Supporting Information for

Deep Learning for Prediction and Optimization of Fast-Flow Peptide Synthesis

Authors: Somesh Mohapatra[1,†], Nina Hartrampf[2,†,#], Mackenzie Poskus[2], Andrei Loas[2], Rafael Gómez-Bombarelli[1,*], Bradley L. Pentelute[2,3,4,5,*]

[1]Massachusetts Institute of Technology, Department of Materials Science and Engineering, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[2]Massachusetts Institute of Technology, Department of Chemistry, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[3]Massachusetts Institute of Technology, Koch Institute, Broad Institute of Harvard and MIT, Center for Environmental Health Sciences, Cambridge, MA, USA

[4]Center for Environmental Health Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[5]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA

[†]These authors contributed equally to this work.

[#]Current address: University of Zurich, Department of Chemistry, Winterthurerstrasse 190, 8057 Zurich, Switzerland

*Correspondence to: blp@mit.edu, rafagb@mit.edu

1 Table of Contents

2 Materials and general methods

2.1 Reagents and solvents.

All reagents were purchased and used as received. Fmoc-protected amino acids (Fmoc-Ala-OH$\times H_2O$, Fmoc-Arg(Pbf)-OH; Fmoc-Asn(Trt)-OH; Fmoc-Asp-(O*t*-Bu)-OH; Fmoc-Cys(Trt)-OH; Fmoc-Gln(Trt)-OH; Fmoc-Glu(O*t*-Bu)-OH; Fmoc-Gly-OH; Fmoc-His(Trt)-OH; Fmoc-Ile-OH; Fmoc-Leu-OH; Fmoc-Lys(Boc)-OH; Fmoc-Met-OH; Fmoc-Phe-OH; Fmoc-Pro-OH; Fmoc-Ser(But)-OH; Fmoc-Thr(*t*-Bu)-OH; Fmoc-Trp(Boc)-OH; Fmoc-Tyr(*t*-Bu)-OH; Fmoc-Val-OH), Fmoc-His(Boc)-OH and backbone protected amino acids were purchased from the Novabiochem-line from Sigma Millipore; O-(7-azabenzotriazol-1-yl)-*N*,*N*,*N*',*N*'-tetramethyluronium hexafluorophosphate (HATU, ≥97.0%), and (7-azabenzotriazol-1-yloxy)tripyrrolidinophosphonium hexa-fluorophosphate (PyAOP, ≥97.0%) were purchased from P3 Biosystems. Biosynthesis OmniSolv® grade *N*,*N*-dimethylformamide (DMF) was purchased from EMD Millipore (DX1732-1). *N*-Methyl-2-pyrrolidone (NMP, ≥99.0%) was purchased from Sigma-Aldrich and dried over PPT Pure Process Technology solvent system. AldraAmine trapping agents (for 1000 – 4000 mL DMF, catalog number Z511706), Diisopropylethylamine (DIEA; 99.5%, biotech grade, catalog number 387649), piperidine (ACS reagent, ≥99.0%), trifluoroacetic acid (HPLC grade, ≥99.0%), triisopropylsilane (≥98.0%), acetonitrile (HPLC grade), formic acid (FA, ≥95.0%) and 1,2-ethanedithiol (EDT, GC grade, ≥98.0%) were purchased from Sigma-Aldrich. H-Rink Amide (0.49 mmol/g and 0.18 mmol/g loading) and HMPB ChemMatrix polyethylene glycol (0.45 mmol/g loading) resin were purchased from PCAS Biomatrix. Water was deionized using a Milli-Q Reference water purification system (Millipore). Nylon 0.22 µm syringe filters were TISCH brand SPEC17984.

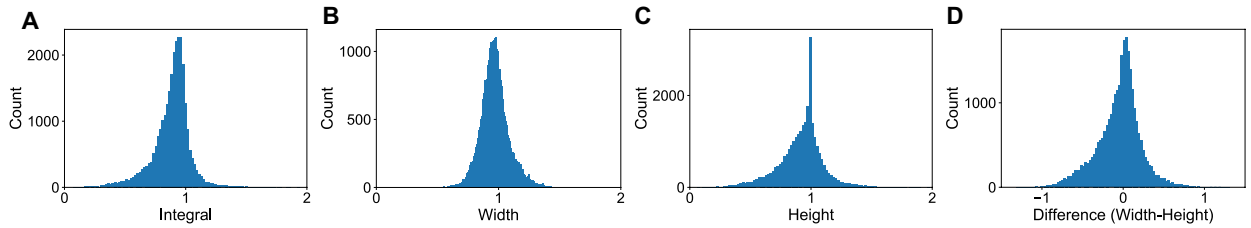
3 Deep learning and optimization

3.1 Data set distribution

The pre-processing of the data set and the number of reaction steps that remained after each step is as follows –

- Raw Dataset – 35427 steps
- Removal of unnatural amino acids – 35327 steps
- Removal of data points lying outside 2 standard deviations across all parameters – 33581 steps
- Removal of reaction steps with missing values – 33565 steps
- Removal of reaction steps for the following conditions – coupling temperature and reactor temperature greater than 200 ^{O}C, deprotection strokes less than 5, flow rate other than 40 ml/min and 80 ml/min – 33159 steps
- Removal of reaction steps without complete syntheses, i.e. missing one or more amino acid from the peptide sequence in the list of reaction steps – 28642 steps
- Averaging over reaction steps with same pre-chain, incoming amino acid and synthesis parameters – 17459 steps

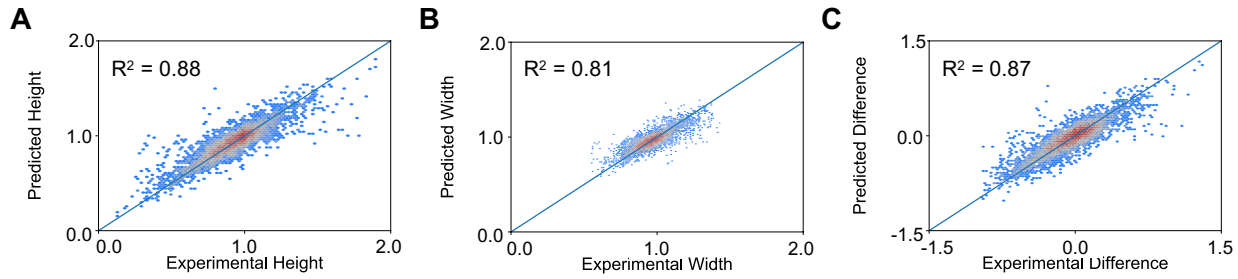
For the processed data set, where all parameters are normalized by the first deprotection step, we see a Gaussian-like distribution for each of the parameters (**SI Figure 1**).



SI Figure 1. Distribution of the data normalized by the first deprotection step for different parameters – **A.** Integral, **B.** Width, **C.** Height and **D.** Difference (Width - Height).

3.2 Training of the model

The model performance is shown in SI Figure 2.



SI Figure 2. The model predictions for a particular reaction step are within 14% error on the validation dataset for **A.** Height, **B.** Width and **C.** Difference.

SI Table 1. The minimum, maximum, standard deviation values, and model performance metrics for different parameters in root mean squared error (RMSE) and % error (RMSE/range) are listed.

Parameter	Minimum	Maximum	Standard Deviation	Training Loss, RMSE	Validation Loss, RMSE	% Error
Area	0.10	1.94	0.15	0.06	0.08	4.12
Width	0.11	1.90	0.18	0.06	0.10	5.76
Height	0.55	1.44	0.12	0.05	0.05	6.07
Difference	-1.07	1.31	0.27	0.09	0.13	5.38

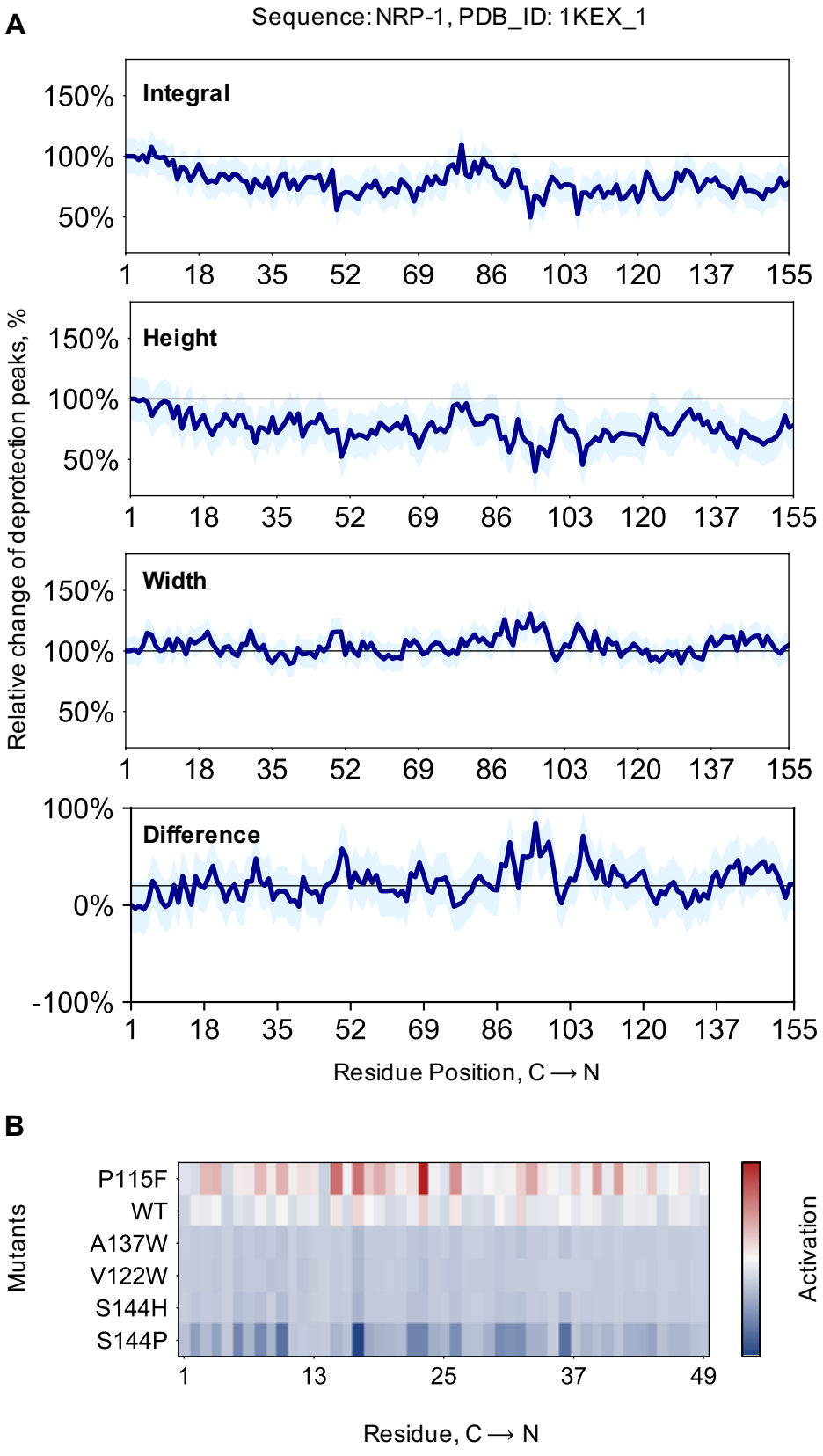
3.3 Automated pipeline for optimization of difficult-to-synthesize sequences

An automated pipeline was set-up for prediction of traces for classical difficult-to-synthesize sequences.[1,2] The individual traces were predicted using the model (**SI Section 3.4**). Gradient activation maps averaged over bit-vectors were obtained for mutants with less aggregation and most aggregating (for negative control) than wild-type sequence (**SI Section 3.4**).

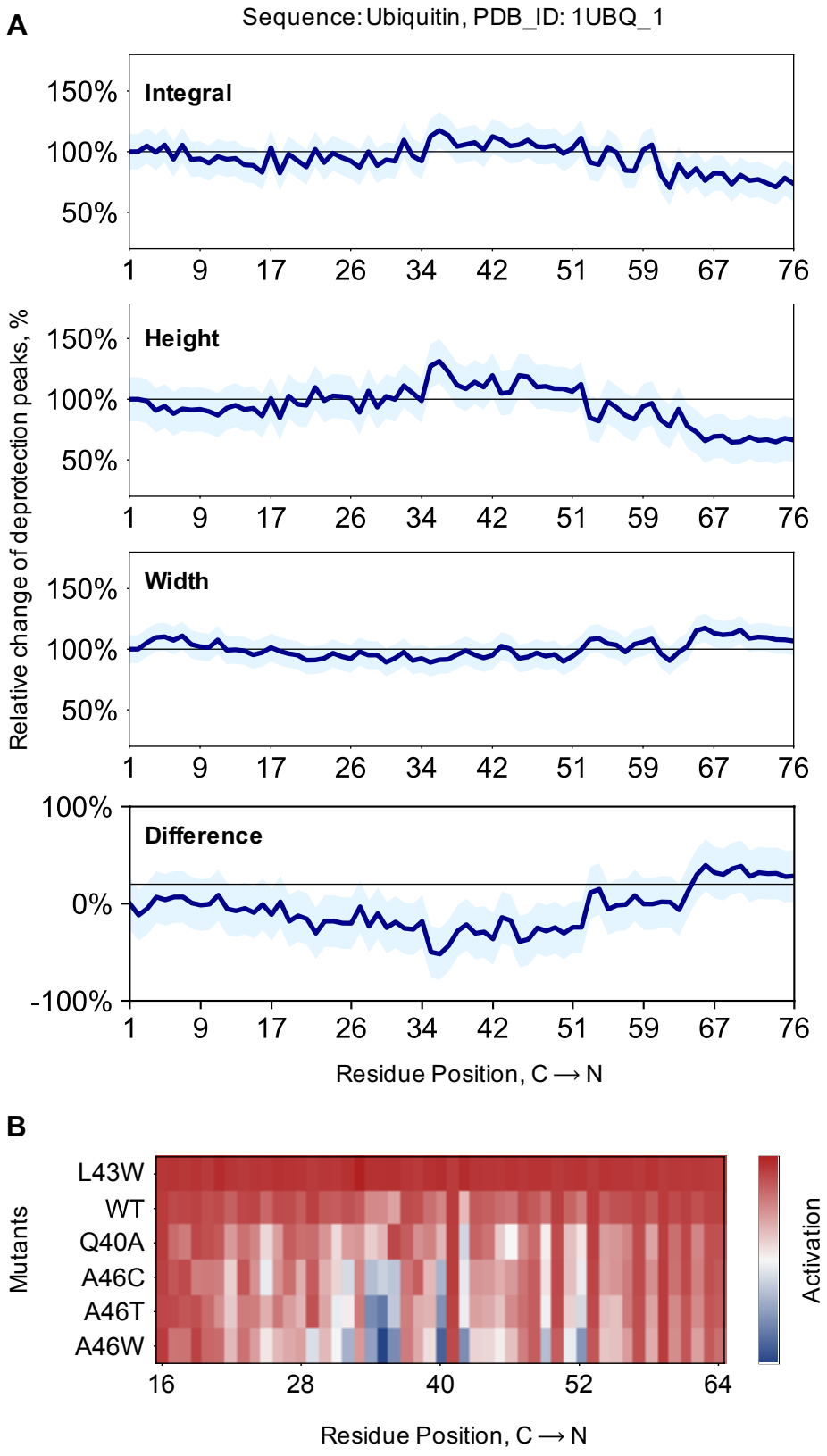
3.4 Prediction of traces and mutants for difficult-to-synthesize sequences

Each figure consists of **A.** predicted integral, width, height and difference traces with error range (1 standard deviation), and **B.** gradient maps for negative control, wild-type and less aggregating sequences, with activation color bar ranging from red to blue indicating the residues contributing most to least towards aggregation for that particular coupling-deprotection step.

