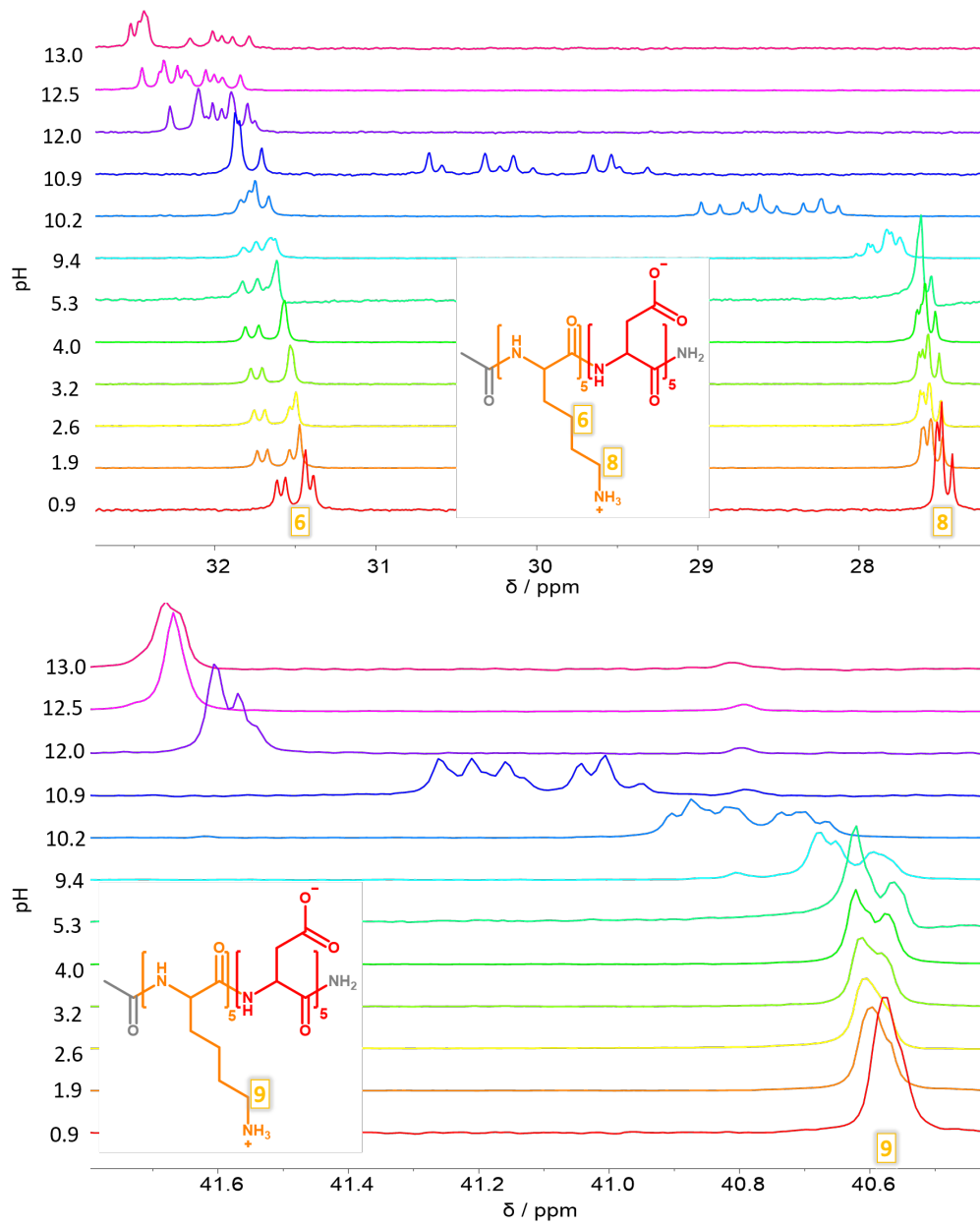Figure S14: Details of NMR spectra of good reporters for His in $Glu_5 - His_5$ at various pH.

(a) ^{1}H

(b) ^{13}C

Figure S15: NMR spectra of $\mathrm{Lys}_5 - \mathrm{Asp}_5$ at various pH.