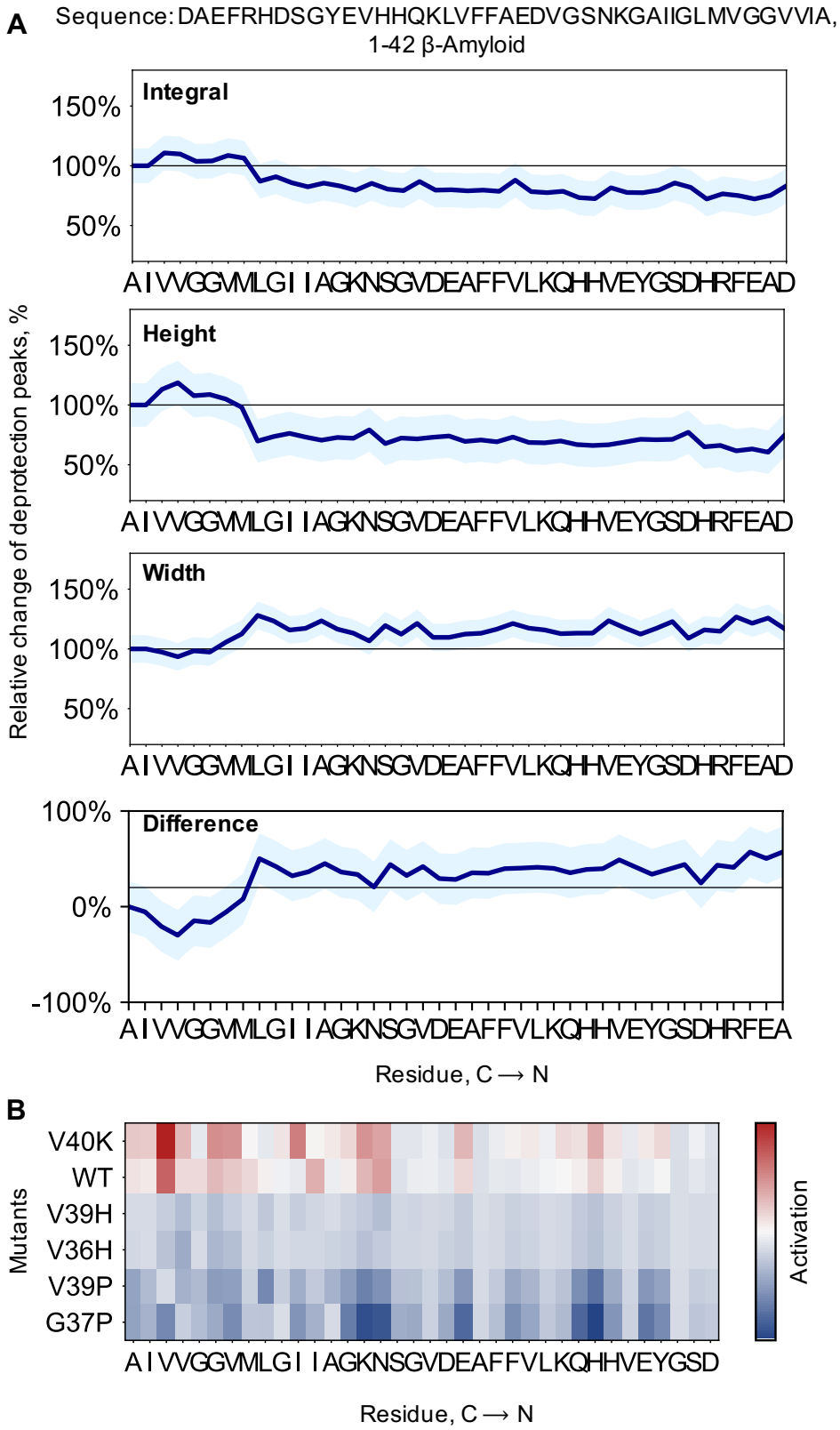
3.4.1 NRP-1, PDB_ID: 1KEX_1



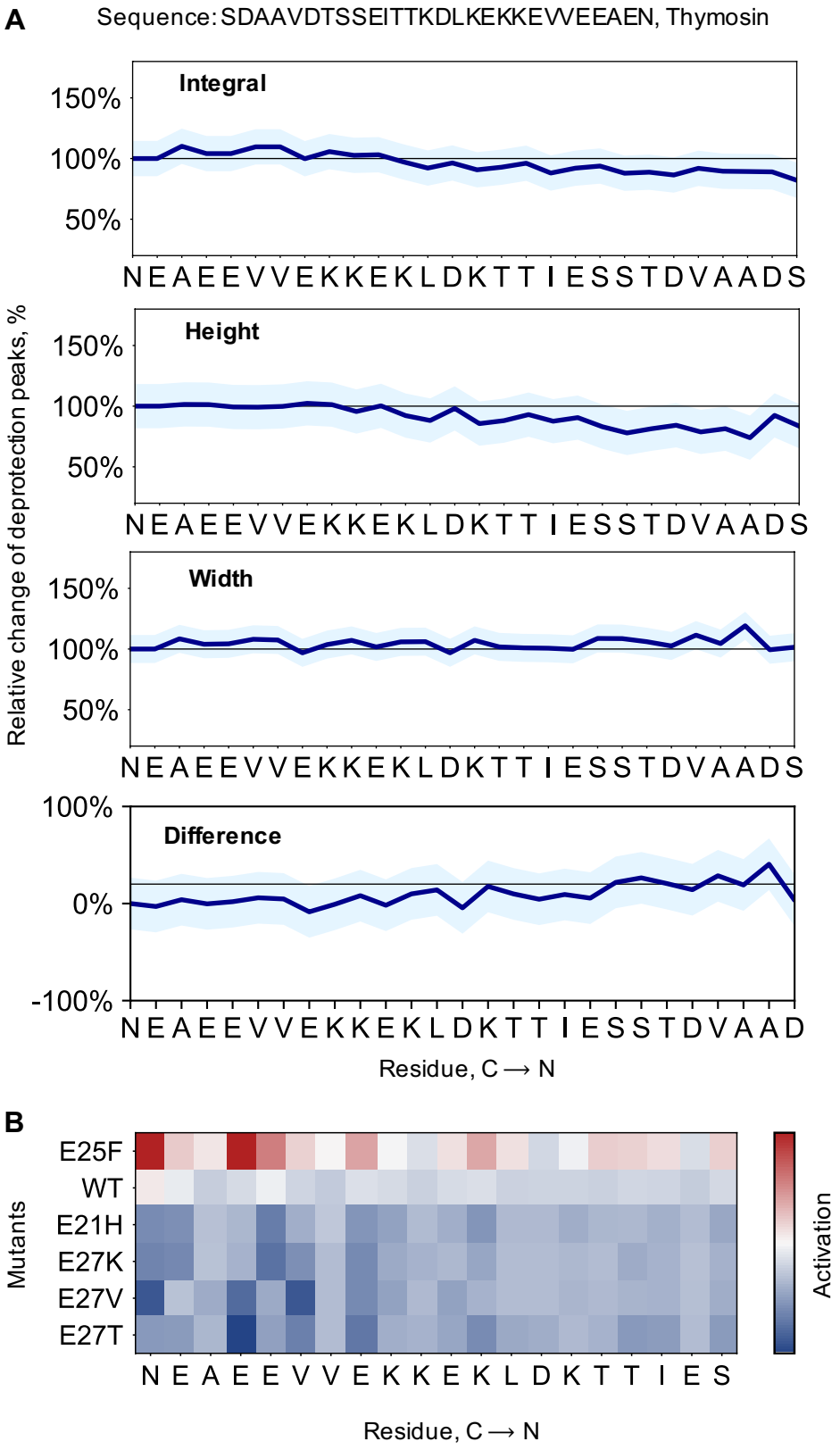
3.4.2 Ubiquitin, PDB_ID: 1UBQ_1



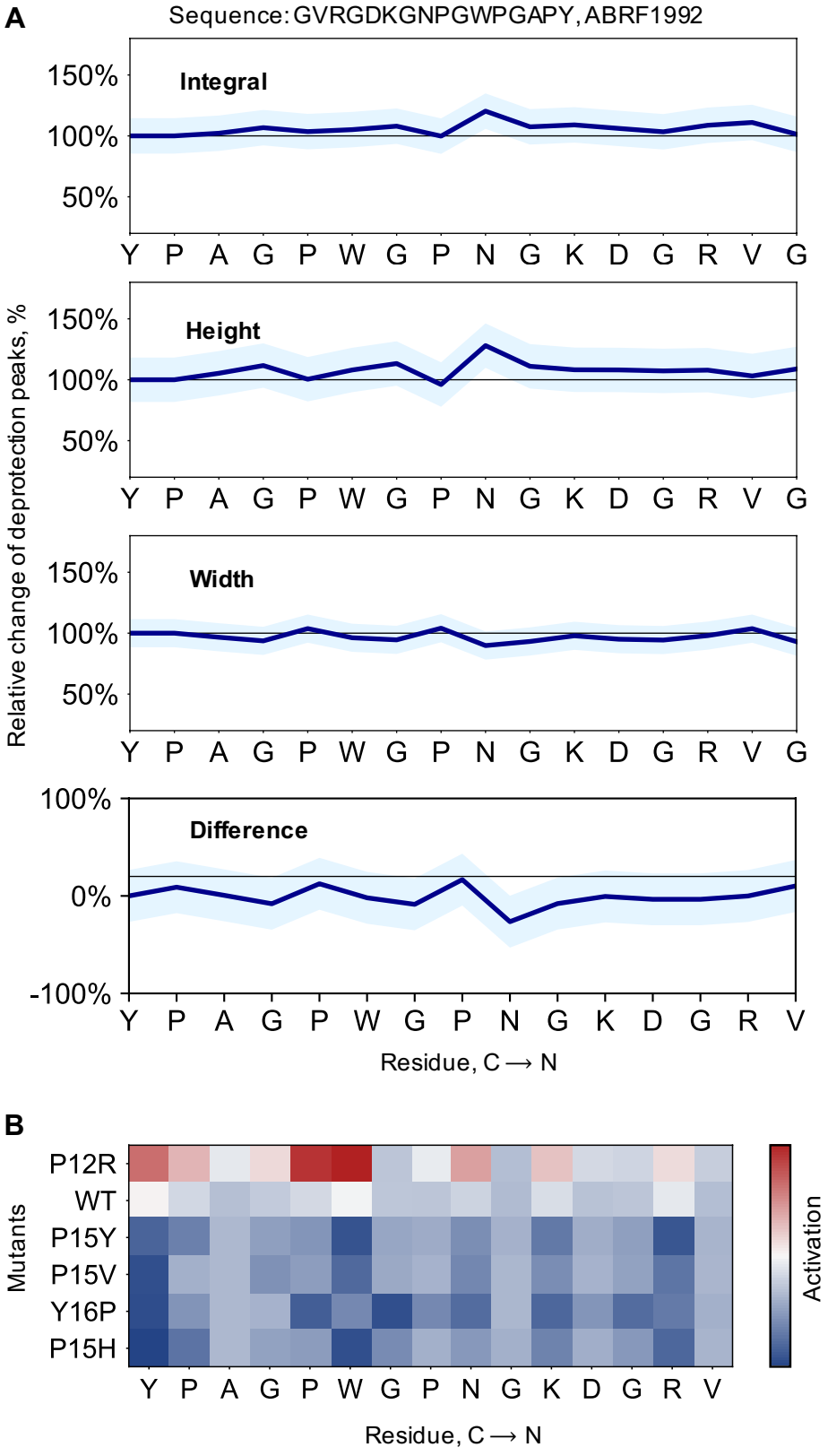
3.4.3 1-42 β-Amyloid



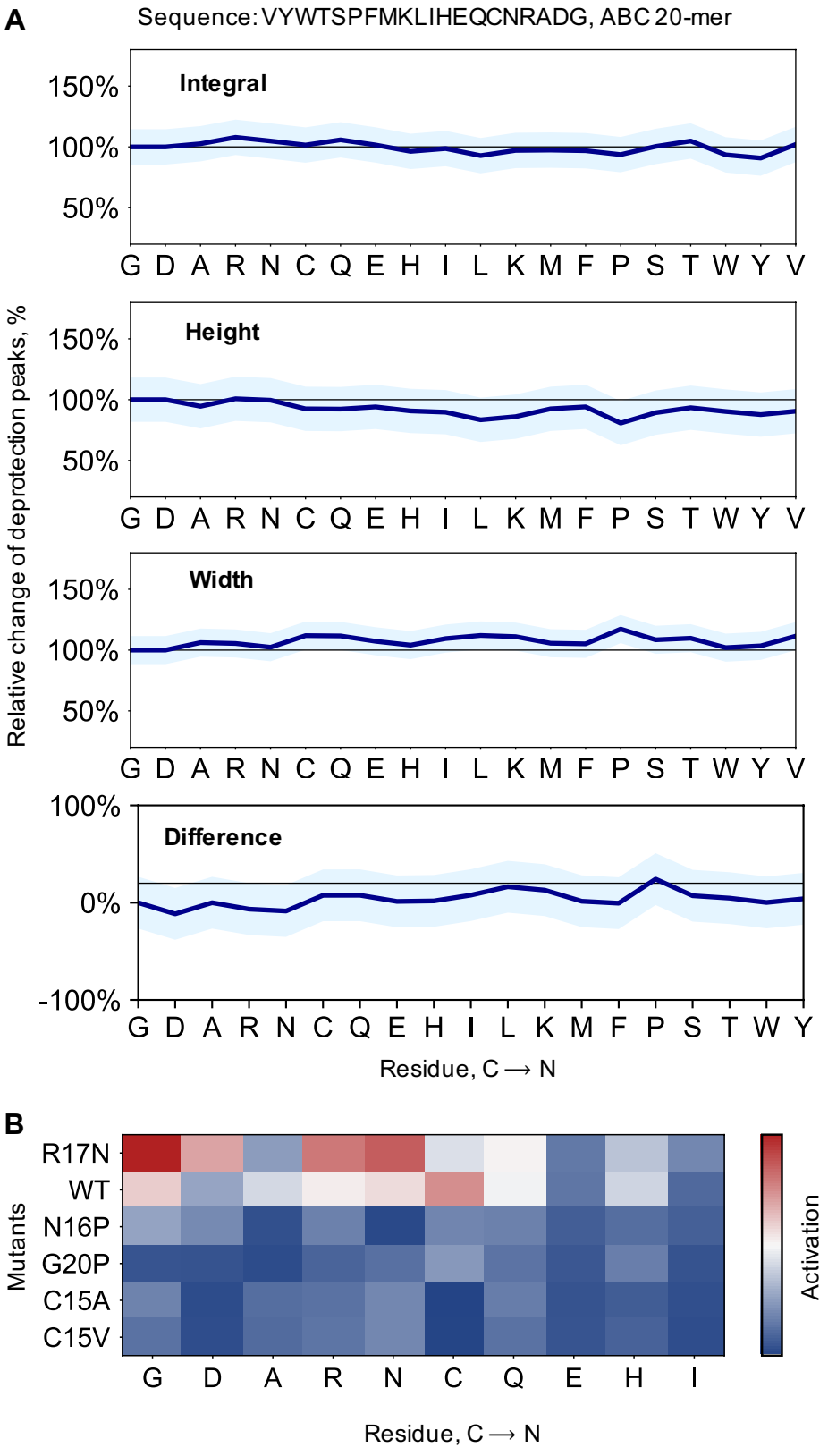
3.4.4 Thymosin



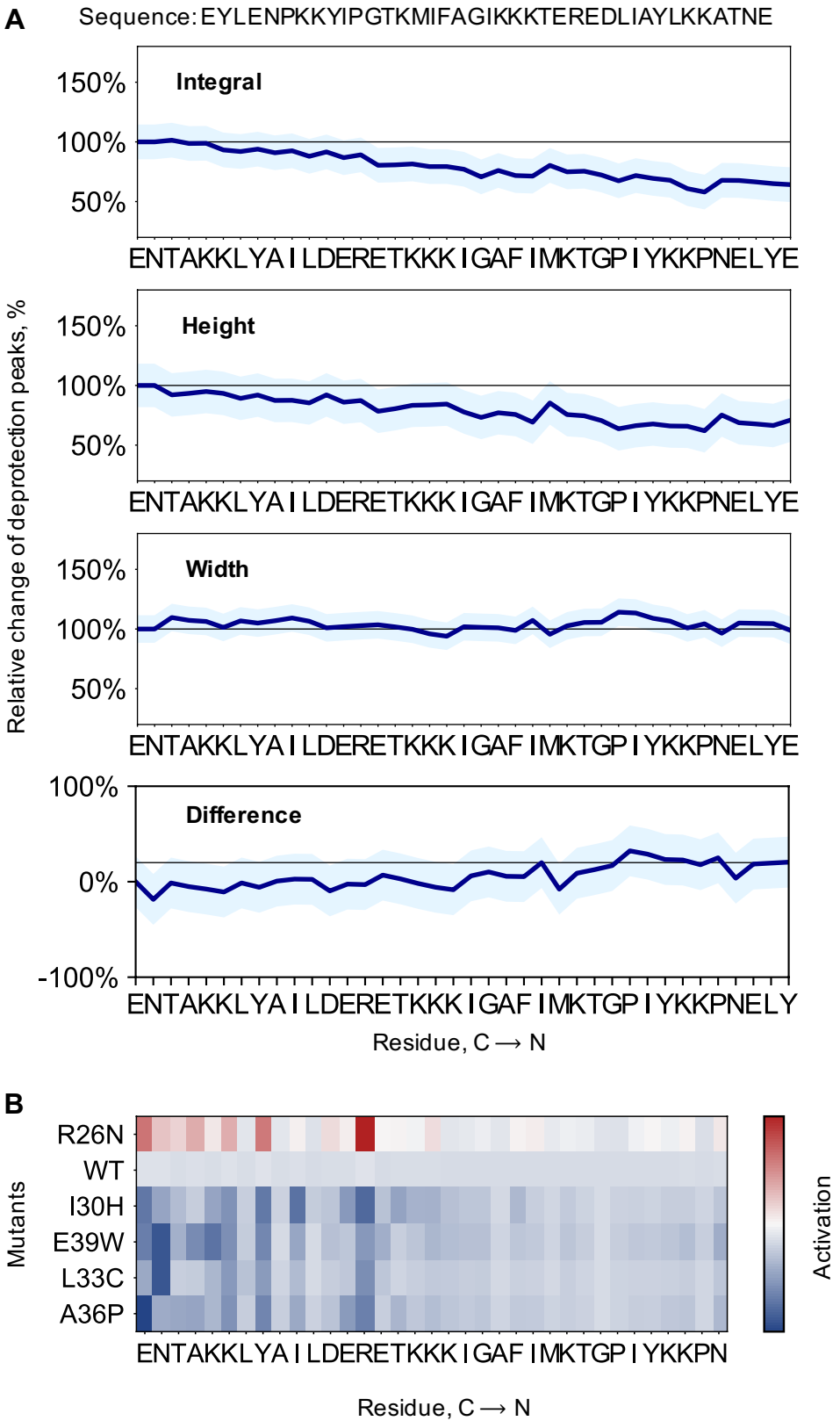
3.4.5 ABRF 1992



3.4.6 ABC 20-mer



3.4.7 Sequence: EYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE



4 Experimental validation of predicted sequences

4.1 Synthesis parameters

Synthesis parameters are described in detail in the literature (SI Table 2).[3]

SI Table 2. Set of optimized synthesis conditions on the AFPS. Pump strokes refer to volumes described in the general synthesis protocol.

Parameter	Conditions
Temperature	85–90 °C in reactor, 90 °C in 10' activation loop (for all other amino acids)**
Flow Rate	40 mL/min
Coupling step	0.40 M amino acids stocks in amine-free DMF 0.38 M activator stocks in amine-free DMF Coupling conditions: HATU (13 pump strokes) except S&A w/ HATU (26 pump strokes) and H, N, Q, V, R, T w/ PyAOP (26 pump strokes)
Deprotection step	40% pip in amine-free DMF with 2% formic acid (13 pump strokes)
Washing steps	Amine-free DMF (40 pump strokes)

**NOTE: during the process of condition optimization C and H coupled were changed to the following optimized protocol: activation with PyAOP (26 pump strokes) at 60 °C in 5' activation loop. However, for the peptides displayed in this manuscript the "old" protocol above was used.

4.2 Cleavage protocol

After synthesis, the peptidyl resin was washed with dichloromethane (3 x 5 mL), dried in a vacuum chamber, and weighed. 50% of the resin was transferred into a 50 mL conical polypropylene tube. For cleavage of peptides we used the following protocol[1]**:**

Approximately 3 mL of cleavage solution (94% TFA, 1% TIPS, 2.5% EDT, 2.5% water) was added to the tube. If needed, more cleavage solution was added to ensure complete submersion. The tube was kept at room temperature for 2 h. Ice cold diethyl ether (45 mL) was added to the cleavage mixture and the precipitate was collected by centrifugation and triturated twice more with cold diethyl ether (45 mL). The supernatant was discarded. Residual ether was allowed to evaporate and the peptide was dissolved in 50% acetonitrile in water with 0.1% TFA (long peptides were dissolved 70% acetonitrile in water with 0.1% TFA). The peptide solution was filtrated with a Nylon 0.22 μm syringe filter and frozen, lyophilized until dry, and weighed.

4.3 Liquid chromatography–mass spectrometry (LC-MS)

For mass analysis, the filtered peptide solution (10 μL of a 1mg/mL solution) was diluted in 50% acetonitrile in water with 0.1% TFA (90 μL) to a final concentration approximately 0.1 mg/mL. LC-MS chromatograms and associated high resolution mass spectra were acquired using an Agilent 6520 Accurate-Mass Q-TOF LC-MS (abbreviated as 6520) or an Agilent 6550 iFunnel Q-TOF LC-MS system (abbreviated as 6550). Solvent compositions used in the LC-MS are 0.1% formic acid in H_2O (solvent A) and 0.1% formic acid in acetonitrile (solvent B). The following LC-MS methods were used:

[1] **Note:** for short peptides, which were soluble in ether (e.g. JR-10), the trituration step was skipped and TFA was evaporated before addition of 50% acetonitrile in water with 0.1% TFA.

- ***1-61% B over 33 min, Phenomenex Jupiter C4 column (6550)***

  LC conditions: Phenomenex Jupiter C4 column: 1.0 × 150 mm, 5 µm, column temperature: 40 °C, gradient: 0-2 min 1% B, 2-30 min 1-91% B, 30-34 min 61-90% B; flow rate: 0.1 mL/min. A final 4-min hold was performed at a flow rate of 0.1 mL/min. The total method time was 38 min. MS is on from 4 to 30 min.

  MS conditions: positive electrospray ionization (ESI) extended dynamic mode in mass range 100–1700 *m*/*z*.

- ***1-91% B over 20 min, Phenomenex Jupiter C4 column (6550)***

  LC conditions: Phenomenex Jupiter C4 column: 1.0 × 150 mm, 5 µm, column temperature: 40 °C, gradient: 0-2 min 1% B, 2-18 min 1-91% B, 18-21 min 91% B; flow rate: 0.1 mL/min. A final 4-min hold was performed at a flow rate of 0.1 mL/min. The total method time was 25 min. MS is on from 4 to 18 min.

  MS conditions: positive electrospray ionization (ESI) extended dynamic mode in mass range 100–1700 *m*/*z*.

- ***1-61% B over 18 min, Luna C18 column (6550)***

  LC conditions: Phenomenex Luna C18 column: 0.5 × 150 mm, 5 µm, column temperature: 40 °C, gradient: 0-2 min 1% B, 2-14 min 1-61% B, 14-18 min 61-91% B; flow rate: 0.1 mL/min. A final 5-min hold was performed at a flow rate of 0.1 mL/min. The total method time was 23 min. MS is on from 4 to 14 min.

  MS conditions: positive electrospray ionization (ESI) extended dynamic mode in mass range 100–1700 *m*/*z*.

- ***1-91% B over 30 min, Luna C18 column (6550)***

  LC conditions: Phenomenex Luna C18 column: 0.5 × 150 mm, 5 µm, column temperature: 40 °C, gradient: 0-2 min 1% B, 2-30 min 1-91% B, 30-34 min 61-90% B; flow rate: 0.1 mL/min. A final 4-min hold was performed at a flow rate of 0.1 mL/min. The total method time was 38 min. MS is on from 4 to 30 min.

  MS conditions: positive electrospray ionization (ESI) extended dynamic mode in mass range 100–1700 *m*/*z*.

Data were processed using Agilent MassHunter Workstation Qualitative Analysis Version B.06.00 with BioConfirm Software.

4.4 Analytical high-performance liquid chromatography (HPLC)

For determination of purity by HPLC, the filtered peptide solution was diluted in 50% acetonitrile in water with 0.1% TFA (100 µL) to a final concentration of approximately 1.0 mg/mL. Peptide samples containing cysteines were diluted in 6M Guanidinium chloride containing 100 mM DTT. The samples were analyzed on Agilent Technologies 1200 Series, which was computer-controlled through Agilent ChemStation software.

For standard analysis of all peptide samples, analytical HPLC spectra were recorded on an analytical Agilent Zorbax 300SB-C3 column (2.1 mm × 150 mm, 5-µm particle size). A linear gradient of acetonitrile with a 0.08% TFA additive (solvent B) in water with a 0.1% TFA additive (solvent A) was used. After a 3-min hold, gradients of 1% B per minute ramped up over 60 min at a flow rate of 0.4 mL/min. Gradients started at 5% B (annotated as "5–65% B over 60 min"). A final 3-min hold was performed. The total method time was 66 min. Crude HPLC purities were determined by manual integration of all signals in the area of 5–60 min.

4.5 Determination of yield

Molecular weight of peptide sequences was determined via ChemDraw, accounting for the weight of a TFA counter-ion for each basic residue (K, R, H) in addition to the N-terminal amine. For example, for a peptide with sequence "KALE" the molecular weight of the peptide as TFA salt is calculated as 916 g/mol (= 688 + 2 × 114).

The weight of lyophilized powders of the peptides was directly measured using analytical scales (XS205DU Analytical Balance, Mettler-Toledo) [note: use of deionizers such as SPI Westek Workstation Still Air Ionizer helps with measurements]. Following folding, protein concentration was measured based on the outlined procedures under "Determination of protein concentration".

Theoretical yield was determined based on weight of the resin, resin loading, and the molecular weight (with TFA) of each peptide.
For example, for the KALE sequence synthesized on 50 mg resin with 0.44 mmol/g loading, theoretical yield is:

$$\text{theoretical yield } = 0.44\frac{\text{mmol}}{\text{g}} \times 50\text{ mg} \times 916\frac{\text{g}}{\text{mol}} = 20\text{ mg}$$

Yield of crude peptide was determined based on the ratio of weight of lyophilized crude peptide (as TFA salt) to theoretical yield multiplied by the purity determined by UV absorption at 280 nm (analytical HPLC).
In the example above, if 10 mg of crude KALE peptide is produced and the purity by analytical HPLC is 50%, synthesis yield is:

$$\text{yield } = \frac{10\text{ mg}}{20\text{ mg}} \times 0.50 \times 100 = 25\%$$

4.6 Computational and analytical data

4.6.1 GLP-1 mutants

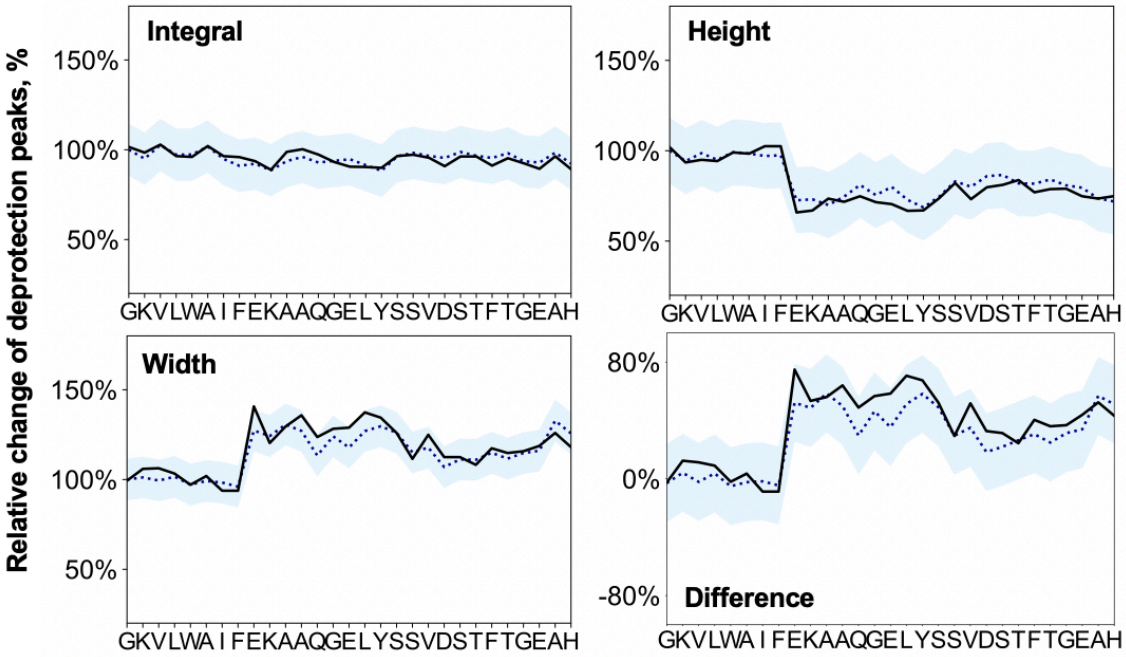
Synthesis Data for GLP-1 (R30S)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGS (30 AA)

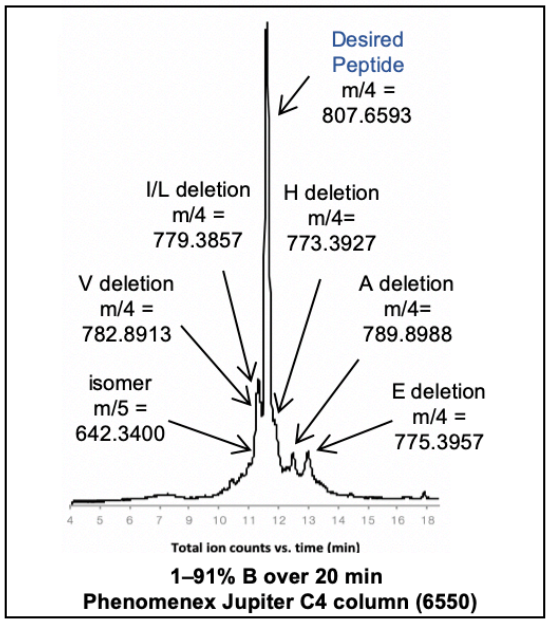
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

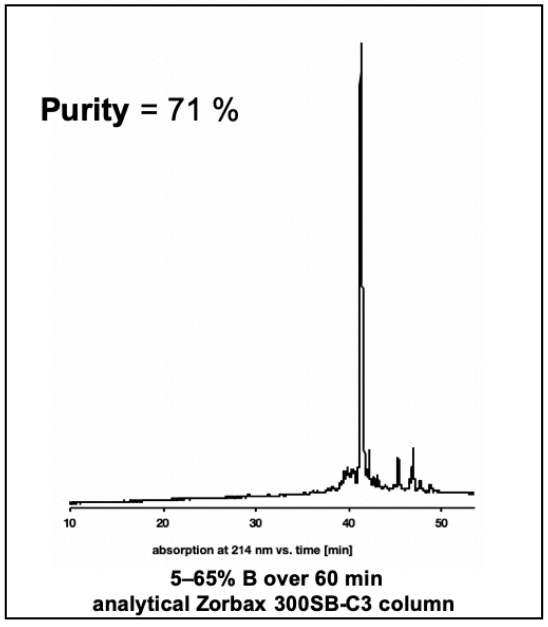
a) LCMS data, TIC



b) LCMS data, TIC



c) HPLC data



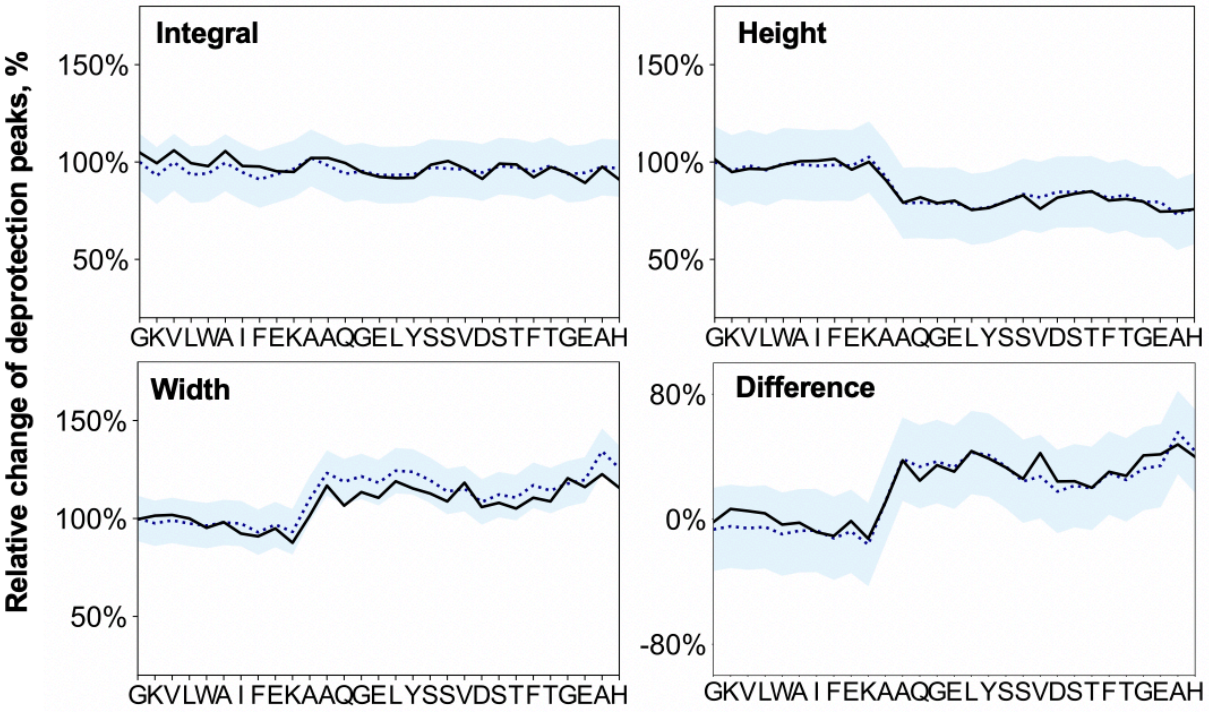
Synthesis Data for GLP-1 (WT)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR (30 AA)