Figure S16: Details of NMR spectra of good reporters for Lys in $\mathrm{Lys}_5 - \mathrm{Asp}_5$ at various pH.
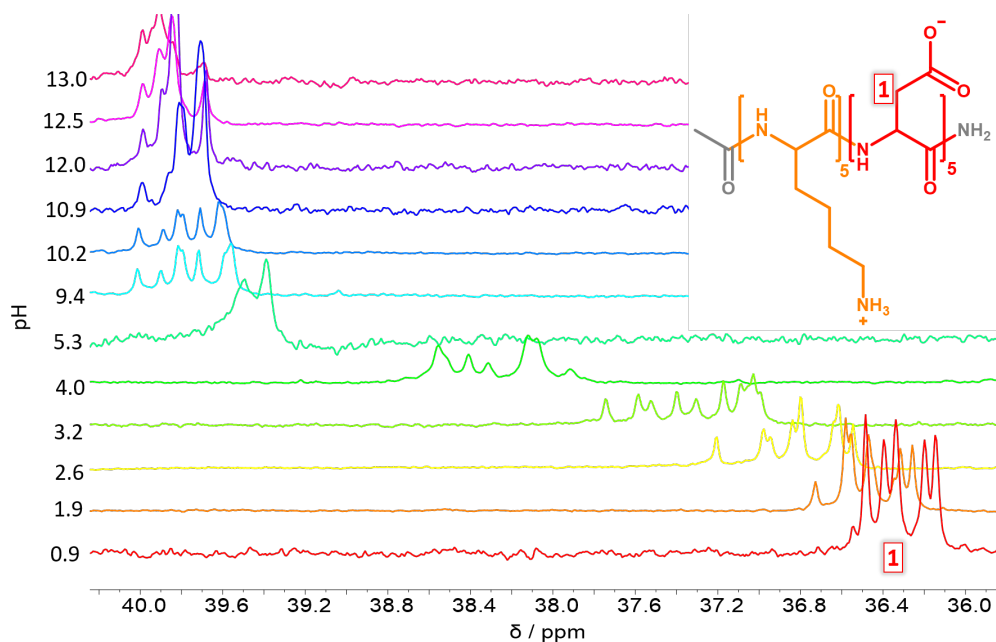
Figure S17: Details of NMR spectra of good reporters for Asp in $\text{Lys}_5 - \text{Asp}_5$ at various pH.

average degree of ionisation of each type of amino acid, we use the centre of mass of the corresponding peak, which should be equivalent to averaging the degree of ionisations of 5 amino acids of the same type. We calculated the degree of ionisation by normalising the chemical shifts as follows:

$$\alpha_{base} = \frac{\delta_{max} - \delta}{\delta_{max} - \delta_{min}} \tag{5}$$