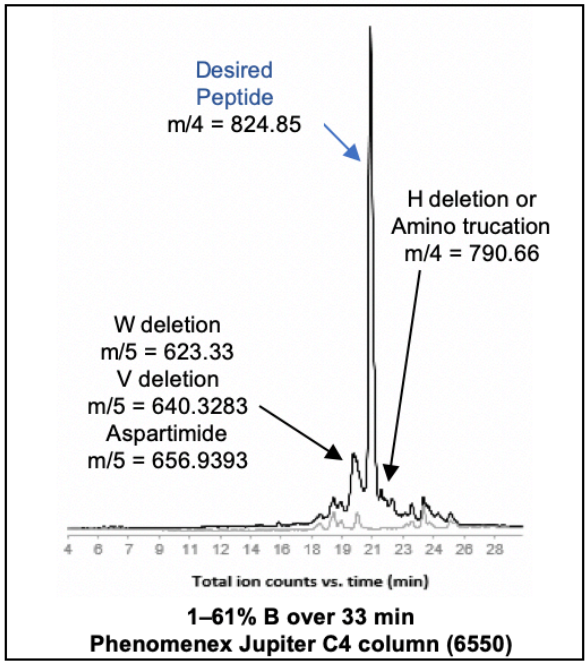
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

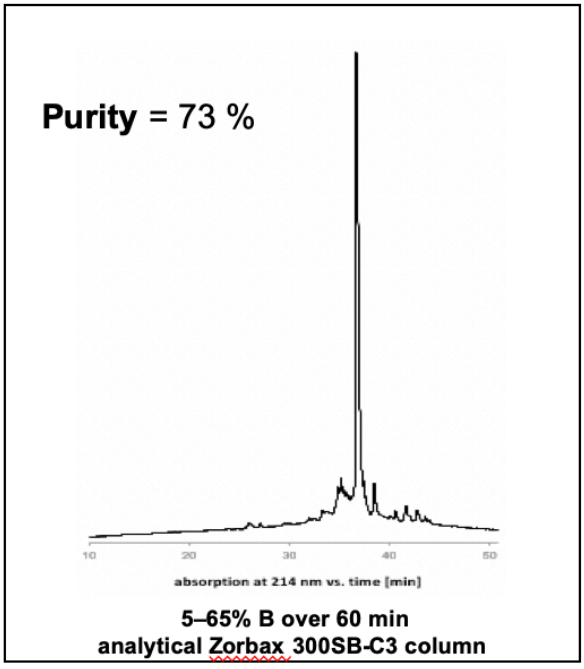


b) LCMS data, TIC



1–61% B over 33 min
Phenomenex Jupiter C4 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

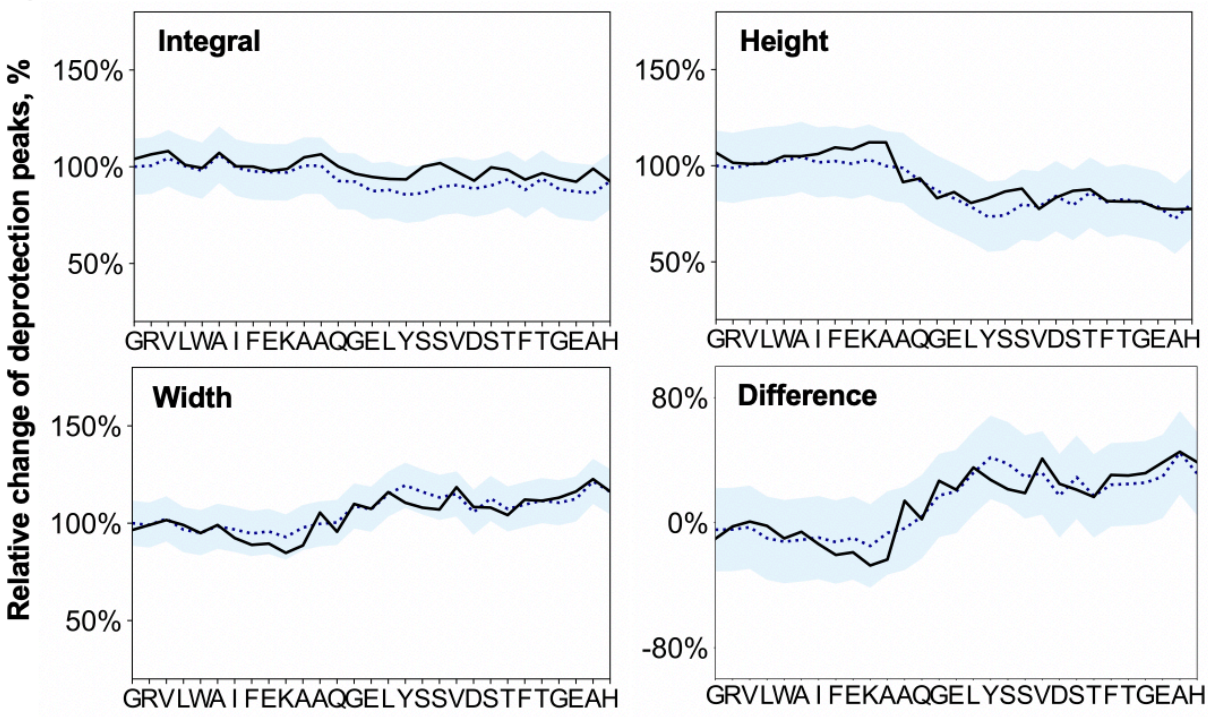
Synthesis Data for GLP-1 K28R

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVRGR (30 AA)

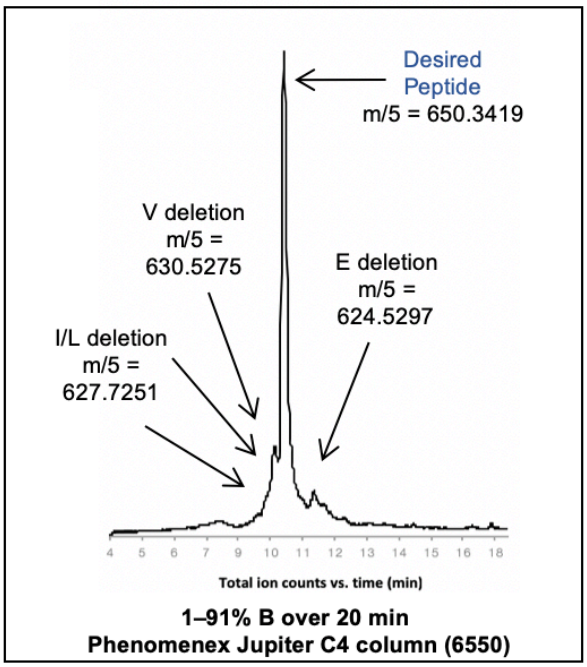
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

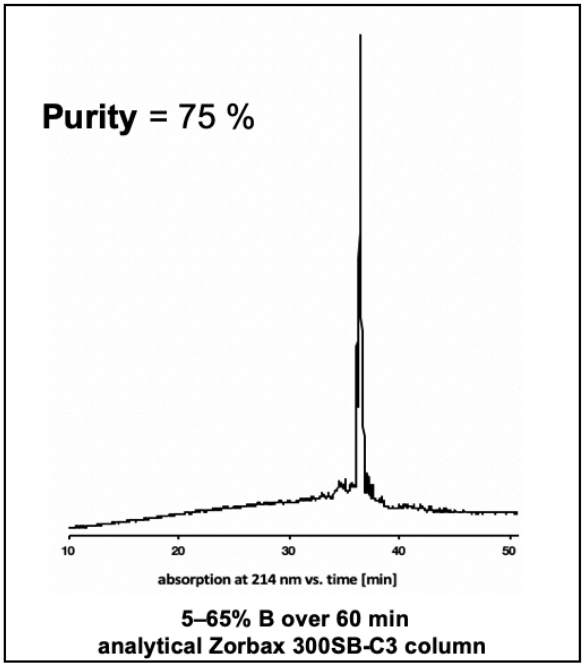


b) LCMS data, TIC



1–91% B over 20 min
Phenomenex Jupiter C4 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

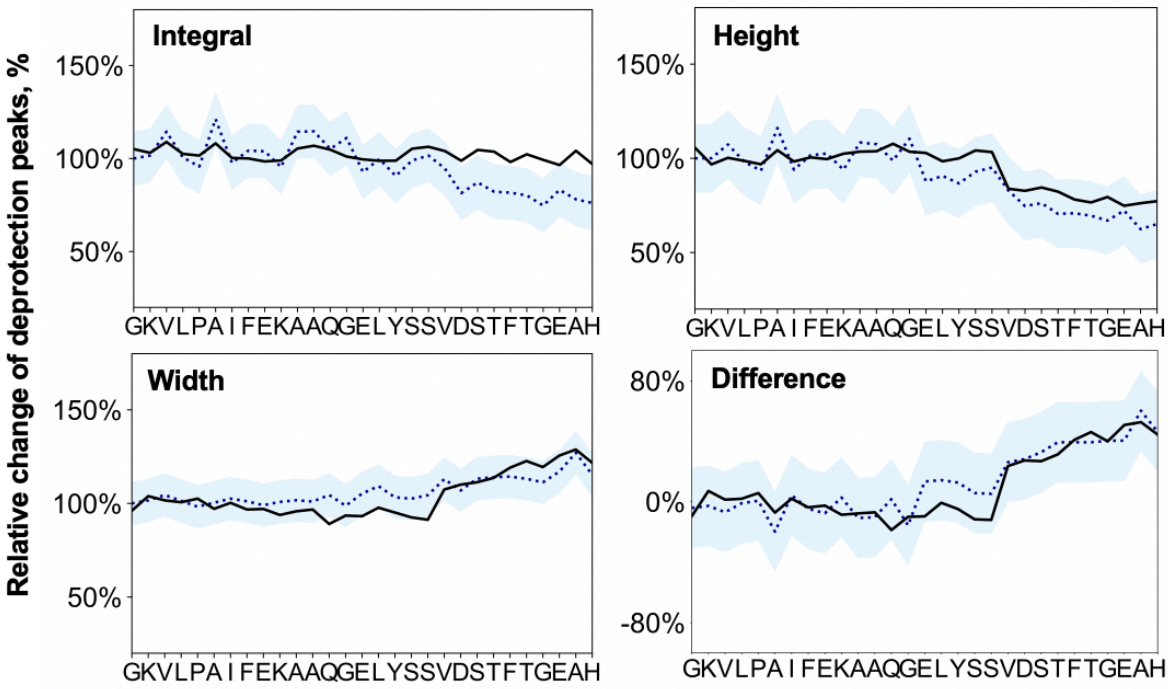
Synthesis Data for GLP-1 (W25P)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAPLVKGR (30 AA)

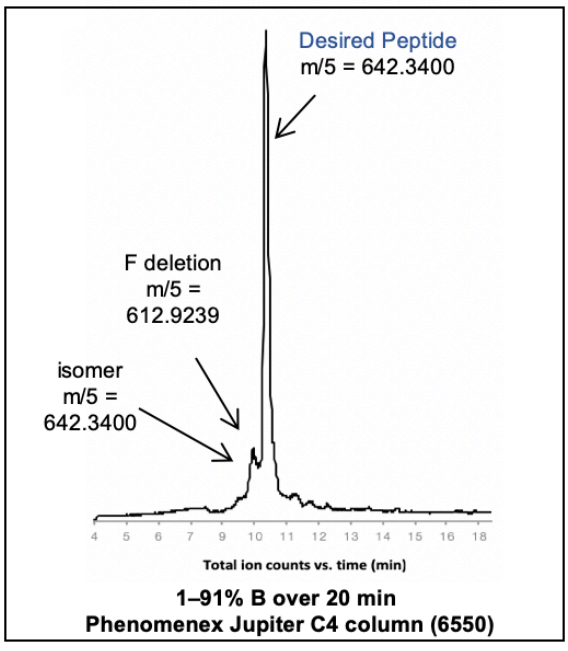
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

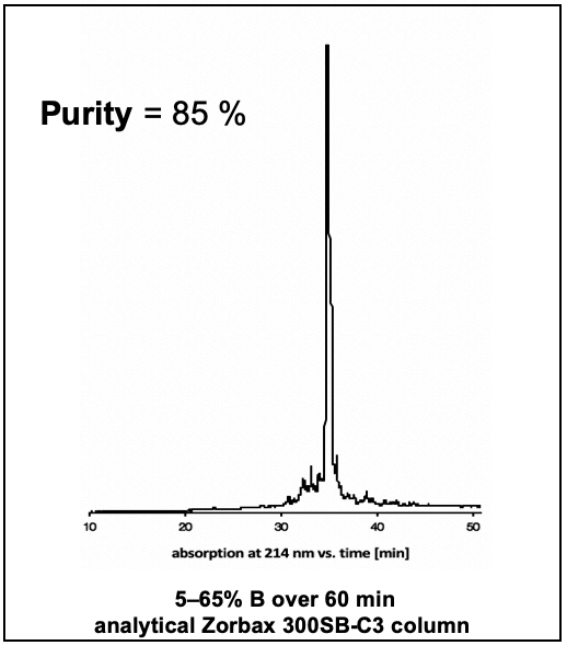


b) LCMS data, TIC



1–91% B over 20 min
Phenomenex Jupiter C4 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

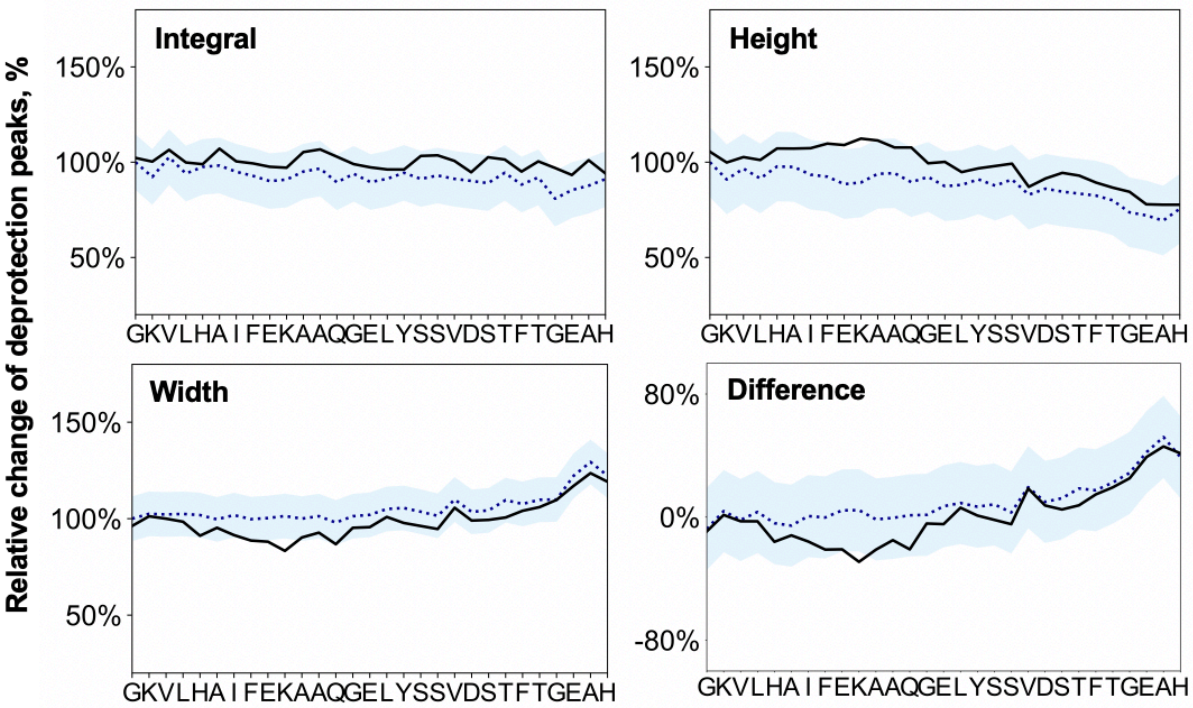
Synthesis Data for GLP-1 (W25H)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAHLVKGR (30 AA)

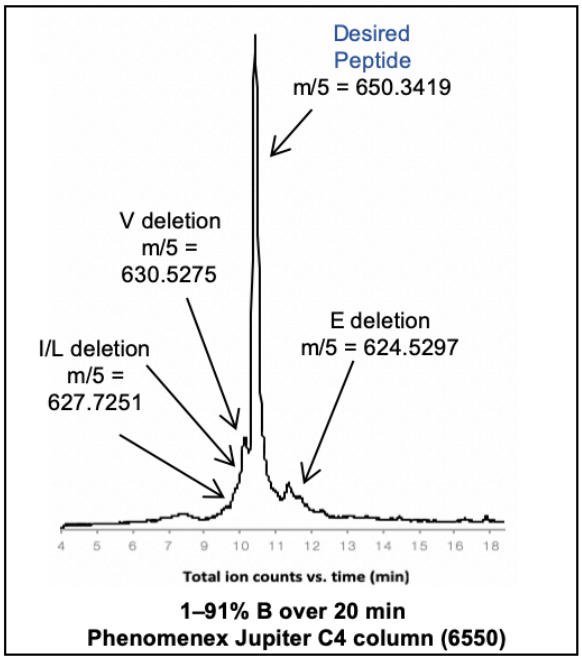
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

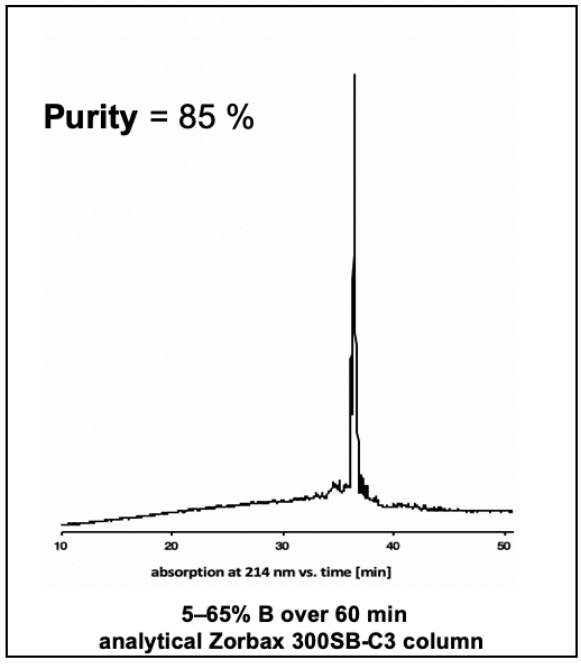


b) LCMS data, TIC



1–91% B over 20 min
Phenomenex Jupiter C4 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

4.6.2 JR-10 mutants

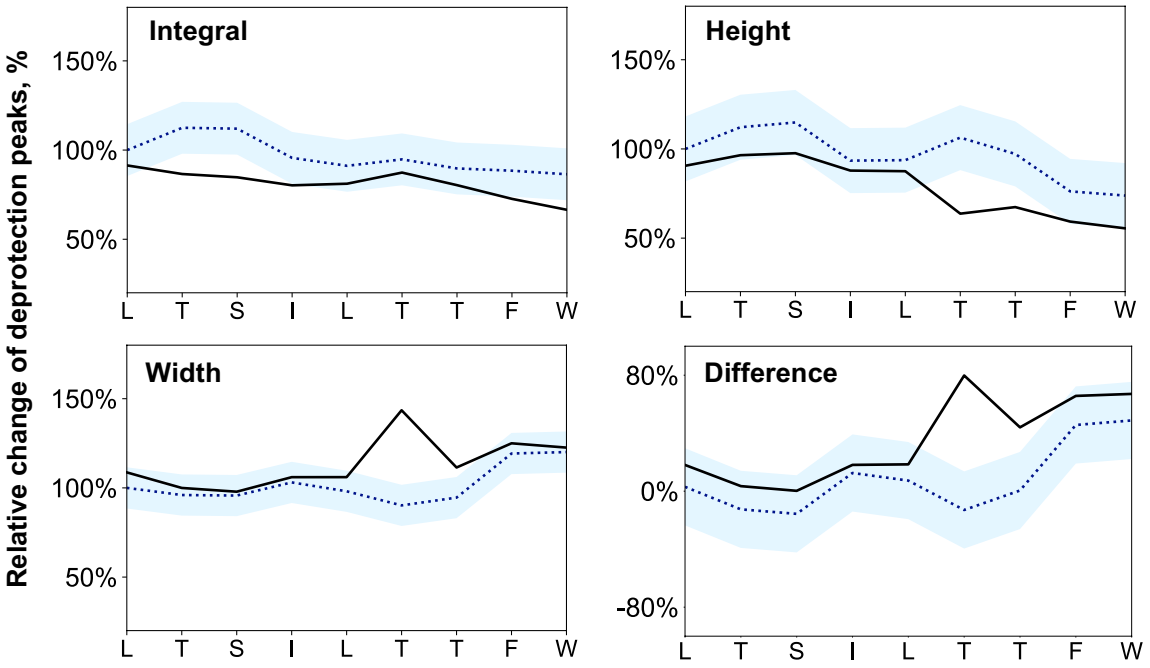
Synthesis Data for JR-10 (I9L)

Sequence: WFFTL ISTLM (10 AA)

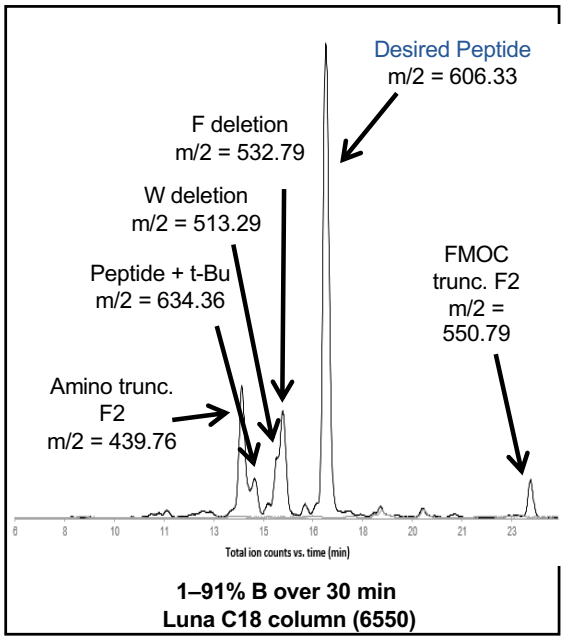
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