for bases (Lys and His) and

$$\alpha_{acid} = \frac{\delta - \delta_{min}}{\delta_{max} - \delta_{min}} \tag{6}$$
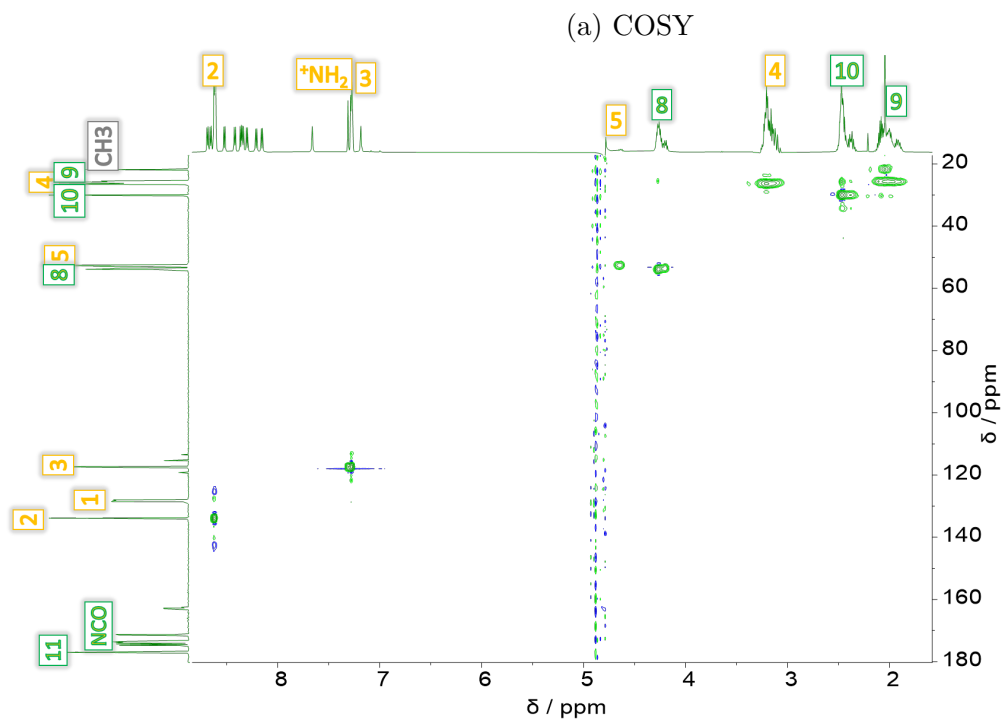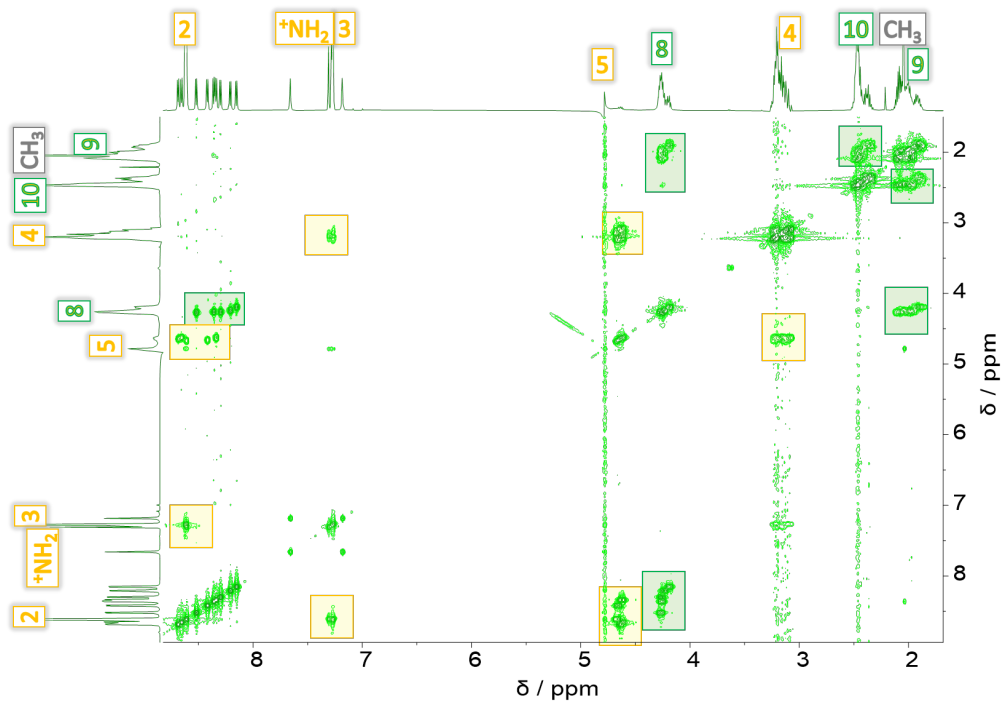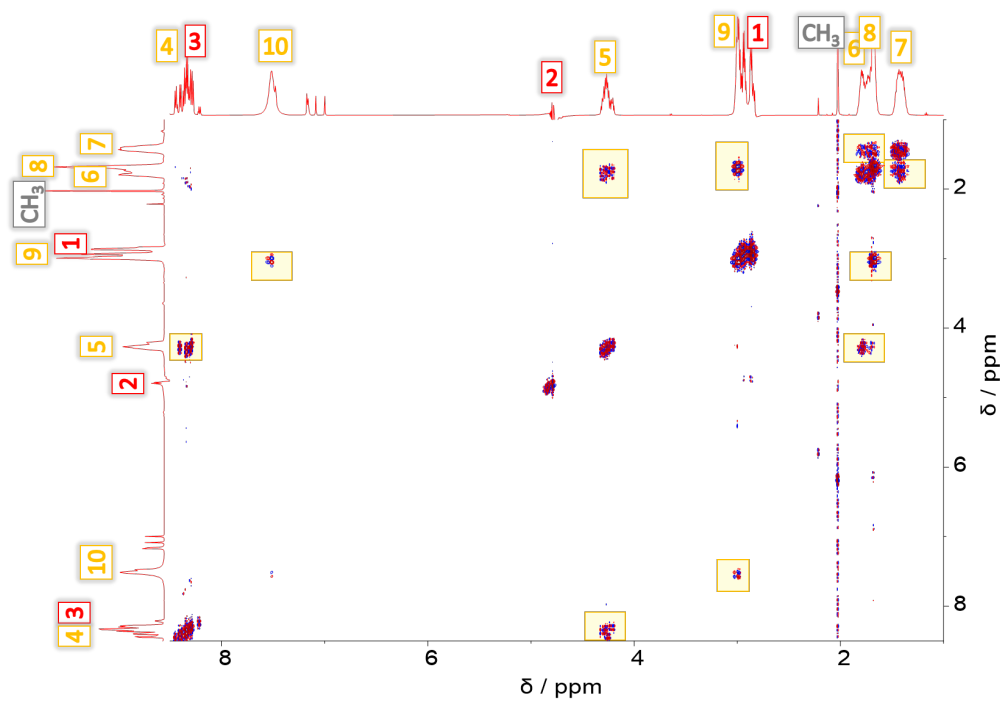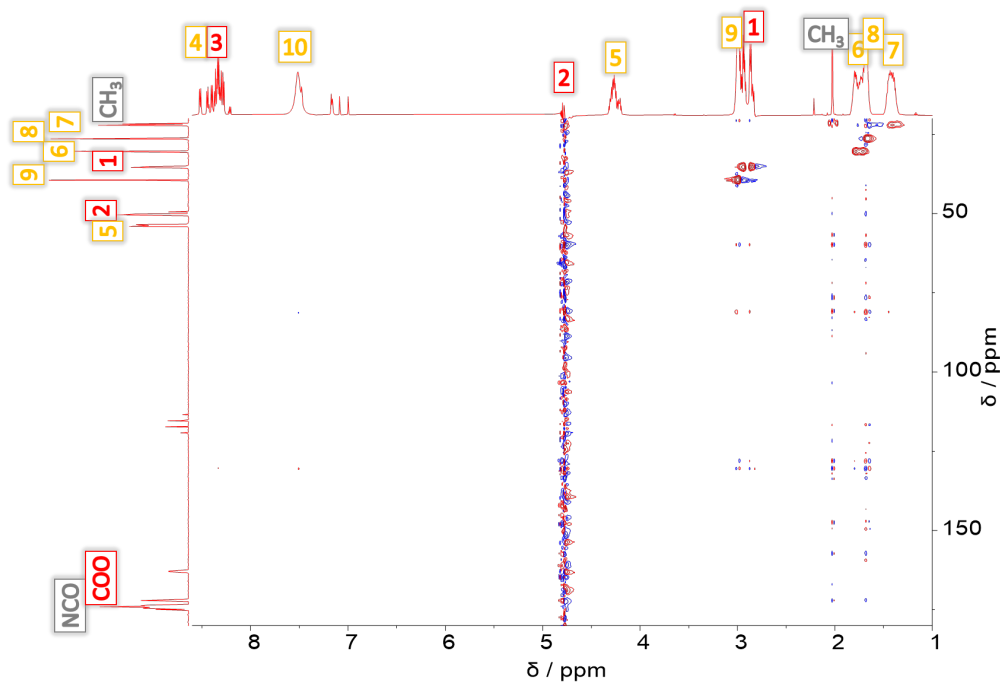
for acids (Asp and Glu).