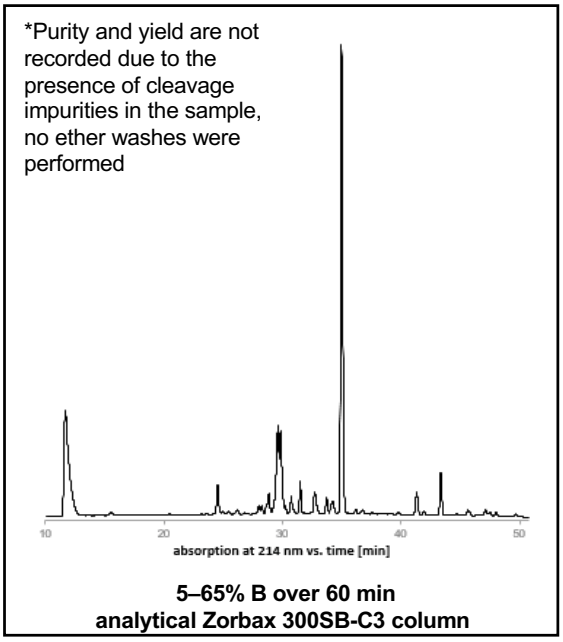


b) LCMS data, TIC



1–91% B over 30 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

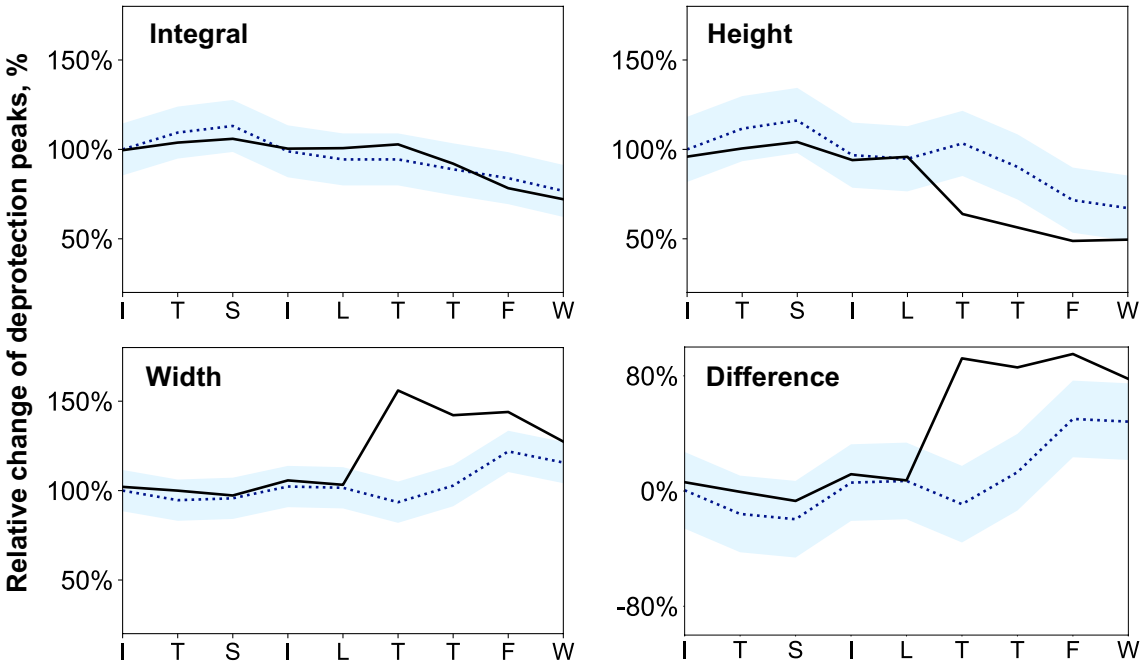
Synthesis Data for JR-10 (WT)

Sequence: WFFTL ISTIM (10 AA)

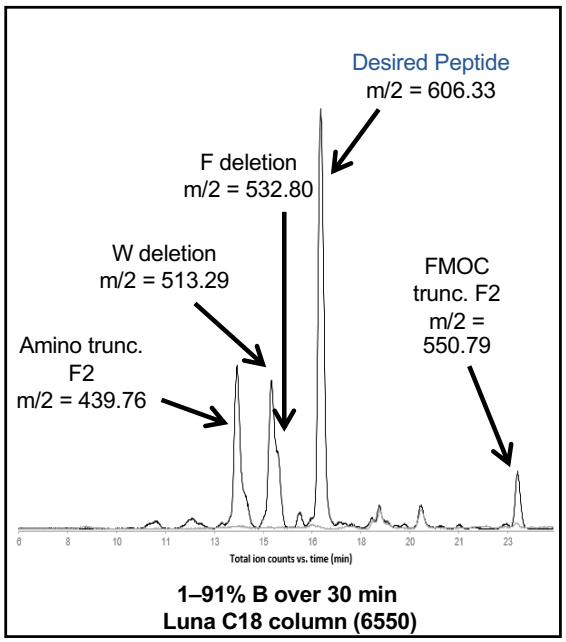
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

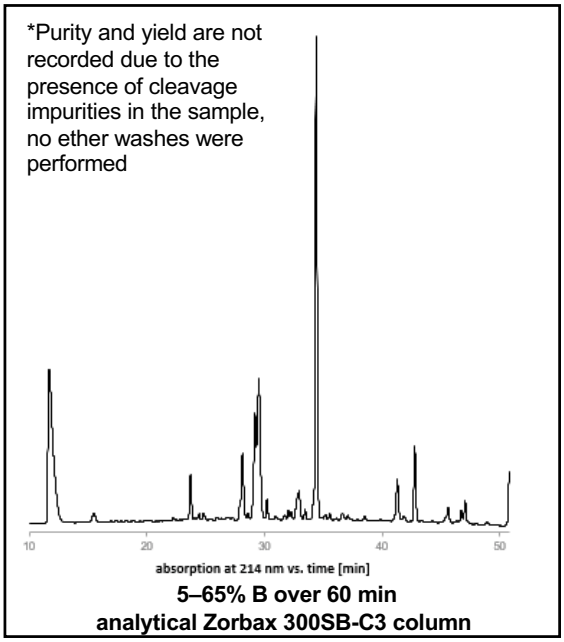


b) LCMS data, TIC



1–91% B over 30 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

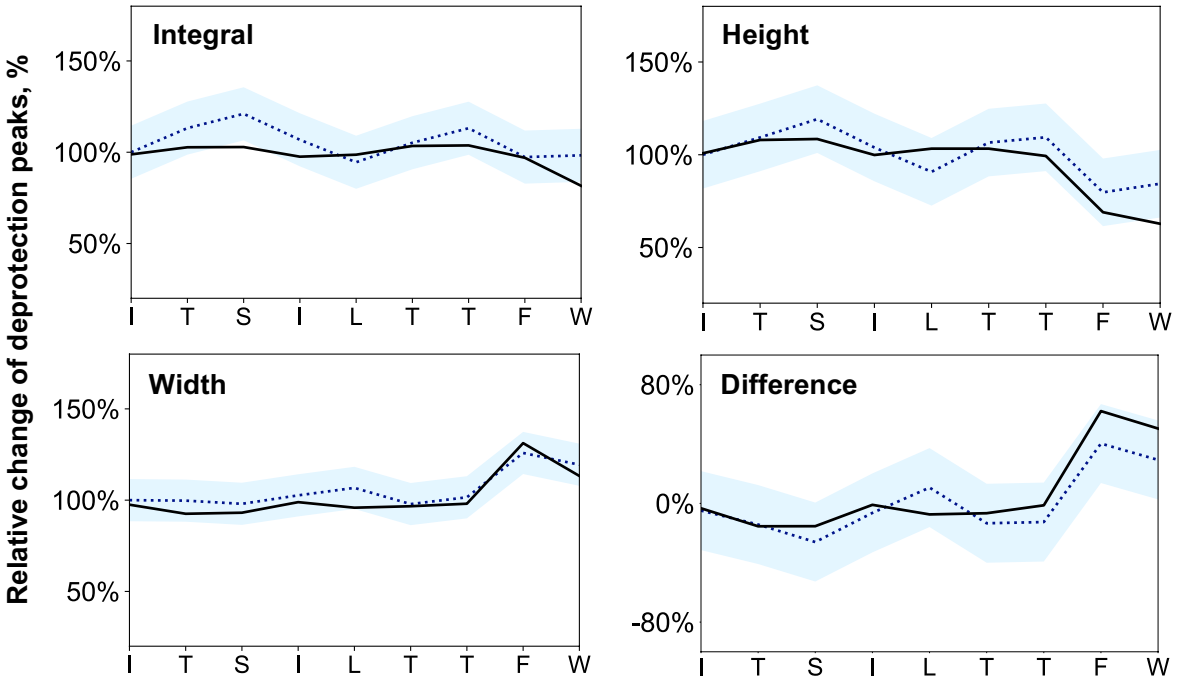
Synthesis Data for JR-10 (M10K)

Sequence: WFFTL ISTIK (10 AA)

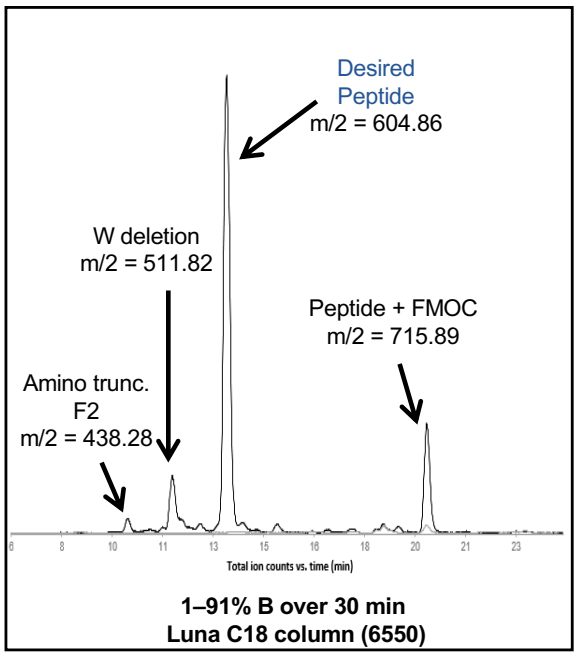
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

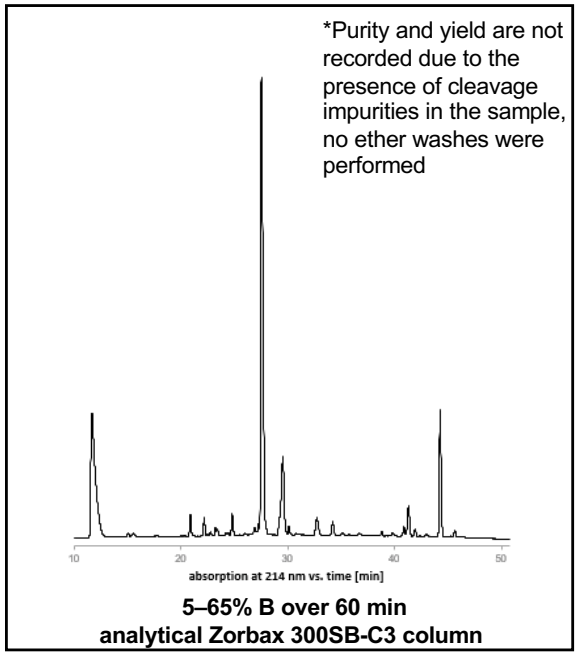


b) LCMS data, TIC



1–91% B over 30 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

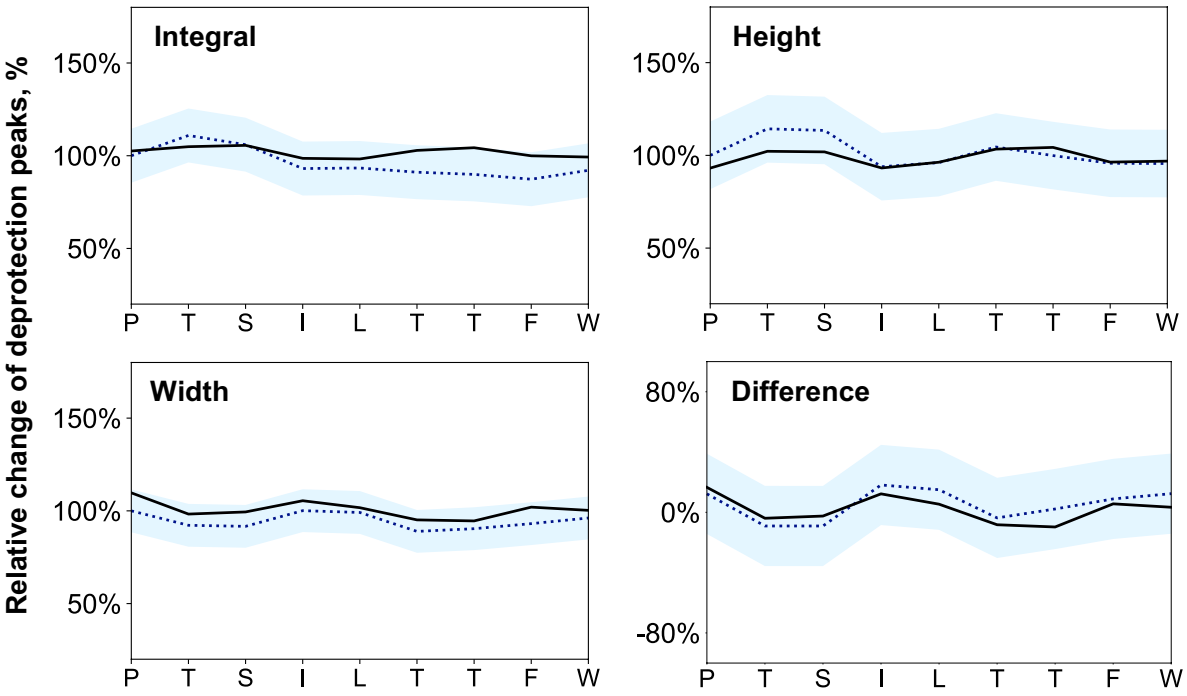
Synthesis Data for JR-10 (I9P)

Sequence: WFFTL ISTPM (10 AA)

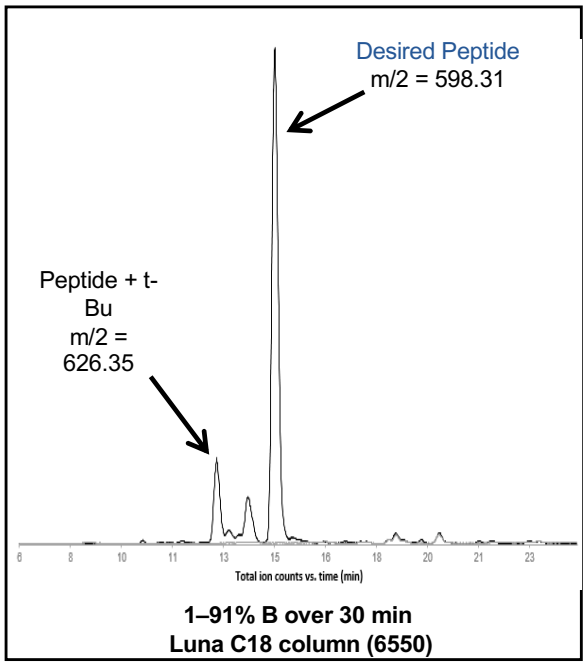
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

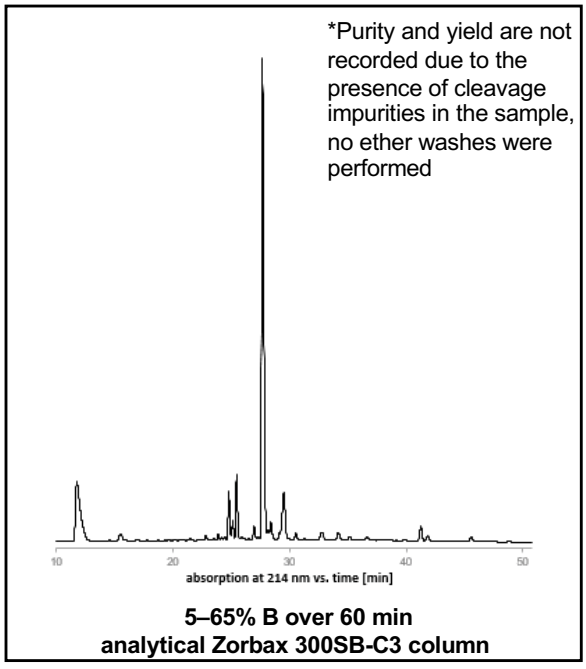


b) LCMS data, TIC



1–91% B over 30 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

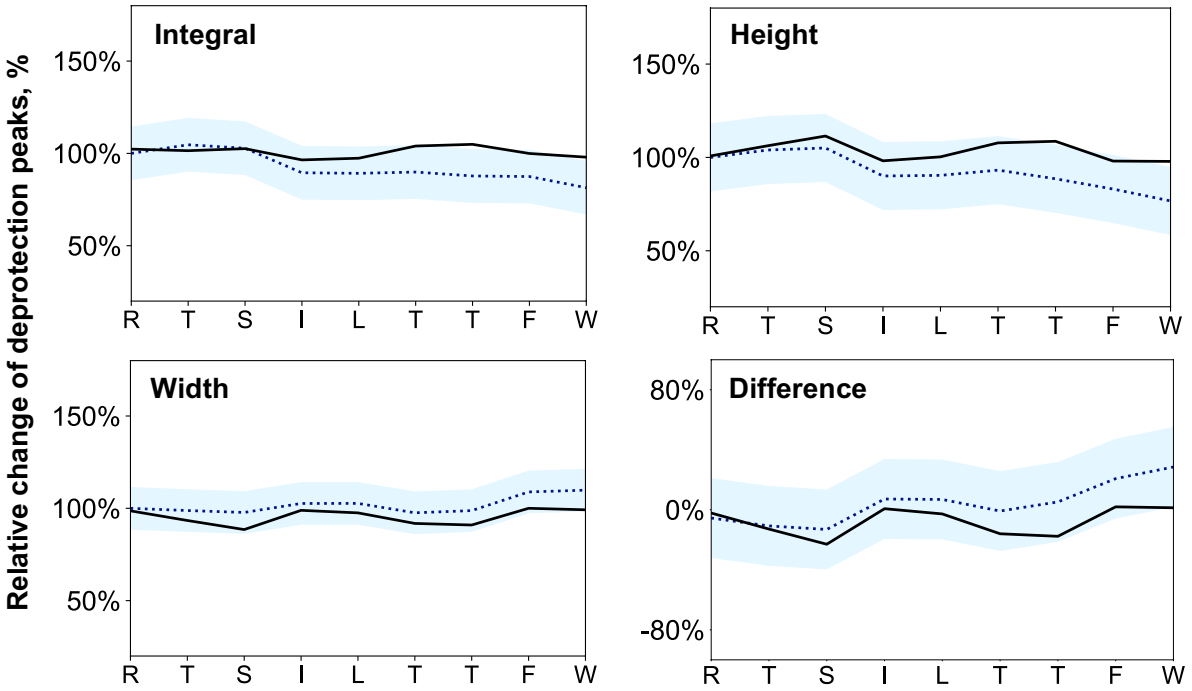
Synthesis Data for JR-10 (I9R)

Sequence: WFFTL ISTRM (10 AA)

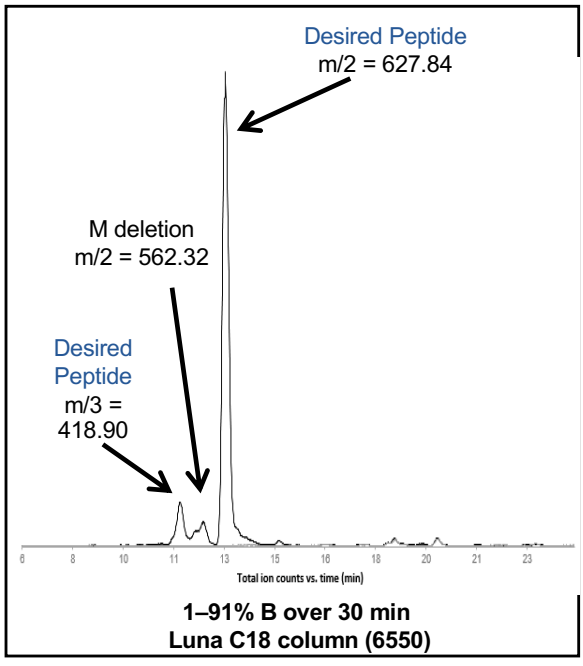
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

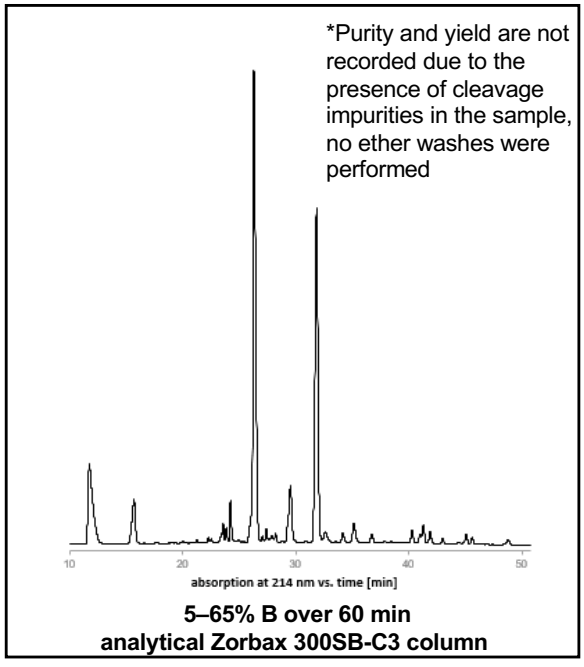


b) LCMS data, TIC



1–91% B over 30 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

4.6.3 Additional sequences

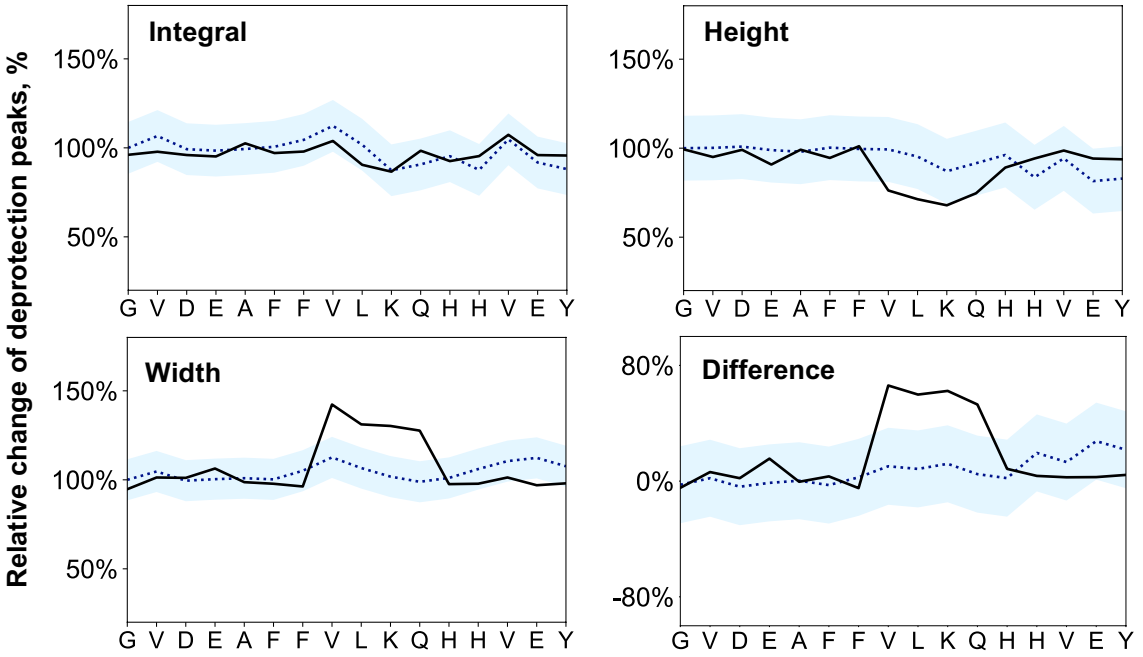
Synthesis Data for Aß[10-26]

Sequence: YEVHHQKLVFF AEDVGS (16 AA)

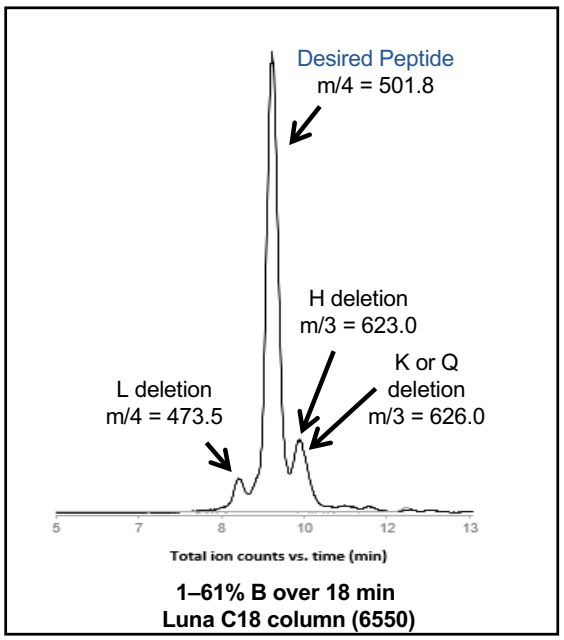
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 38 min

a) LCMS data, TIC

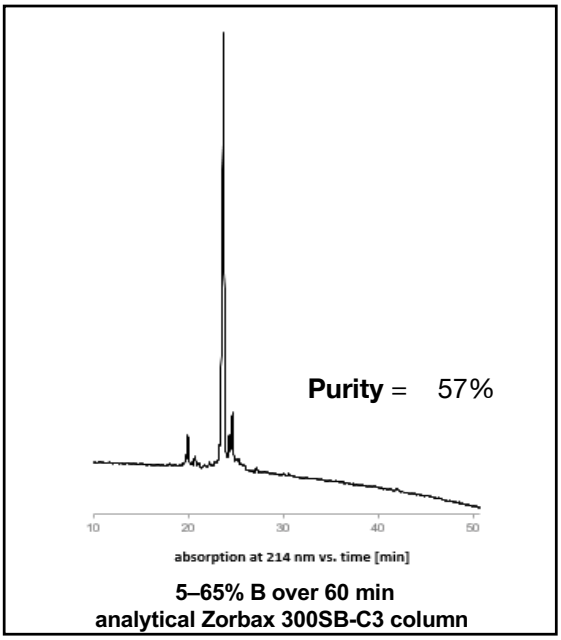


b) LCMS data, TIC



1–61% B over 18 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

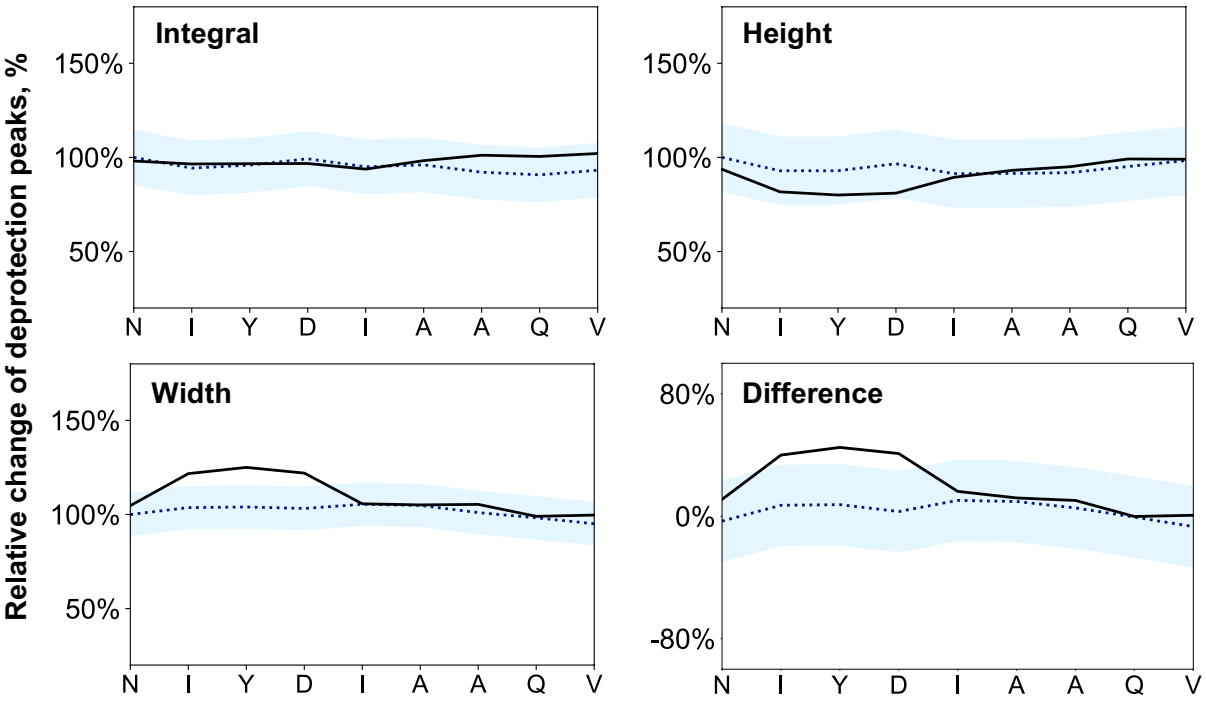
Synthesis Data for ACP[65-74]

Sequence: VQAAIDYING (10 AA)

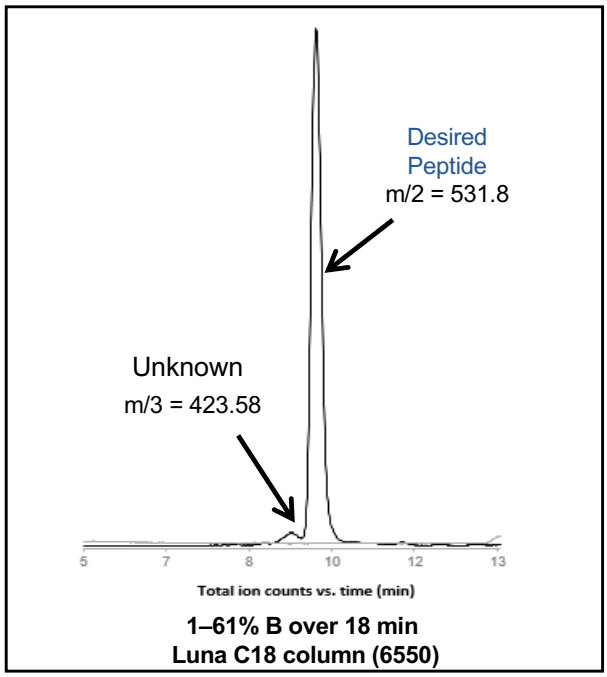
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

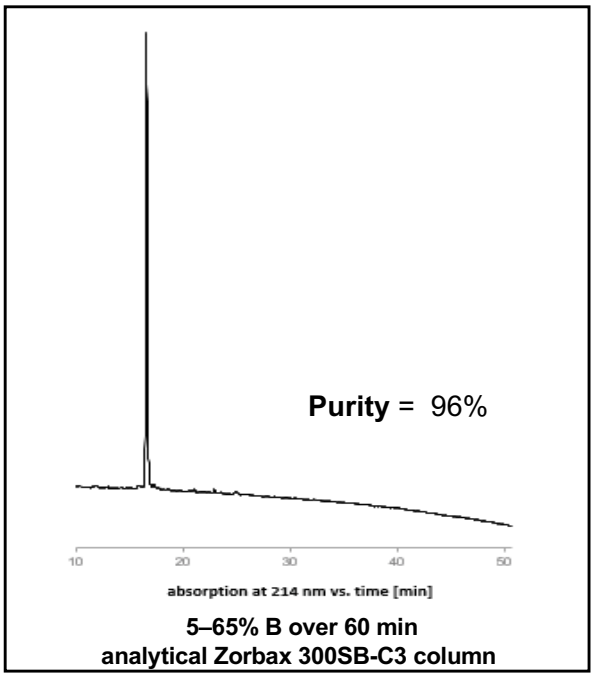


b) LCMS data, TIC



1–61% B over 18 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

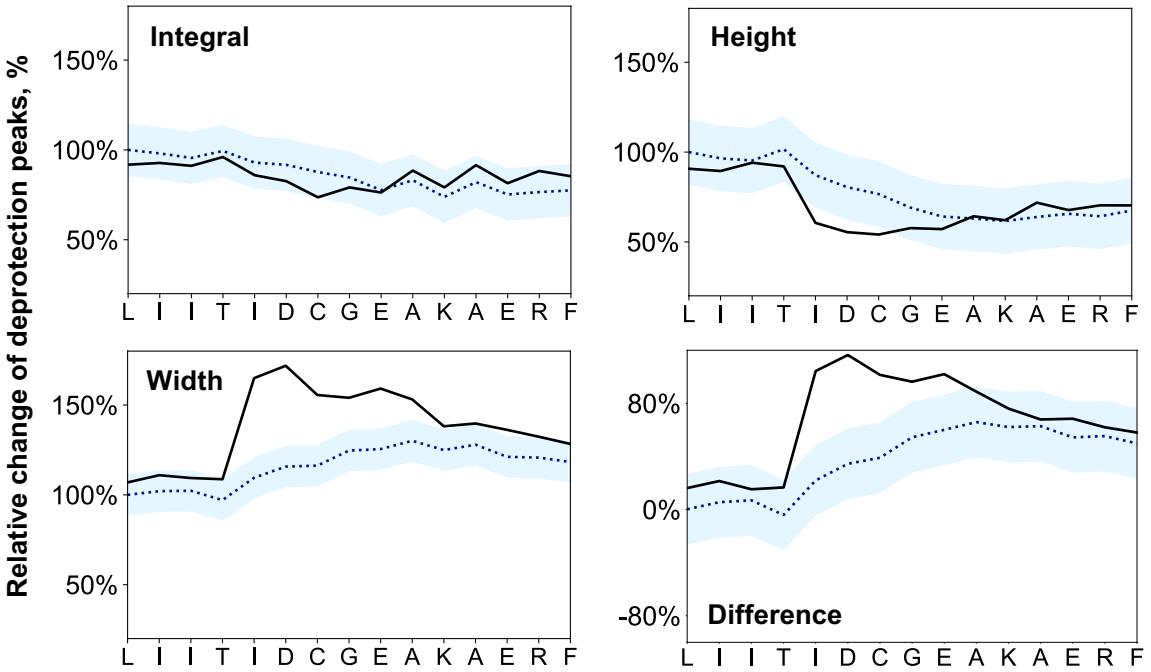
Synthesis Data for barstar[75-90]

Sequence: FREAKAEGCD ITIILS (16 AA)

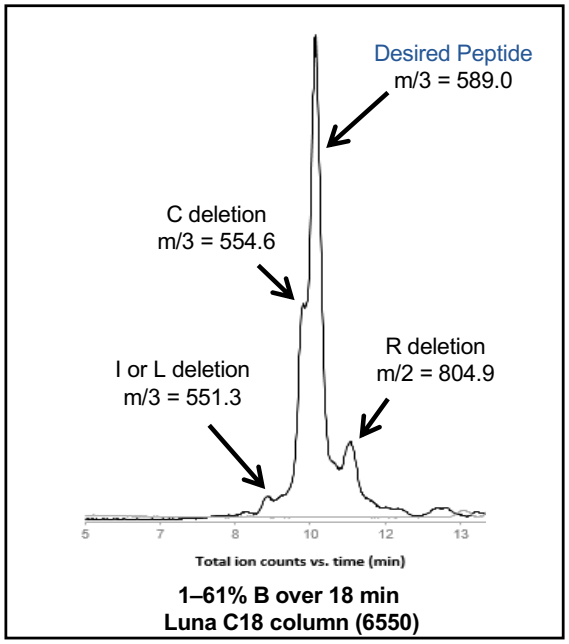
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 38 min

a) LCMS data, TIC

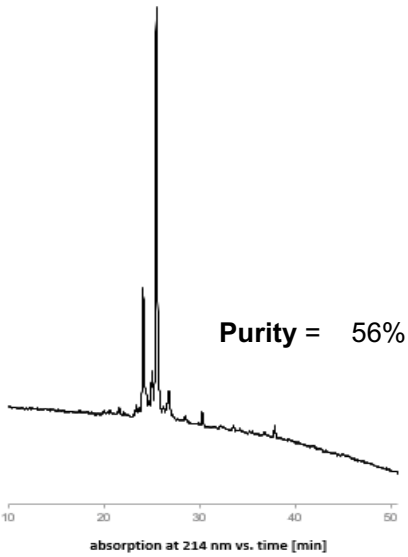


b) LCMS data, TIC



1–61% B over 18 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

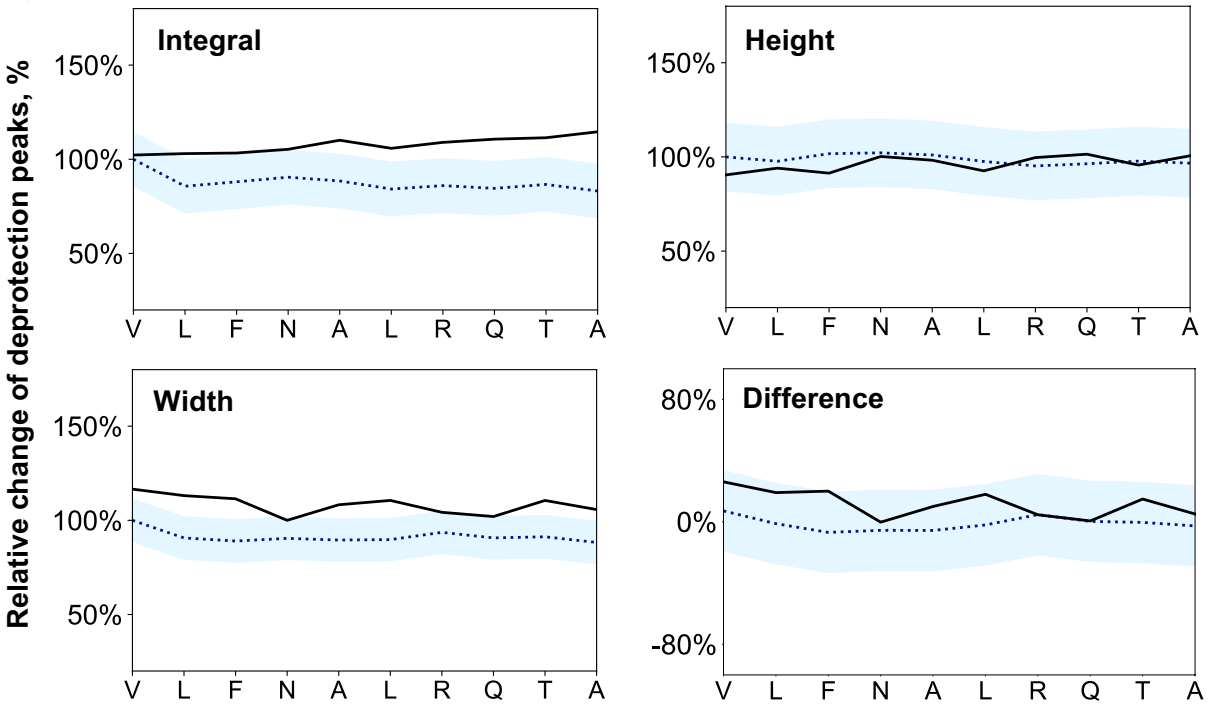
Synthesis Data for IAPP[1-18]

Sequence: ATQRLANFLV H (11 AA)

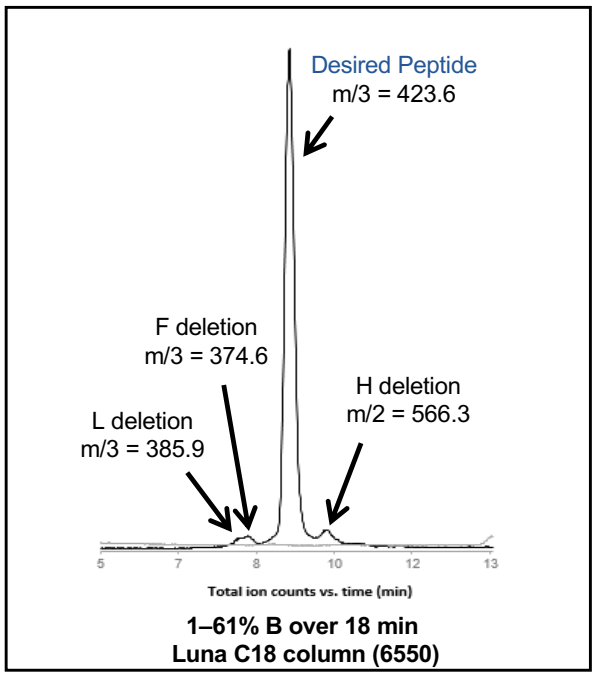
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 25 min

a) LCMS data, TIC

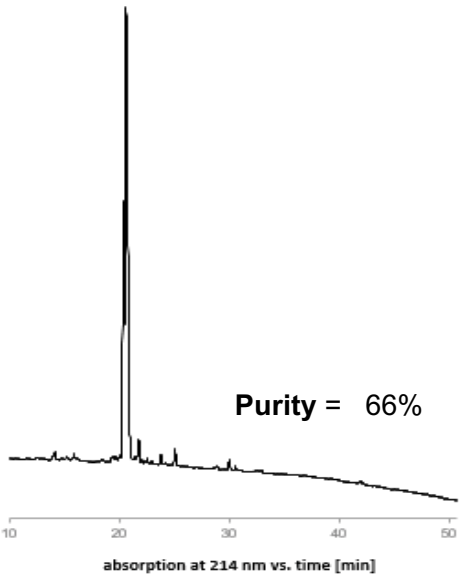


b) LCMS data, TIC



1–61% B over 18 min
Luna C18 column (6550)

c) HPLC data



5–65% B over 60 min
analytical Zorbax 300SB-C3 column

4.6.4 Backbone-modified peptides

How does backbone protection affect GLP-1 synthesis? And can we predict the synthesis outcome for residues that are new to the model?

SI Table 3. Point mutations in GLP-1 with building blocks with are new to the model.

Amidator No.	Sequence	Add. Building Block (= i)	HPLC purity	Crude yield
1412	HAEGT FTSDV SSYLE GQAAK EFIAW LVKGR		77%	30%
1413	HAEGT FTSDV SSYLE GQAAK EFIAW LVKGR synthesized as: HAEGT FTSDV SSYLE iQAAK EFIAW LVKGR	DMB-Gly (synthesized)	60%	26%
1417	HAEGT FTSDV SSYLE GQAAK EFIAW LVKGR synthesized as: HAEGT FTSDV SSYLE GQAAK EFIAW LVKiR	DMB-Gly (synthesized)	74%	29%
1415	HAEGT FTSDV SSYLE GQAAK EFIAW LVKGR synthesized as: HAEGT FTSDV iYLE GQAAK EFIAW LVKGR	Fmoc-Ser(*t*-Bu)-Ser($\Psi^{Me,Me}$pro)-OH	79%	36%
1416	HAEGT FTSDV SSYLE GQAAK EFIAW LVKGR synthesized as: HAEGT iSDV SSYLE GQAAK EFIAW LVKGR	Fmoc-Phe-Thr($\Psi^{Me,Me}$pro)-OH	77%	34%

Outcome: Fmoc-Ser(*t*-Bu)-Ser($\Psi^{Me,Me}$pro)-OH improves GLP-1 synthesis and minimizes H deletions. DMB-Gly seems to reduce deletions in the aggregation area (QAAK) – combination of both could lead to better results.

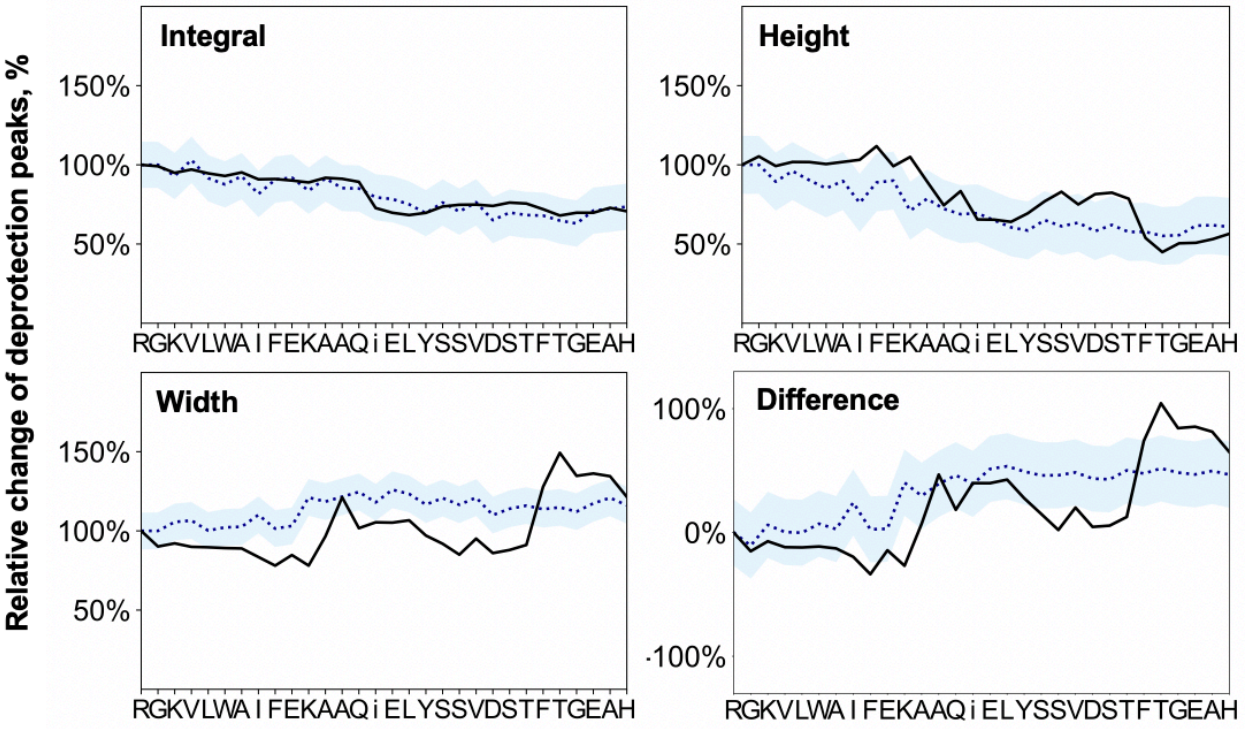
Synthesis Data for GLP-1 (Fmoc-(DMB)Gly-OH 1)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR (30 AA);
G = Fmoc-(DMB)Gly-OH

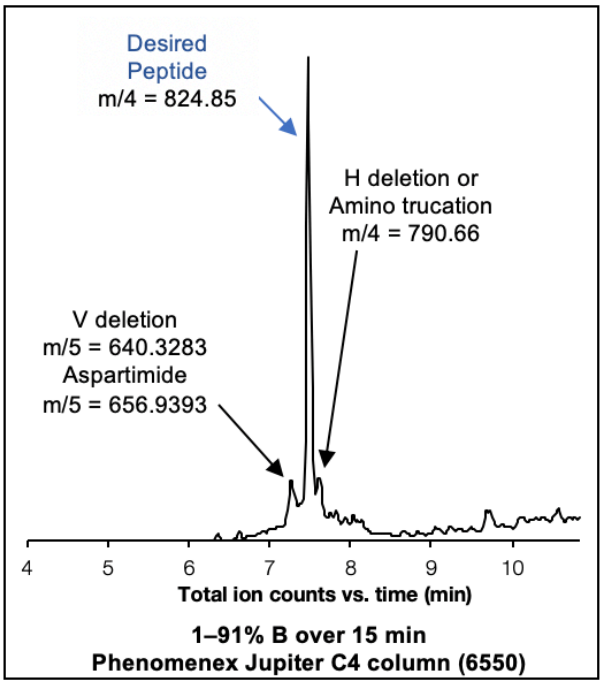
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