(a) COSY

(b) ^{1}H-^{13}C HSQC

Figure S18: 2D NMR spectra of $Glu_5 - His_5$ at pH=2.

(a) COSY

(b) ^{1}H-^{13}C HSQC

Figure S19: 2D NMR spectra of $\text{Lys}_5 - \text{Asp}_5$ at pH=2.
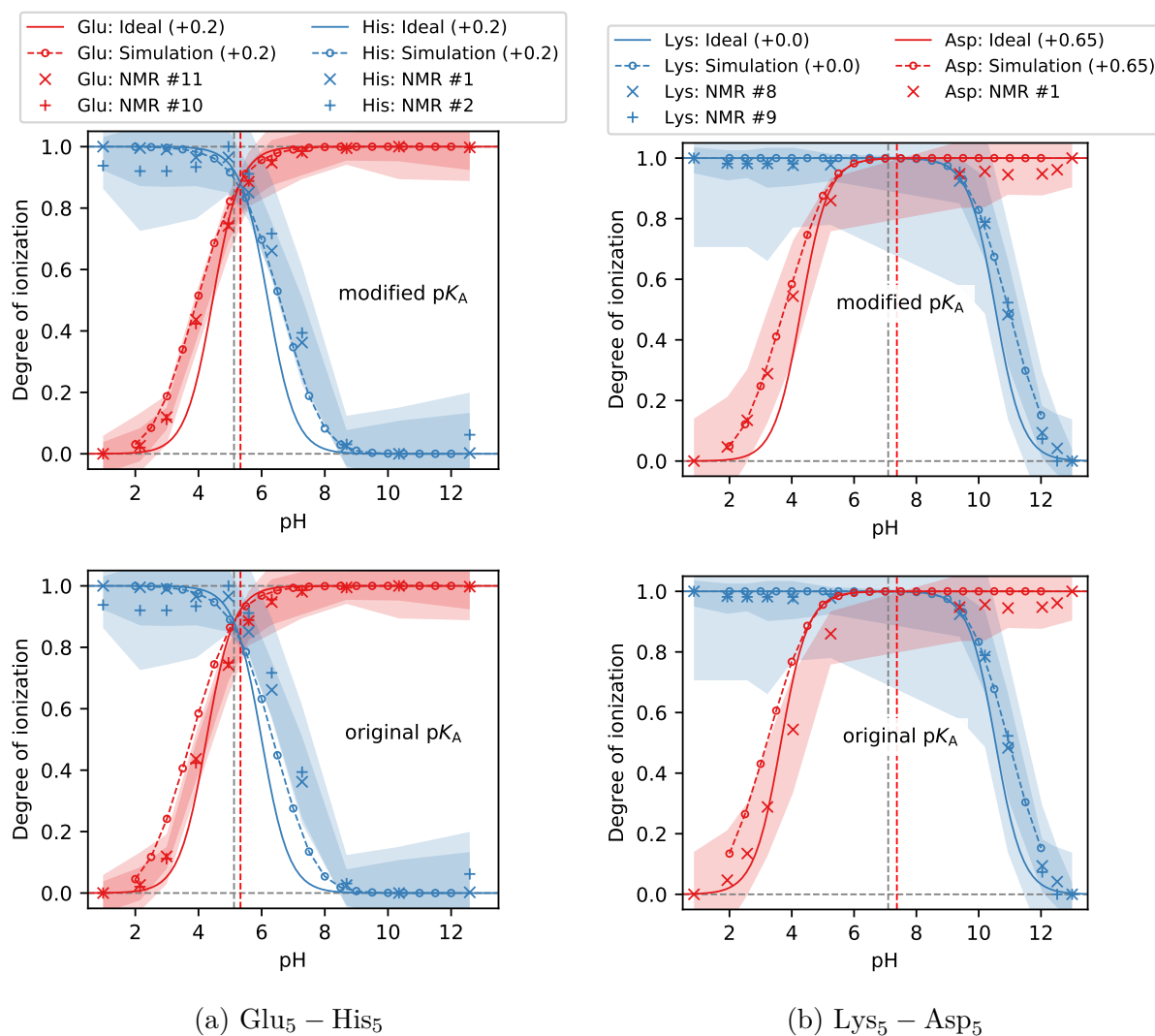
S-37

(a) $Glu_5 - His_5$

(b) $Lys_5 - Asp_5$

Figure S20: The degree of ionisation of acid and base groups on the peptides predicted from simulations, determined from NMR measurements and from the Henderson-Hasselbalch equation. Individual reporter atoms are indicated in the legend. Shaded areas indicate the spread of peaks in the NMR spectra. Red and grey vertical lines represent the isoelectric point determined from CZE and the ideal isoelectric point. Panels in the top row show the ideal and simulation data using the modified pK_A values to correct for the effect of incorporating amino acids into the peptide (see Table S7). Numbers in the legend indicate the reporter atom id in NMR (see Fig. S12 and Fig. S15) and the amount by which the pK_A values were modified in the simulations. The bottom row shows the original uncorrected data for comparison (denoted as "Lit1" in Table S7).