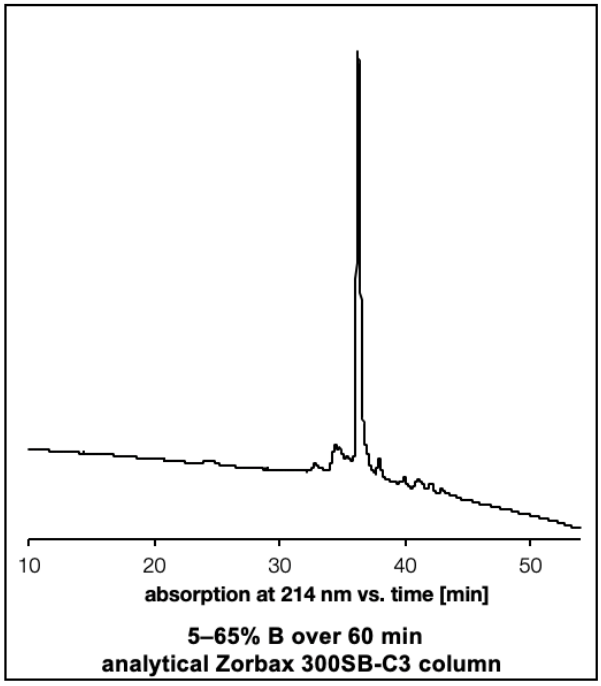
a) LCMS data, TIC



b) LCMS data, TIC



c) HPLC data



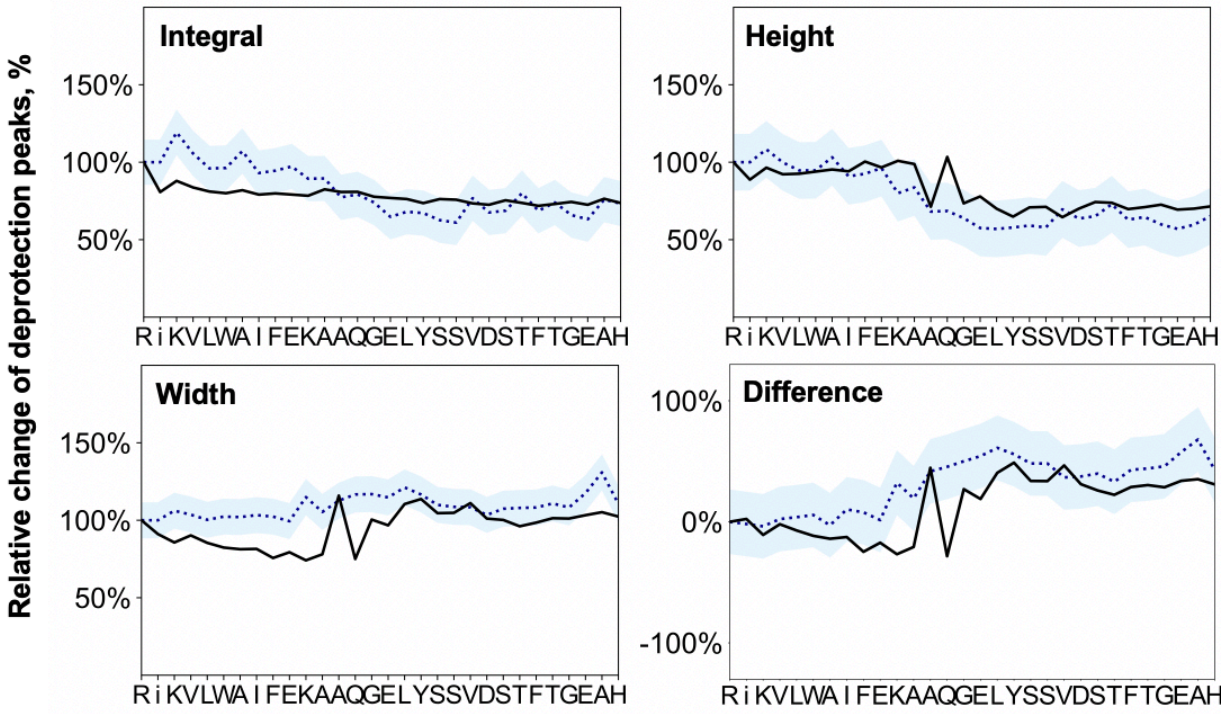
Synthesis Data for GLP-1 (Fmoc-(DMB)Gly-OH 2)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR (30 AA);
G = Fmoc-(DMB)Gly-OH

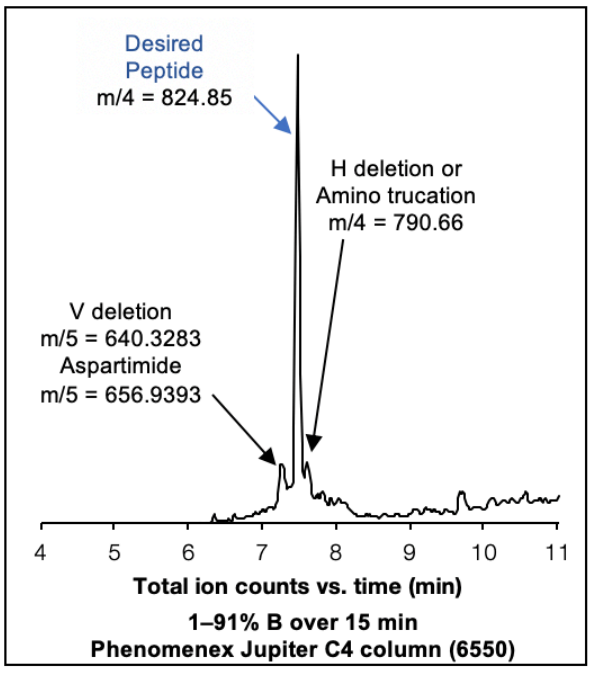
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

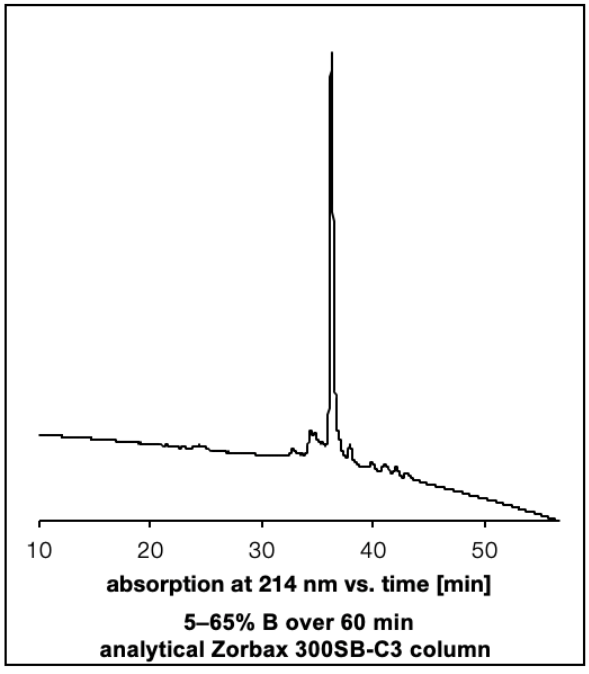
a) LCMS data, TIC



b) LCMS data, TIC



c) HPLC data



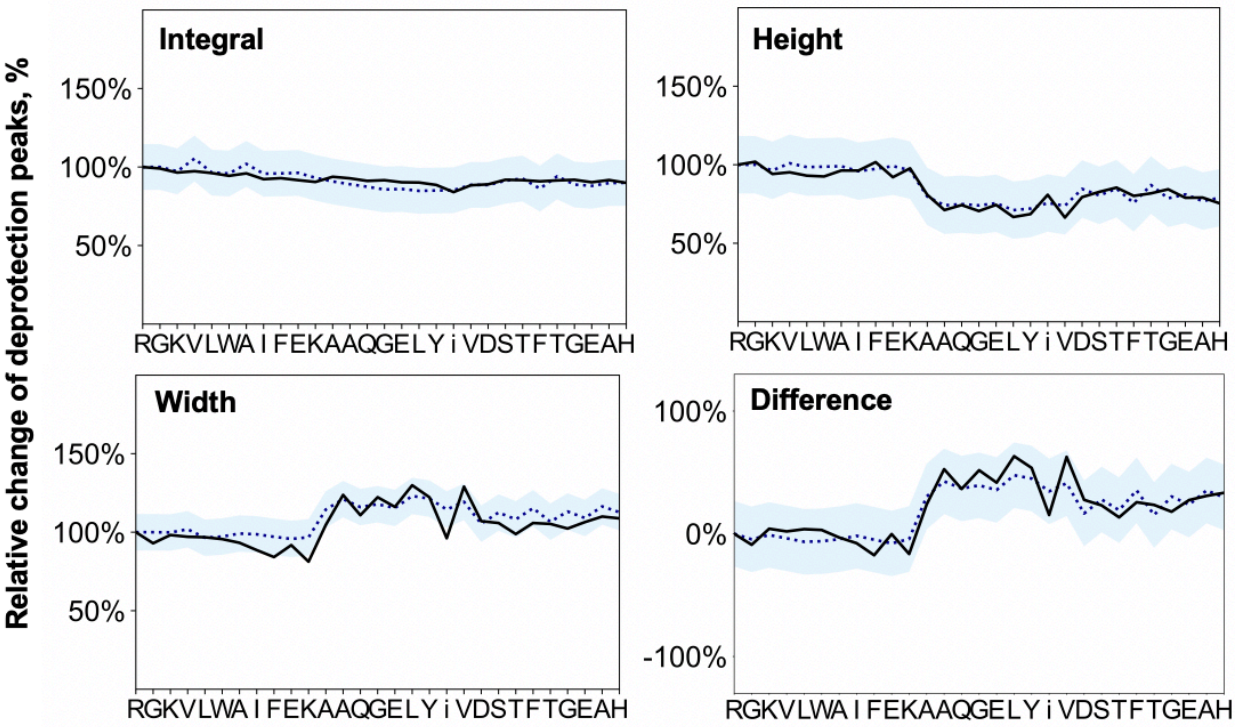
Synthesis Data for GLP-1 (Fmoc-Ser(*t*-Bu)-Ser($\Psi^{Me,Me}$pro)-OH)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR (30 AA);
SS = Fmoc-Ser(*t*-Bu)-Ser($\Psi^{Me,Me}$pro)-OH

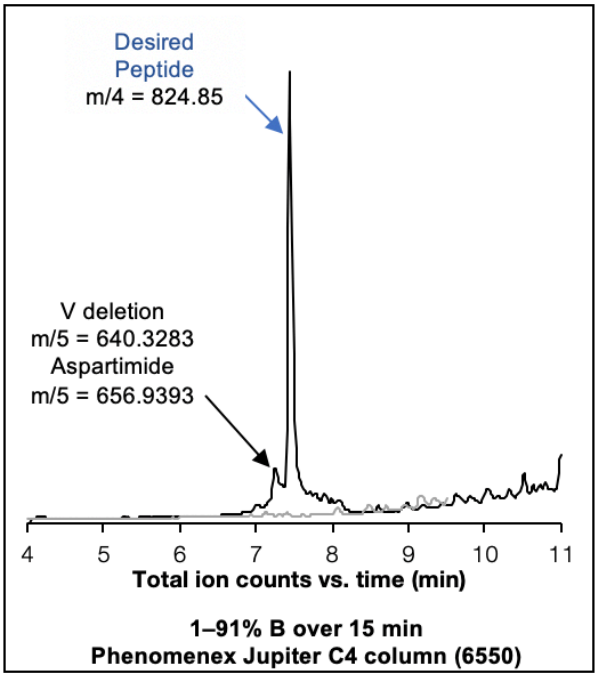
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

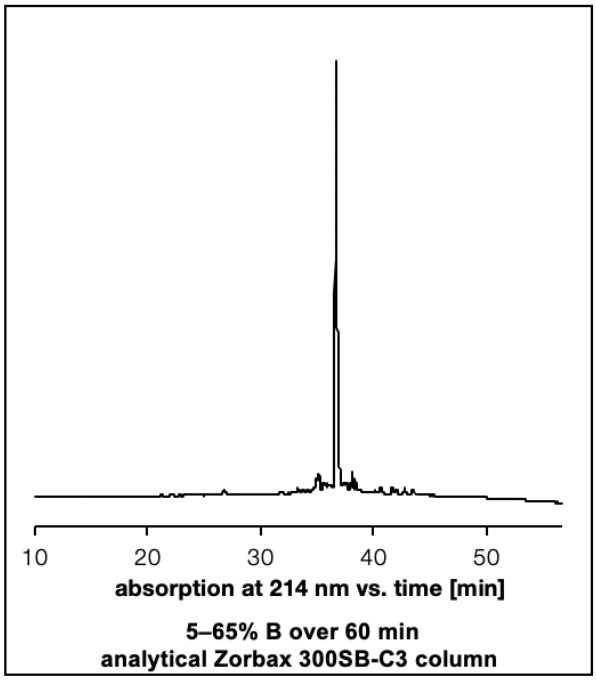
a) LCMS data, TIC



b) LCMS data, TIC



c) HPLC data



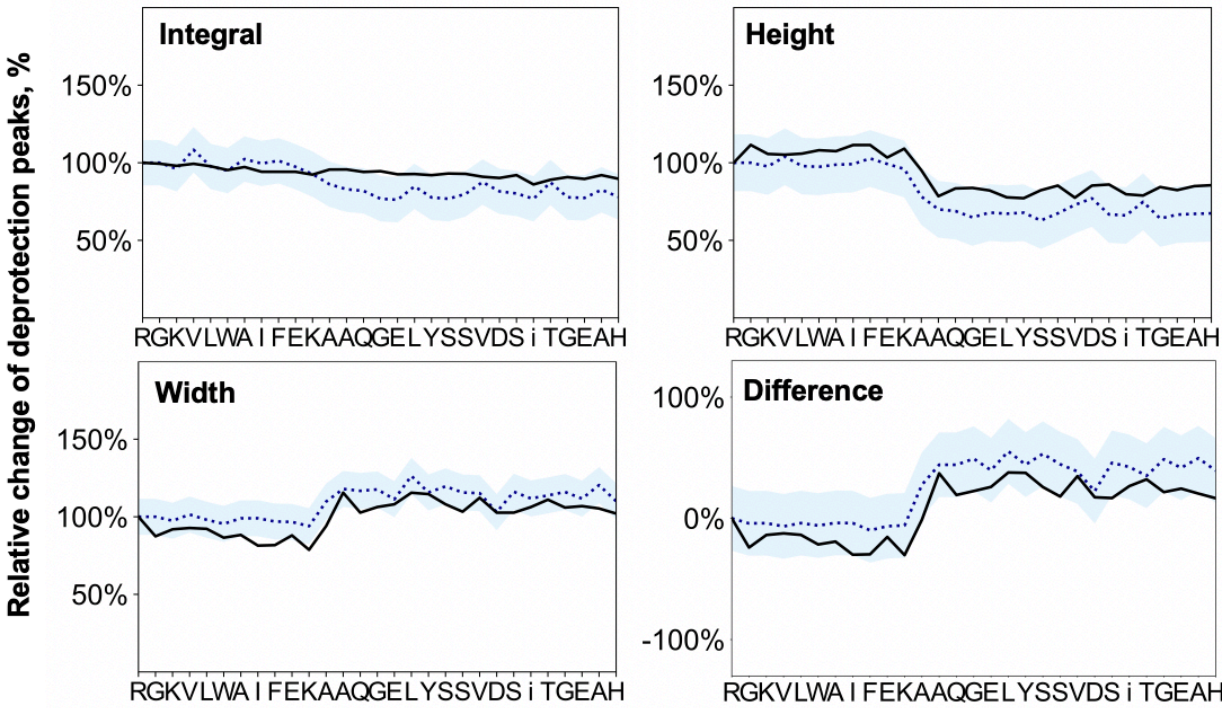
Synthesis Data for GLP-1 (Fmoc-Phe-Thr($\Psi^{Me,Me}$pro)-OH)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR (30 AA);
FT = Fmoc-Phe-Thr($\Psi^{Me,Me}$pro)-OH

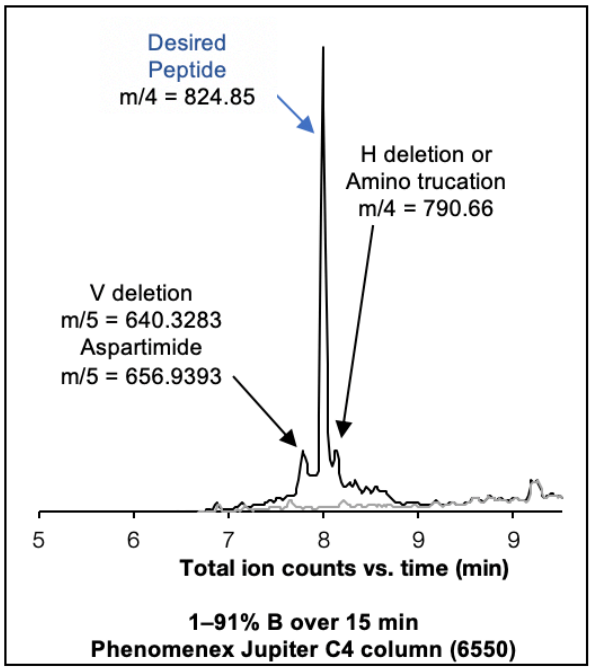
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