Fig.S20 shows that the degree of ionisations determined using different good reporters agree with each other. However, individual reporters may considerably differ in the spread of the peaks due to the non-equivalence of amino acids of the same type. The panels in the bottom row of Fig.S20a show that the degree of ionisations of both Glu and His exhibit the same trend as that assessed in simulations, albeit shifted to higher pH values, in line with the shift of the isoelectric point between the ideal and CZE results ΔpI. In contrast, Fig.S20b shows that the degree of ionisation of Asp is shifted with respect to the simulation results approximately twice as much as the difference in isoelectric points, while the degree of ionisation of Lys matches the simulations. We attributed these differences to the uncertainty in the pK_A values caused by different substituents. To assess this hypothesis, we performed a new set of CG simulations using a modified set of pK_A values: In $Glu_5 - His_5$, we increased the pK_A of both Glu and His by ΔpI; In $Lys_5 - Asp_5$, we increased the pK_A of Lys by 2ΔpI (see Table S7). Consequently, the isoelectric point from simulations using the modified pK_A values matches pI_{cze}. With the modified pK_A, we obtained an almost perfect match between the simulation and the NMR results, as shown in the top row of the panels in Fig.S20b. This observation supports our hypothesis that the previously observed differences between the pI from simulations and experiments could be attributed to the uncertainty in choosing the right pK_A values for the simulation.

As an alternative hypothesis, one could claim that using different literature sources of

Table S7: Difference between the pK_A values of free amino acids reported in the literature and our estimates of the pK_A values of the same amino acids incorporated in the peptides $Glu_5 - His_5$ and $Lys_5 - Asp_5$.

Abbreviation	Glu	His	Asp	Lys	Source
Original pK_A (Lit1)	4.25	6.00	3.65	10.53	CRC Handbook 1991[S15]
Lit2	4.30	6.00	3.90	10.80	Concepts in Biochemistry 1988[S16]
Lit3	4.15	6.04	3.71	10.67	CRC Handbook 2015[S17]
Modified pK_A	4.45	6.20	4.21	10.53	Our estimate from NMR, Fig. S20

pK_A values of free amino acids might have the same effect as the modifications proposed above because the values reported in the literature are not entirely consistent. Table S7 outlines the pK_A values from several literature sources and the modified pK_A values obtained as described in the previous paragraph. The source labeled "Lit1" was used for the original pK_A values in the manuscript, while the other sources were included only for comparison. The reported pK_A values for His and Glu are quite consistent among different sources and do not vary by more than 0.1, while our estimation suggests that the modified pK_A values should be higher than the original values by $\Delta pI(Glu_5 - His_5) \approx 0.2$. The reported p$K_A$ values for Lys and Asp are less consistent among different sources and do not vary by more than 0.1. Our estimation suggests that the modified pK_A values of Asp should be higher than the original value by $2\Delta pI(Glu_5 - His_5) \approx 0.56$, while the p$K_A$ of Lys should remain unchanged. Thus, the differences between the original and modified pK_A values are greater than the inconsistencies in pK_A values reported in various literature sources.