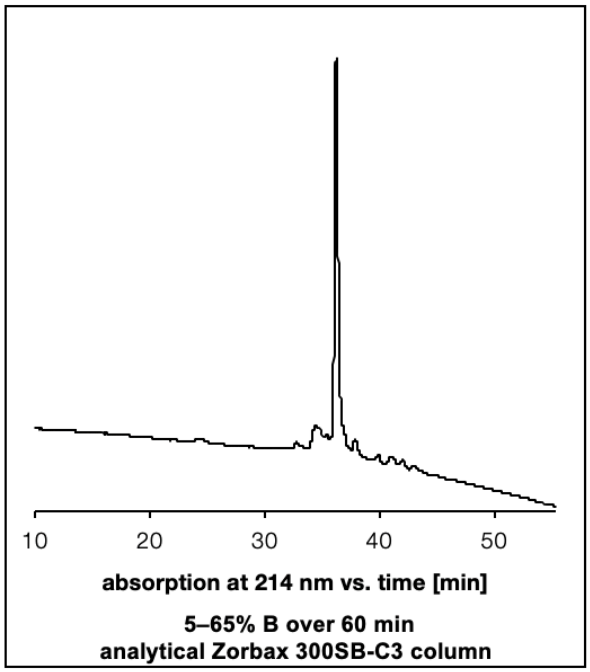
a) LCMS data, TIC



b) LCMS data, TIC



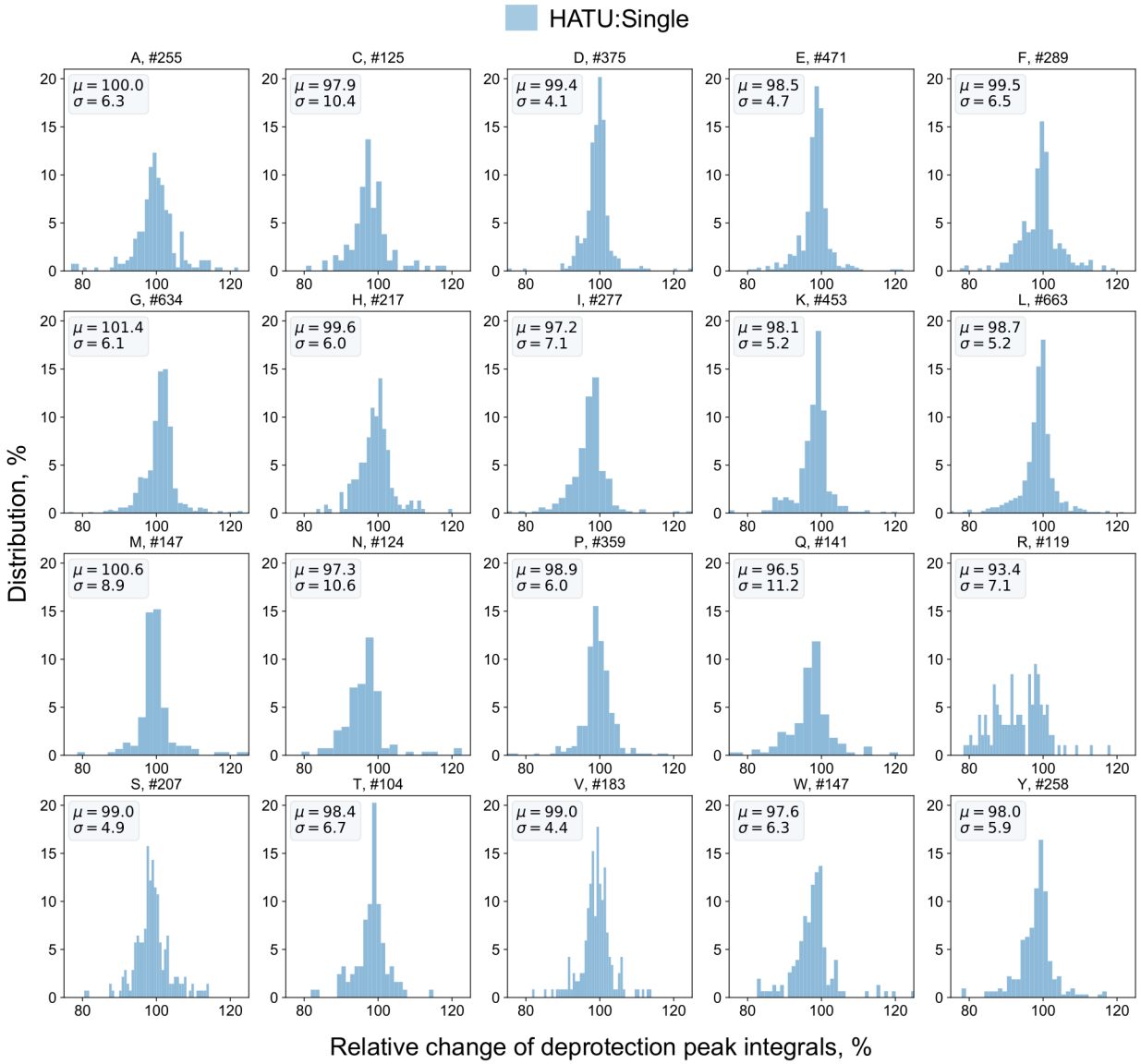
c) HPLC data



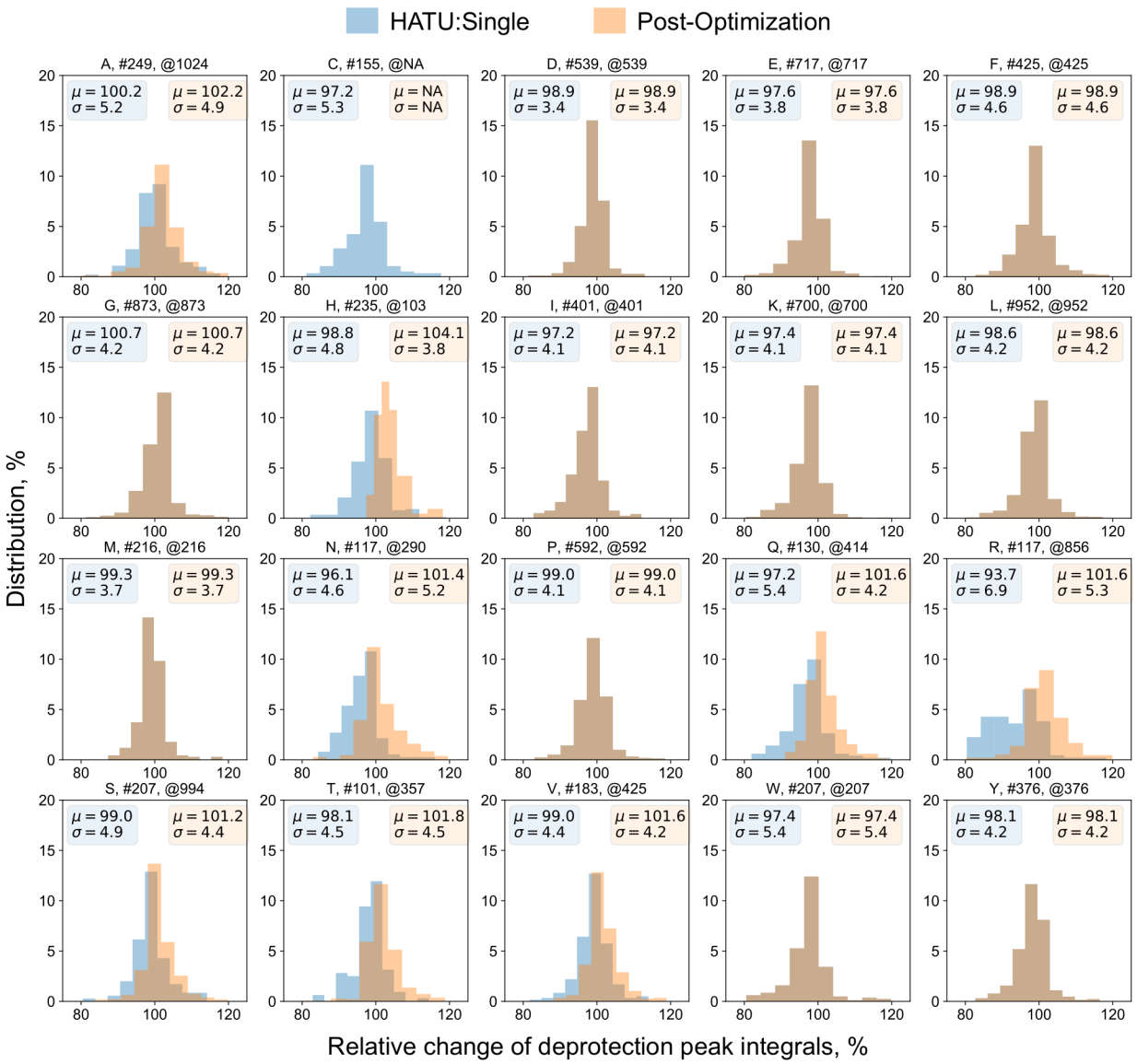
5 Statistical analysis of AFPS data set

5.1 Distribution of integrals for different synthesis parameters

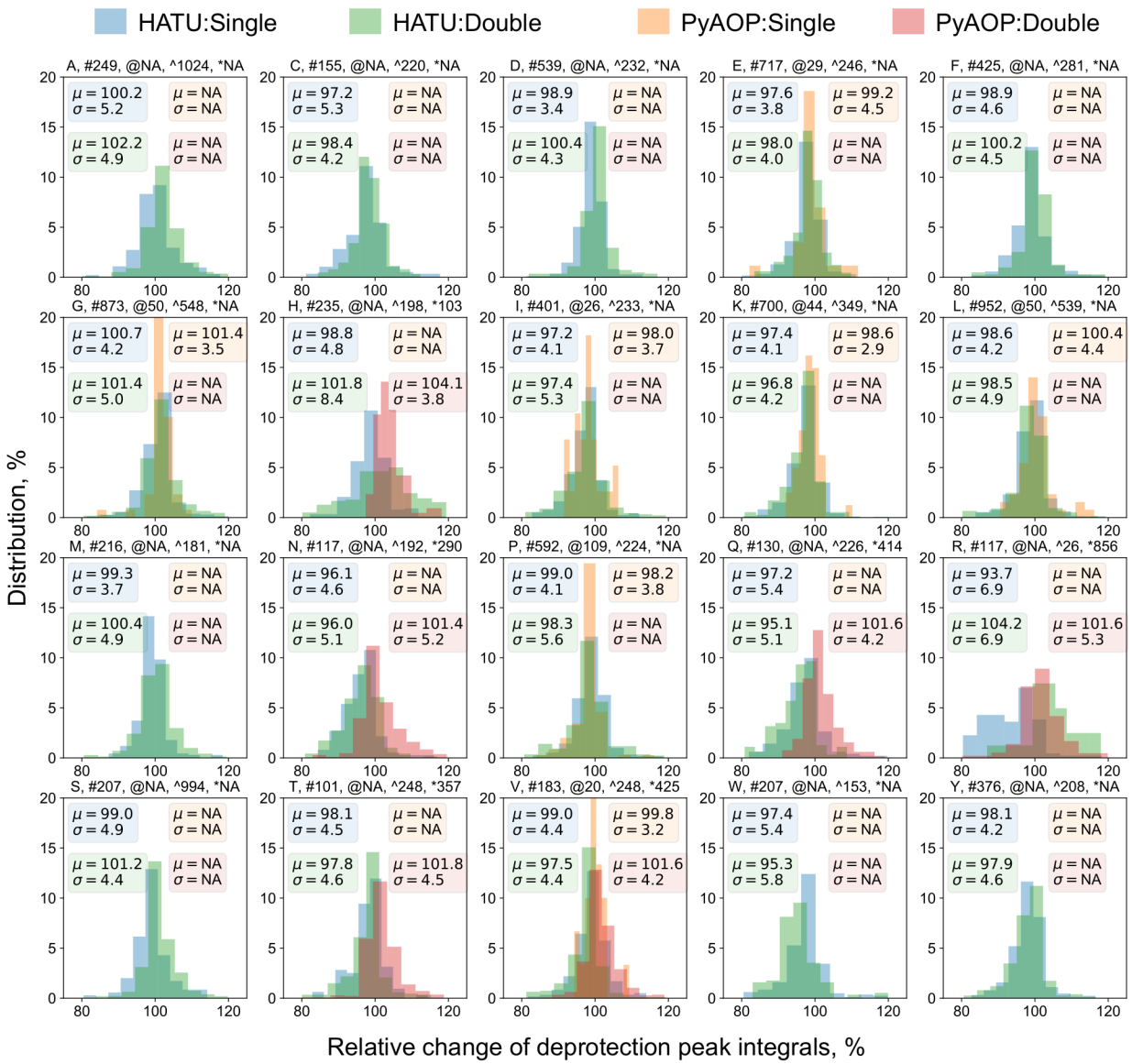
Deprotection peak integrals were analyzed for different combinations of synthesis parameters for coupling of each incoming amino acid (**SI Figure 3-5, Appendix 2**). On an average, PyAOP and double coupling strokes are seen to be more effective than HATU and single coupling strokes respectively. Apart from the mean of the distribution being shifted towards the right, indicating better coupling, the spread is also narrower thereby indicating more consistency of coupling thus reproducibility.



SI Figure 3. Distribution of integrals for optimized recipe by amino acid. Relative change of deprotection peak integrals for reaction steps with HATU coupling agent, single coupling stroke and 40 mL/min flow rate by amino acid. The number of data points for each amino have been mentioned after #.



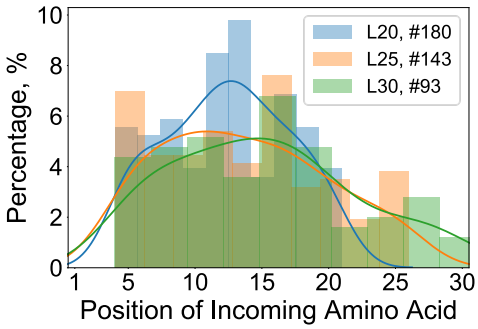
SI Figure 4. Distribution of deprotection peak integrals by amino acid for HATU coupling agent and single coupling stroke (blue) and after optimization parameters (red).[3] In the optimized synthesis protocol, A and S were coupled with HATU (double coupling) and H, N, Q, R, T and V were coupled with PyAOP (double couplings). It is noteworthy, that C was also coupled with PyAOP (double couplings) under the final conditions, however, only a few data points existed for this amino acid. The number of data points for each combination are noted above the specific distribution after the symbol notation - # for HATU:Single and @ for optimized parameters. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable).



SI Figure 5. Distribution of deprotection peak integrals by amino acid for different combinations of coupling agent (HATU, PyAOP) and coupling strokes (Single, Double). The number of data points for each combination are noted above the specific distribution after the symbol notation - # for HATU:Single, @ for PyAOP:Single, ^ for HATU:Double and * for PyAOP:Double. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable).

5.2 Onset of aggregation

To identify the position from C-terminus where aggregation starts, we analyzed all pre-chains of aggregating sequences (**SI Figure 6**). Position for onset of aggregation was defined as the first coupling-deprotection step where the difference between width and height in the deprotection trace was greater than 20%. Sequences of lengths greater than 20, 25 and 30 were analyzed individually. The maximum position at which aggregation can start was restricted to the minimum length for each case. For instance, in the analysis of all sequences greater than 20 amino acids, the data set was restricted to the sequences where the aggregation starts before 20 amino acids. For subsequent lengths, the dataset for analysis included the peptides with lengths shorter than the threshold, such as analysis of sequences less than 30 amino acids included sequences with less than 20 amino acids. It was found that 49%, 60% and 69% of sequences of lengths greater than 20, 25 and 30, respectively, are aggregating, according to our definition.

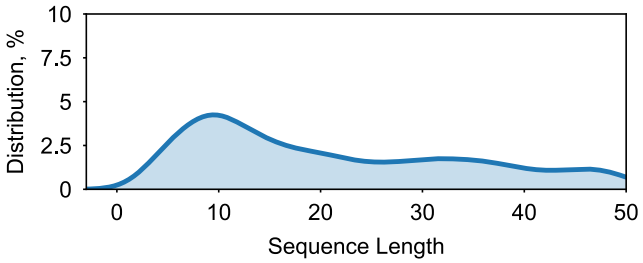


SI Figure 6. The onset of aggregation is analyzed for sequences of length (L) greater than 20, 25 and 30. # followed by the numerical quantity indicates the number of sequences in the dataset with the unique pre-chain where aggregation starts.

6 Statistical analysis of PDB data set

6.1 Downloading and pre-processing of data set

The PDB dataset was downloaded and pre-processed (accessed on April 17, 2020).[4] From the FASTA file, only sequences with less than equal to 50 amino acids were selected. Redundant sequences and sequences with unnatural residues were removed. A total of 8441 out of 33982 sequences remained after the pre-processing, and were used for further analysis (**SI Figure 7**).



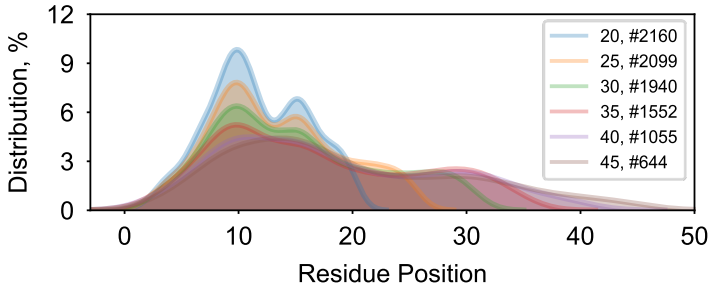
SI Figure 7. Distribution of sequences for different sequence lengths.

6.2 Prediction of aggregation

Complete traces for difference were obtained for all sequences using the pre-trained model. If difference at all coupling-deprotection steps was less than 20%, then the sequence was marked as a non-aggregating sequence. If aggregation was seen at a particular step, then the pre-chain of that step was added to the list of aggregating pre-chains and the sequence was added to the list of aggregating sequences. Based on the heuristic definition of aggregation used in the current study, 3815 out of 8441, or 45% of the sequences were predicted to have at least one aggregating coupling-deprotection step.

6.3 Onset of aggregation

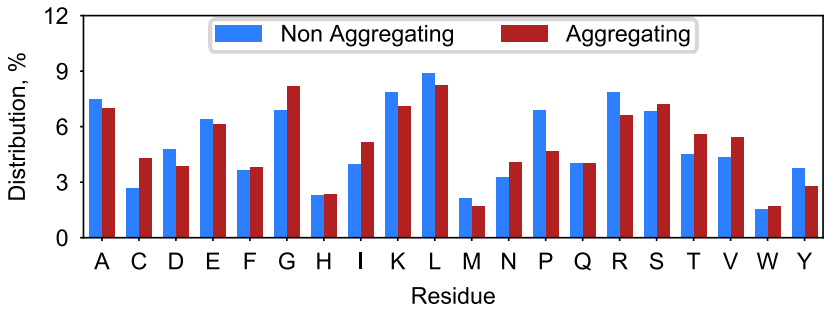
Onset of aggregation for the PDB sequences was calculated for sequences with different minimum lengths, similar to SI Section 5.2 (**SI Figure 8**).



SI Figure 8. The onset of aggregation is analyzed for PDB sequences of different minimum lengths. # followed by the numerical quantity indicates the number of sequences in the dataset with the unique pre-chain where aggregation starts.

6.4 Distribution of amino acids

Distribution of amino acids in non-aggregating sequences and pre-chain at the aggregating step of aggregating sequences was similar (**SI Figure 9**). Based on this, it may be said that aggregation is mostly independent of the residue composition.

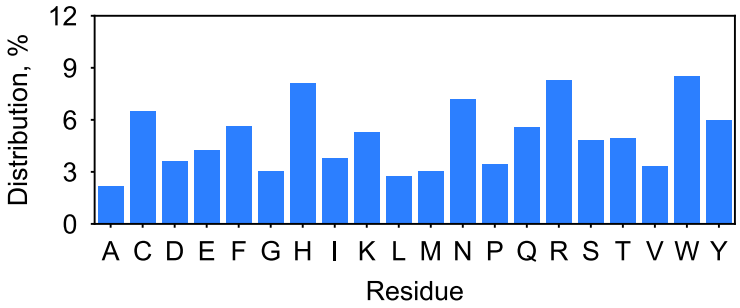


SI Figure 9. Distribution of amino acids in non-aggregating sequences and pre-chain at the aggregating step of aggregating sequences is similar.

6.5 Activation analysis

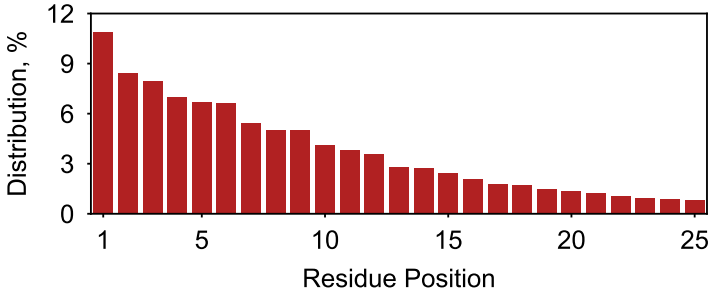
Activation maps for pre-chains of all aggregating sequences were calculated. For analysis, sum of all activations by position, residue and bit-indices for each residue was done for sequences with greater than equal to 25 residues.

Residues and side chain protecting groups with aryl groups are found to be the most activated for aggregation, consistent with previous findings. (**SI Figure 10; SI Table 3**).



SI Figure 10. Distribution of residues responsible for aggregation, as calculated from gradient activation maps.

Residues at the C-terminus are predicted to be the principal contributors to aggregation (**SI Figure 11**). There is a polynomial decay in the contribution of aggregation from other positions for this specific data set. This result is consistent with the mutations of GLP-1 and JR-10 (**SI Section 4**) which were predicted and experimentally validated. A majority of mutants, both less and more aggregating, had single point mutations for C-terminal residues. Further, the predicted mutants for the difficult-to-synthesize sequences (**SI Section 3.8**) demonstrate a similar trend.



SI Figure 11. Distribution of residue positions responsible for aggregation, as calculated from gradient activation maps.

SI Table 4. List of indices for the most activated substructures for the residues contributing most to aggregation. Barring indices redundant across all amino acids and those belong to the amino acid scaffold, bulkier protecting groups are most activated.

Most activated indices	Trp	Arg	His	Asn	Cys	Tyr	Phe
1	25	16	70	104	83	78	42
2	77	93	73	101	85	66	79
3	101	68	87	11	37	73	32
4	93	53	13	36	116	54	73
5	26	44	17	56	88	61	64
6	37	111	79	122	2	79	89
7	116	88	107	9	45	89	22
8	121	21	111	127	61	36	121
9	45	121	115	47	47	22	39
10	111	34	112	106	9	60	5

7 References

(1) Collins, J. M.; Porter, K. A.; Singh, S. K.; Vanier, G. S. High-Efficiency Solid Phase Peptide Synthesis (He -Spps). *Org. Lett.* **2014**, *16* (3), 940–943. https://doi.org/10.1021/ol4036825.

(2) Atherton, E.; Woolley, V.; Sheppard, R. C. Internal Association in Solid Phase Peptide Synthesis. Synthesis of Cytochrome C Residues 66-104 on Polyamide Supports. *J. Chem. Soc. Chem. Commun.* **1980**, No. 20, 970–971. https://doi.org/10.1039/C39800000970.

(3) Hartrampf, N.; Saebi, A.; Poskus, M.; Gates, Z. P.; Callahan, A. J.; Cowfer, A. E.; Hanna, S.; Antilla, S.; Schissel, C. K.; Quartararo, A. J.; et al. Synthesis of Proteins by Automated Flow Chemistry. *ChemRxiv. Prepr.* **2020**, No. 2. https://doi.org/10.26434/chemrxiv.11833503.v1.

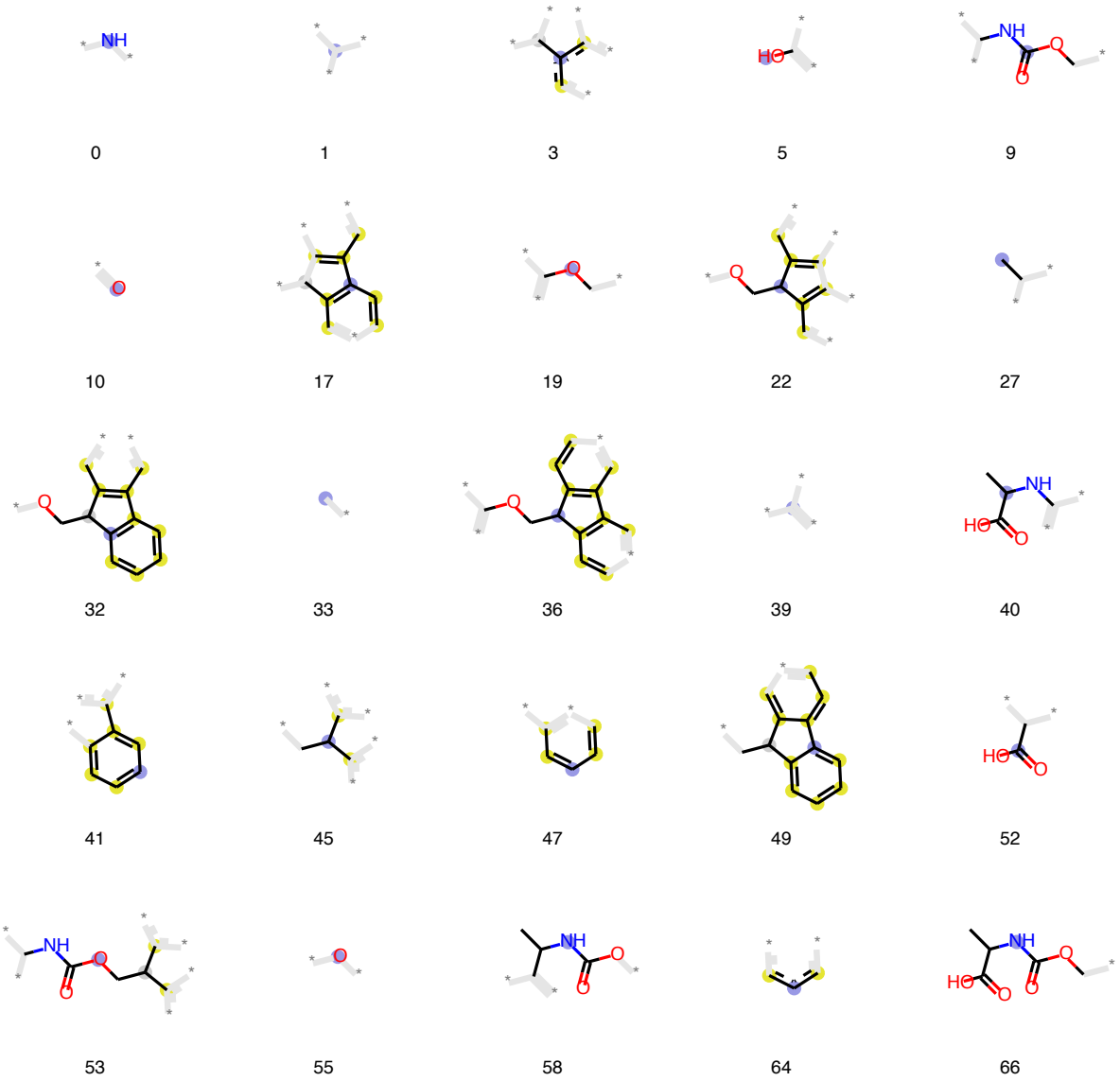
(4) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

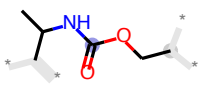
8 Appendix 1

The substructures used in the training of the model have been indexed by their respective bit-vector indices. The blue shaded circle represents the node atom of the substructure and dark bonds depict the topological exploration of the n-nearest neighbors. The bonds and atoms that are not a part of the specific topological exploration are in grey color. Atoms which are a part of an aromatic ring have a yellow shaded circle to differentiate them from the rest.