To assess how different pK_A values from the literature might affect our simulation results, we performed a set of CG simulations using each literature source listed in Table S7. Fig. S21 shows that simulations performed with these sets of pK_A values differ only marginally, with much larger systematic differences between all simulation and CZE results. Thus, we conclude that the differences between simulations and NMR or CZE results cannot be attributed to the uncertainty in choosing a literature source of pK_A values of free amino acids. Instead, they should be attributed to a systematic shift in pK_A caused by the replacement of some substituents on the amino acids upon their incorporation into the peptide.
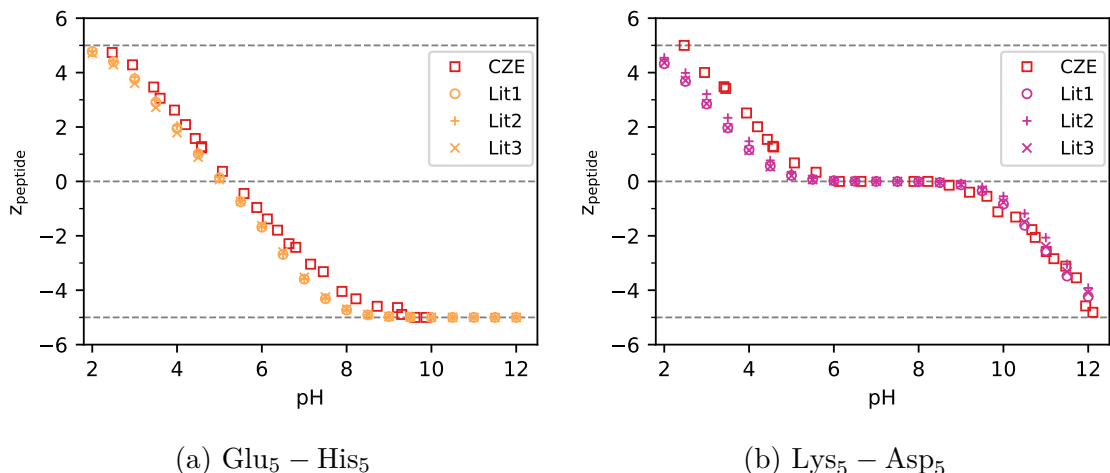
(a) $Glu_5 - His_5$　　　　　　　　(b) $Lys_5 - Asp_5$

Figure S21: Simulation predictions of the total charge of the peptides, compared with experimental data from capillary zone electrophoresis (CZE). The pK_A values used in individual simulations are listed in Table S7.

2.4.4 Diffusion coefficients from DOSY NMR

The DOSY experiment made it possible to calculate the diffusion coefficients, D, of the oligopeptides as function of pH using GNAT software by integrating the peaks. Fig. S22 shows that the values of diffusion coefficients, $D(pH)$, do not exhibit any visible trend, and they do not deviate from the average value beyond the range of the estimated error. The error bar was determined as a standard deviation of D determined from each individual peak of the spectrum. Thus, we conclude that the diffusion coefficients are approximately constant. In the whole pH range, they do not deviate from the average values by more than 10%.

References

(S1) Janke, W. Statistical Analysis of Simulations: Data Correlations and Error Estimation. *Quantum Simulations of Complex Many-Body Systems: from Theory to Algorithms* **2002**, 423–445.
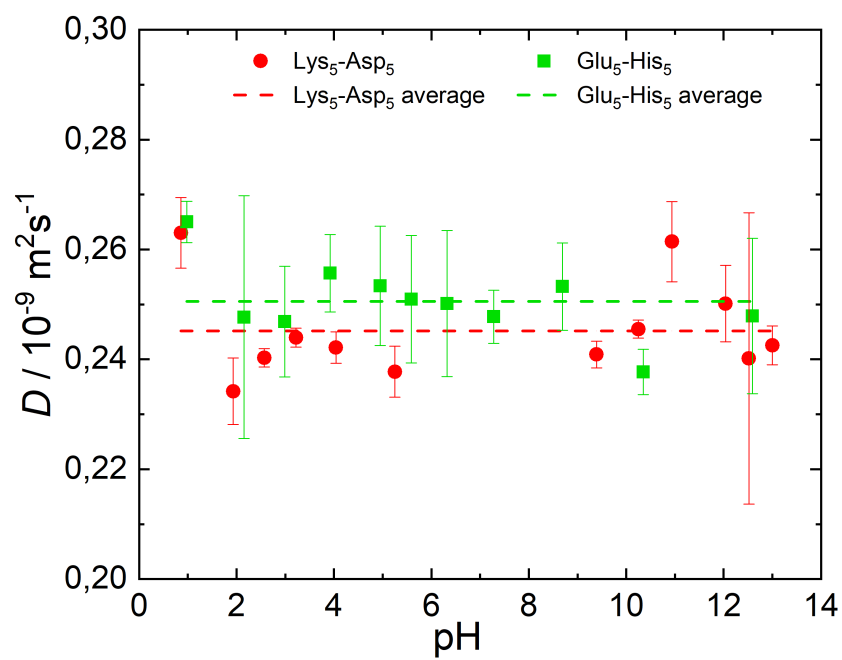
(S2) http://www.gromacs.org/.

Figure S22: Diffusion coefficients for peptides as a function of pH, determined from DOSY NMR. Horizontal lines represent the diffusion coefficient averaged over all pH values.

(S3) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25, DOI: `10.1016/j.softx.2015.06.001`.

(S4) `https://avogadro.cc`.

(S5) Hruška, V.; Svobodová, J.; Beneš, M.; Gaš, B. A nonlinear electrophoretic model for PeakMaster: Part III. Electromigration dispersion in systems that contain a neutral complex-forming agent and a fully charged analyte. Theory. *Journal of Chromatography A* **2012**, *1267*, 102–108, DOI: `10.1016/j.chroma.2012.06.086`.

(S6) Beneš, M.; Svobodová, J.; Hruška, V.; Dvořák, M.; Zusková, I.; Gaš, B. A nonlinear electrophoretic model for PeakMaster: Part IV. Electromigration dispersion in systems that contain a neutral complex-forming agent and a fully charged analyte. Experimental verification. *Journal of Chromatography A* **2012**, *1267*, 109–115, DOI: `10.1016/j.chroma.2012.06.053`.

(S7) Dubský, P.; Ördögová, M.; Malý, M.; Riesová, M. CEval: All-in-one software for data processing and statistical evaluations in affinity capillary electrophoresis. *Journal of Chromatography A* **2016**, *1445*, 158–165, DOI: `10.1016/j.chroma.2016.04.004`.

(S8) Lobaskin, V.; Dünweg, B.; Holm, C. Electrophoretic mobility of a charged colloidal particle: a computer simulation study. *Journal of Physics: Condensed Matter* **2004**, *16*, S4063–S4073, DOI: `10.1088/0953-8984/16/38/021`.

(S9) Hill, R. J.; Saville, D.; Russel, W. Electrophoresis of spherical polymer-coated colloidal particles. *Journal of Colloid and Interface Science* **2003**, *258*, 56–74, DOI: `10.1016/s0021-9797(02)00043-7`.

(S10) Griffiths, P. C.; Paul, A.; Stilbs, P.; Petterson, E. Charge on Poly(ethylene

imine): Comparing Electrophoretic NMR Measurements and pH Titrations. *Macro-molecules* **2005**, *38*, 3539–3542, DOI: `10.1021/ma0478409`.

(S11) Hwang, T.; Shaka, A. Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *Journal of Magnetic Resonance, Series A* **1995**, *112*, 275–279, DOI: `10.1006/jmra.1995.1047`.

(S12) Jerschow, A.; Müller, N. Suppression of Convection Artifacts in Stimulated-Echo Diffusion Experiments. Double-Stimulated-Echo Experiments. *Journal of Magnetic Resonance* **1997**, *125*, 372–375, DOI: `10.1006/jmre.1997.1123`.

(S13) Castañar, L.; Poggetto, G. D.; Colbourne, A. A.; Morris, G. A.; Nilsson, M. The GNAT: A new tool for processing NMR data. *Magnetic Resonance in Chemistry* **2018**, *56*, 546–558, DOI: `10.1002/mrc.4717`.

(S14) Hass, M. A.; Mulder, F. A. Contemporary NMR Studies of Protein Electrostatics. *Annual Review of Biophysics* **2015**, *44*, 53–75, DOI: `10.1146/annurev-biophys-083012-130351`, tex.ids: hass2015a.

(S15) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 72nd ed.; CRC Press: New York, 1991.

(S16) Stephenson, W. K. *Concepts in Biochemistry*, 3rd ed.; Wiley: New York, 1988.

(S17) Haynes, W. *CRC Handbook of Chemistry and Physics*, 96th ed.; CRC Press: New York, 2015.