8.1 Substructures for incoming amino acids

Alanine



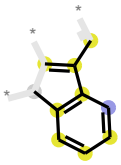


67

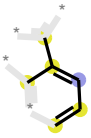
80

81

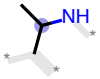
86



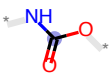
87



88

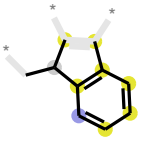


93



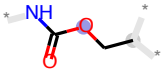
94

96



97

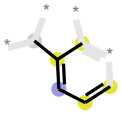
100



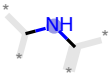
109

110

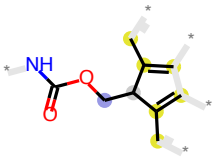
112



115



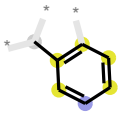
117



120

123

125



127

Cysteine

0

1

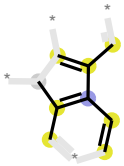
2

3

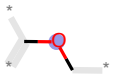
5

9

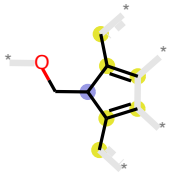
10



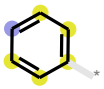
17



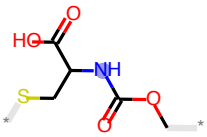
19



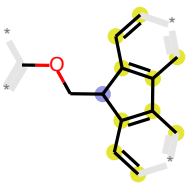
22



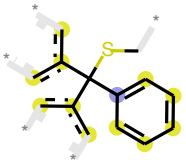
27



32

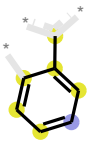


36



37

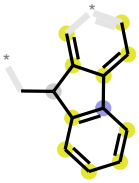
39



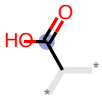
41

45

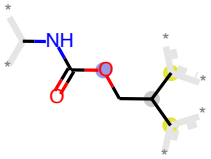
47



49

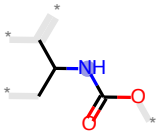


52

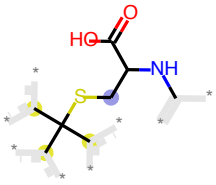


53

55

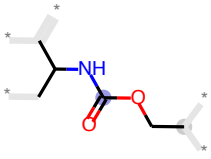


57



61

64

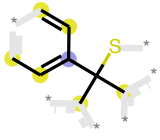


66

69

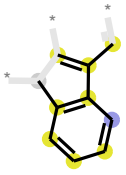
80

81

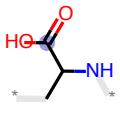


83

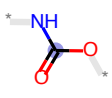
86



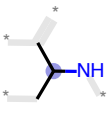
87



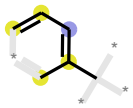
88



94

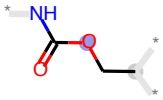


96

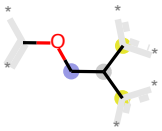


97

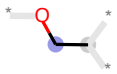
100



109

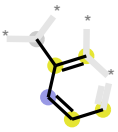


110



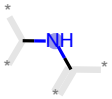
112

114

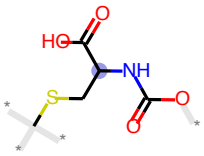


115

116



117

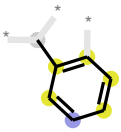


120

123

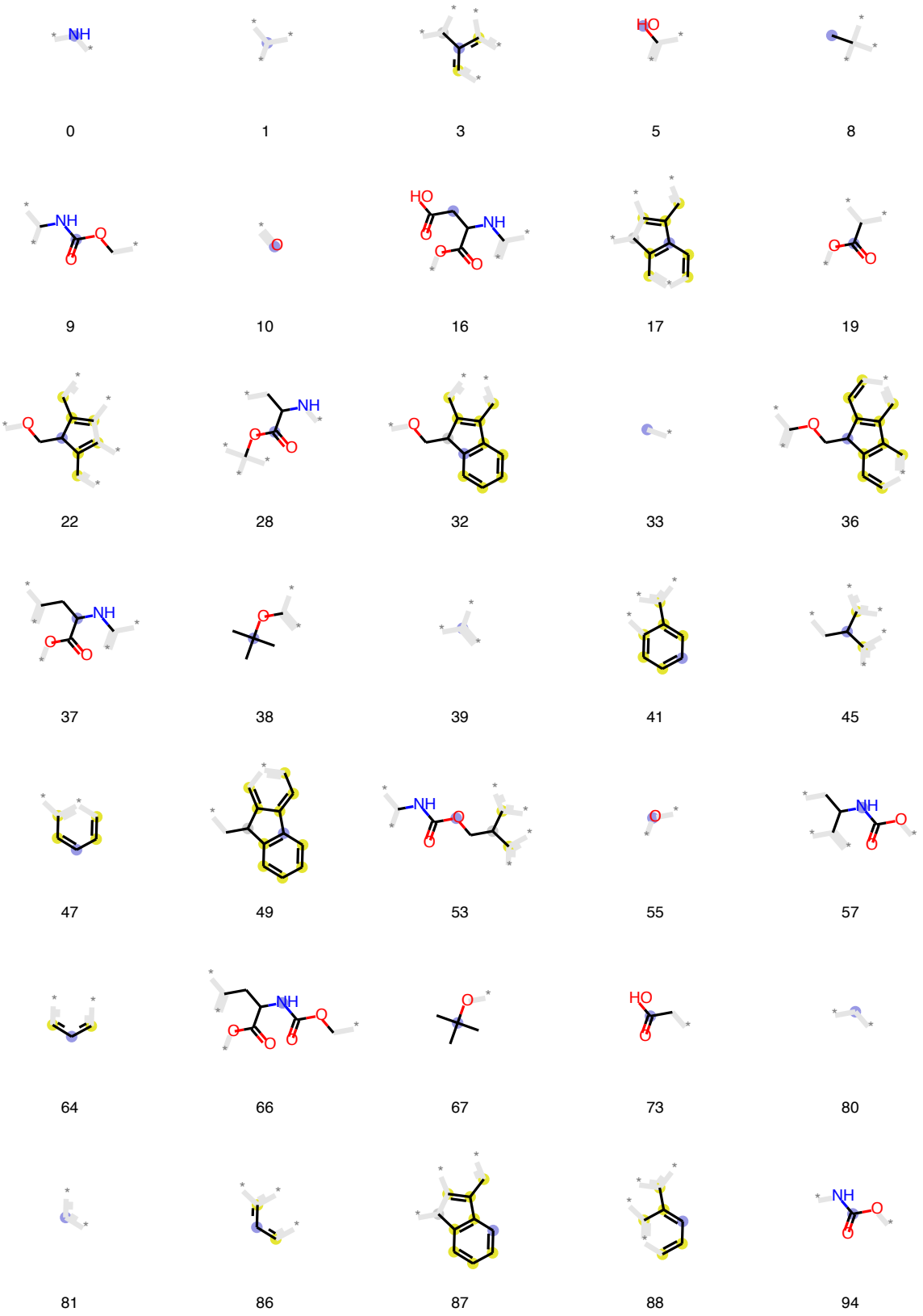
124

125

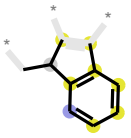


127

Aspartic acid



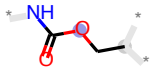
96



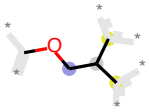
97

100

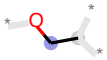
103



109

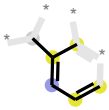


110

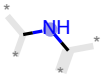


112

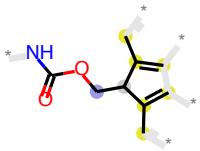
114



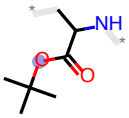
115



117



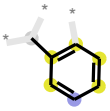
120



122

123

125



127

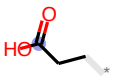
Glutamic acid

0

1

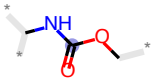
3

5



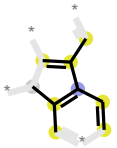
7

8



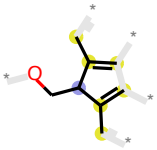
9

10



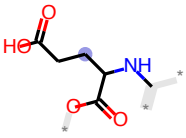
17

19

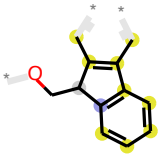


22

28

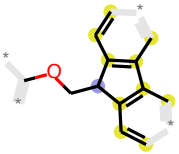


29



32

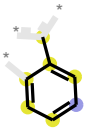
33



36

38

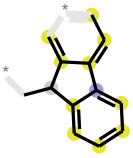
39



41

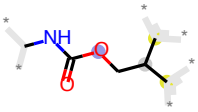
45

47



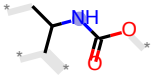
49

52

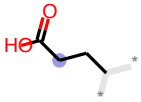


53

55



57

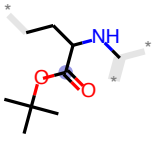


59

64

66

67



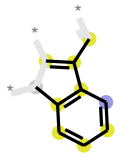
70

73

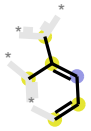
80

81

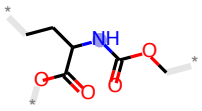
86



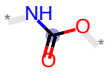
87



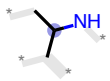
88



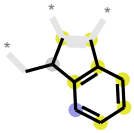
90



94



96

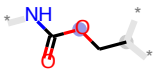


97

99

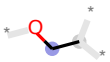
100

103



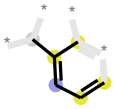
109

110

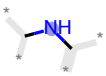


112

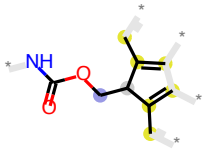
114



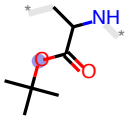
115



117



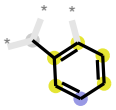
120



122

123

125



127

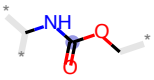
Phenylalanine

0

1

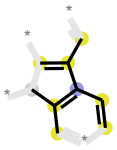
3

5



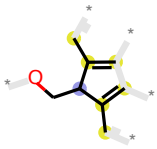
9

10



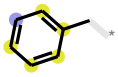
17

19

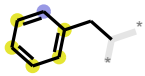


22

27

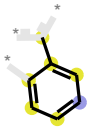


32

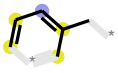


36

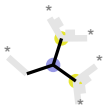
39



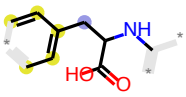
41



42

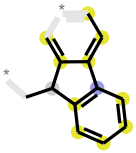


45



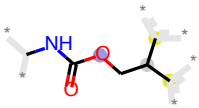
46

47



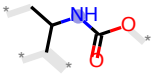
49

52



53

55



57

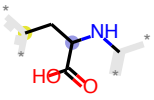
64

66

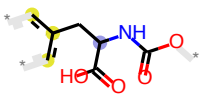
79

80

81

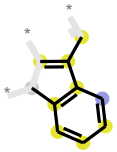


83



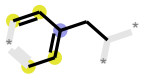
84

86



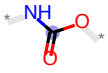
87

88

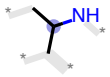


89

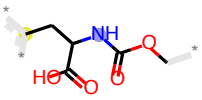
90



94

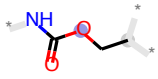


96

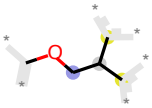


97

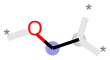
100



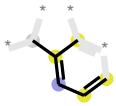
109



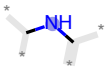
110



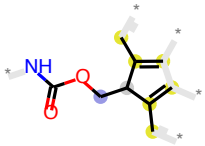
112



115



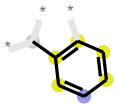
117



120

123

125

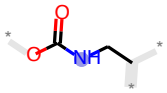


127

Glycine

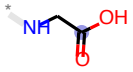
0

3



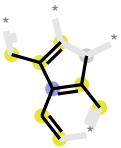
4

5

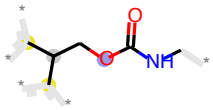


6

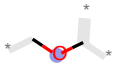
10



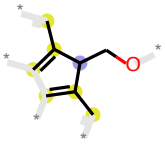
17



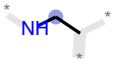
18



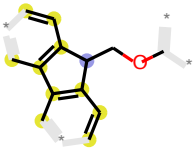
19



22

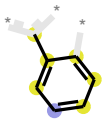


32

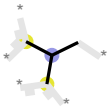


36

39

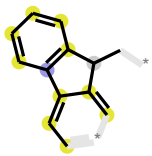


41



45

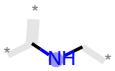
47



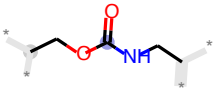
49

55

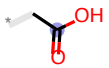
64



69



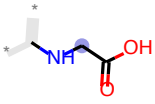
71



73

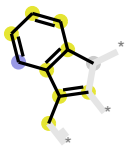
80

81

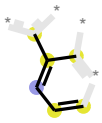


83

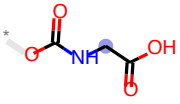
86



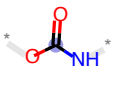
87



88

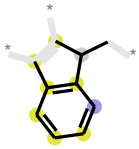


91



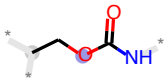
94

96

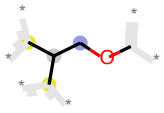


97

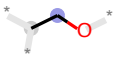
100



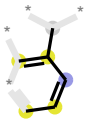
109



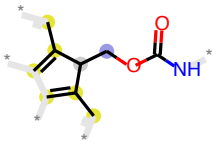
110



112



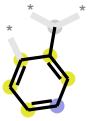
115



120

123

125

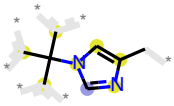


127

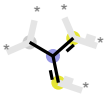
Histidine

0

1

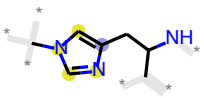


2



3

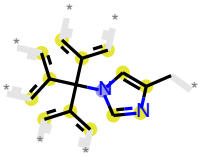
5



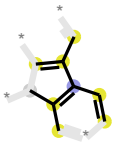
6

9

10

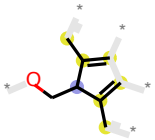


16

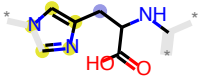


17

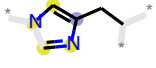
19



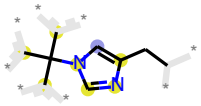
22



27



29

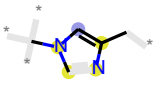


31

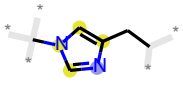
32

36

39

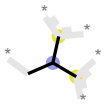


40

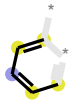


41

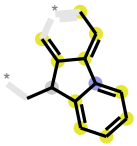
44



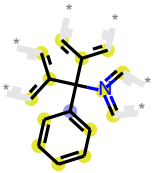
45



47

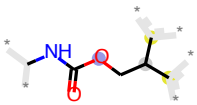


49



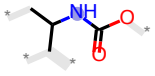
50

52

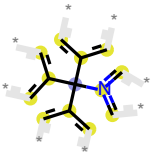


53

55

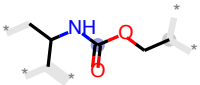


57

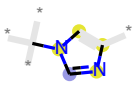


61

64



66

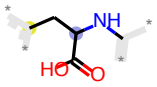


69

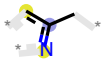
79

80

81

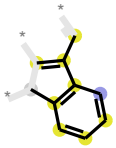


83

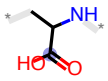


84

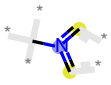
86



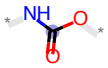
87



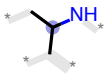
88



93



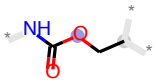
94



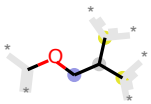
96

97

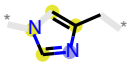
100



109



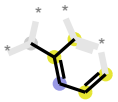
110



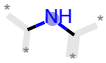
111

112

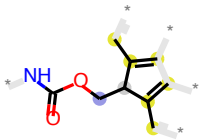
114



115



117

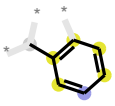


120

122

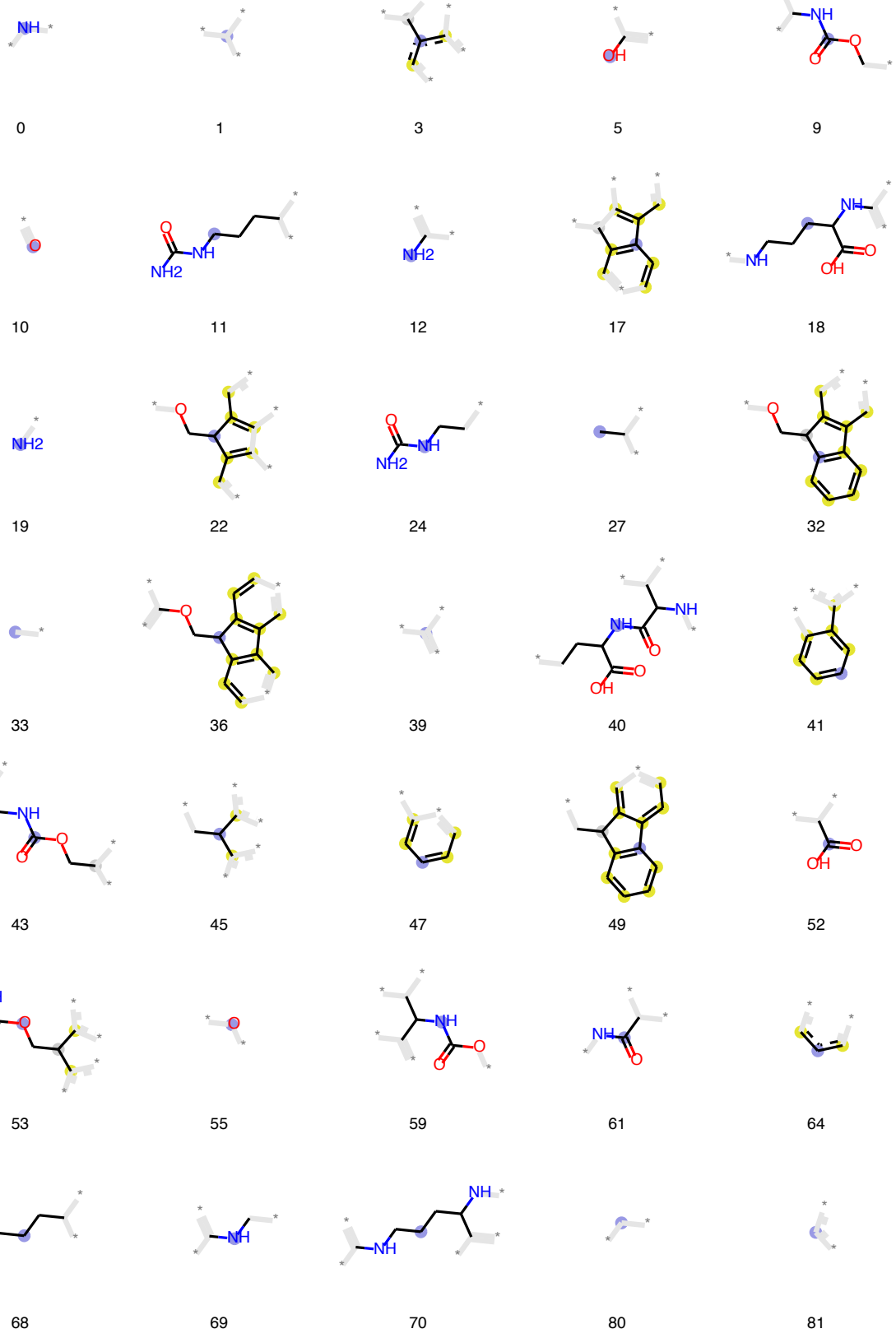
123

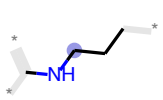
125



127

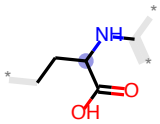
Isoleucine


0
1
3
5
9
10
11
12
17
18
19
22
24
27
32
33
36
39
40
41
43
45
47
49
52
53
55
59
61
64
68
69
70
80
81



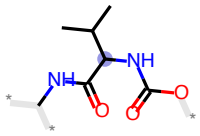
84

86

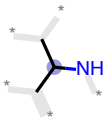


87

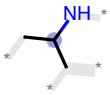
88



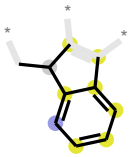
91



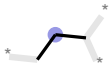
94



96



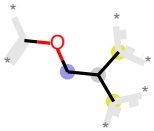
97



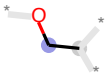
99

100

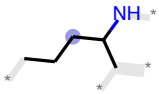
109



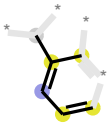
110



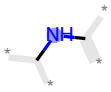
112



113



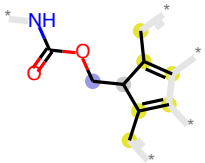
115



117

118

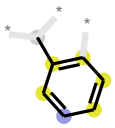
119



120

123

125



127

Lysine

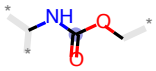
0

1

3

5

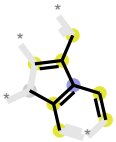
8



9

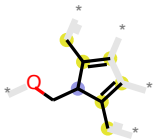
10

16

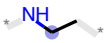


17

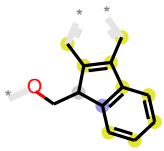
19



22

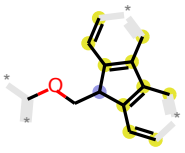


27

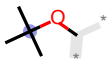


32

33

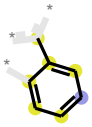


36

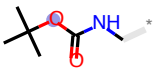


38

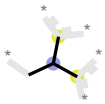
39



41

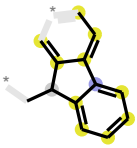


42



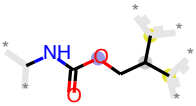
45

47



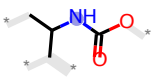
49

52



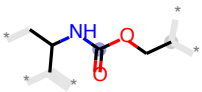
53

55



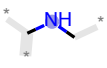
57

64



66

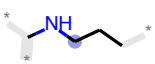
67



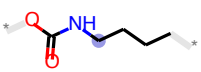
69

80

81

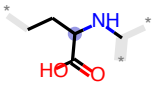


84

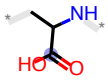


85

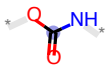
86



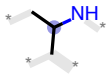
87



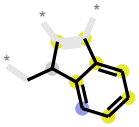
88



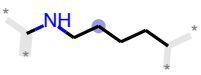
94



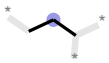
96



97



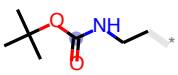
98



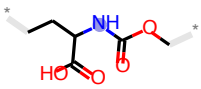
99

100

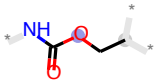
103



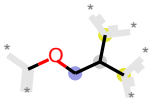
105



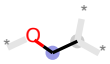
108



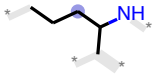
109



110

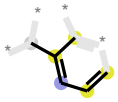


112

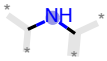


113

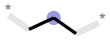
114



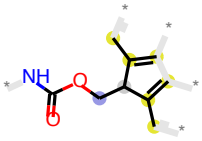
115



117



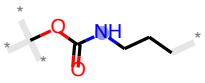
119



120

123

125



127

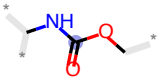
Leucine

0

1

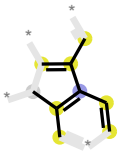
3

5



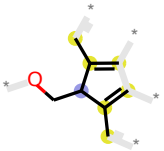
9

10



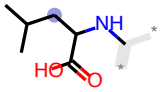
17

19

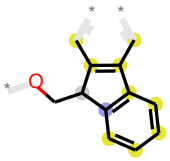


22

27

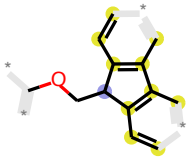


30



32

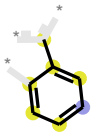
33



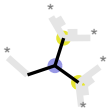
36

39

40

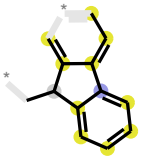


41



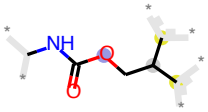
45

47



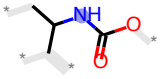
49

52



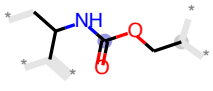
53

55

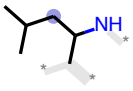


57

64



66

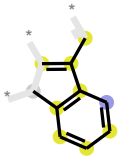


74

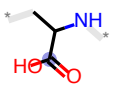
80

81

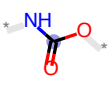
86



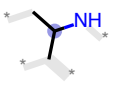
87



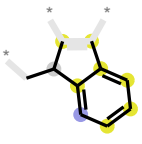
88



94

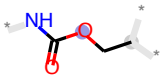


96

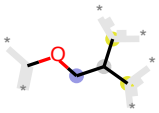


97

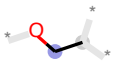
100



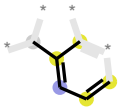
109



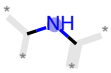
110



112

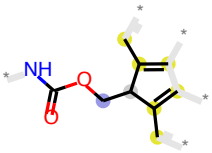


115



117

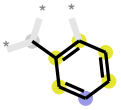
119



120

123

125



127

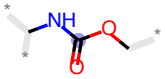
Methionine

0

1

3

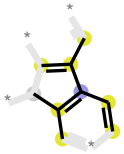
5



9

10

11



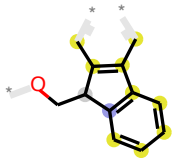
17

19

22

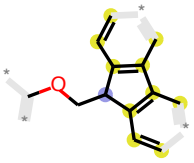
25

30



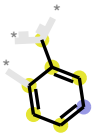
32

33



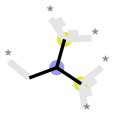
36

39



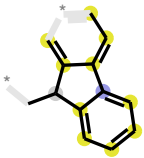
41

44



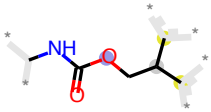
45

47



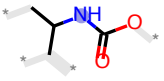
49

52

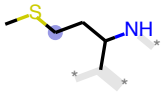


53

55

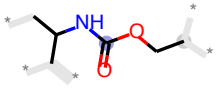


57



62

64

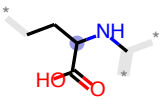


66

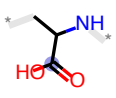
80

81

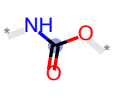
86



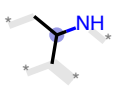
87



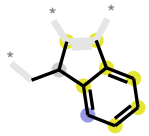
88



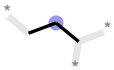
94



96

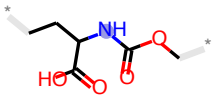


97

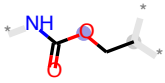


99

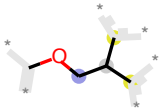
100



108

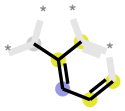


109



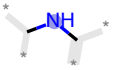
110

112

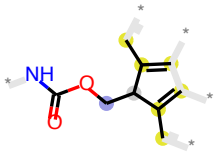


115

116



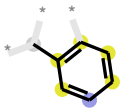
117



120

123

125



127

Asparagine

0

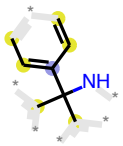
1

3

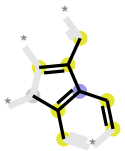
5

9

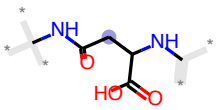
10



11

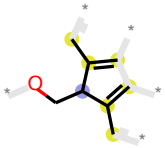


17

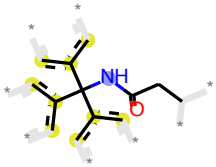


18

19

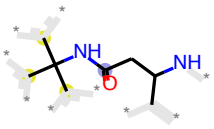


22

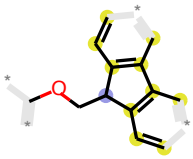


24

27



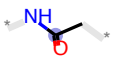
32



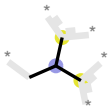
36

39

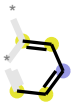
40



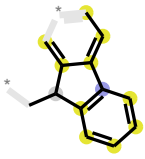
41



45

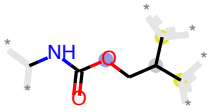


47



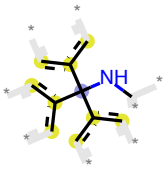
49

52

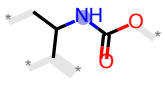


53

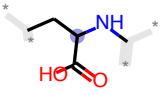
55



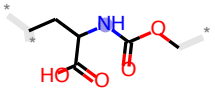
56



57

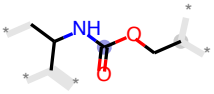


58

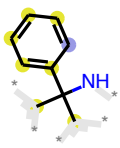


59

64



66

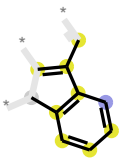


69

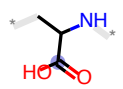
80

81

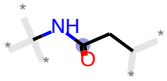
86



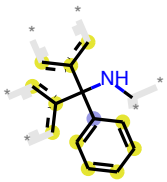
87



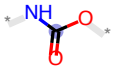
88



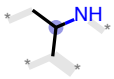
89



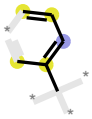
91



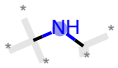
94



96



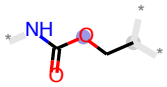
97



98

100

105

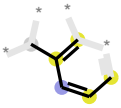


109

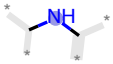
110

112

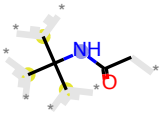
114



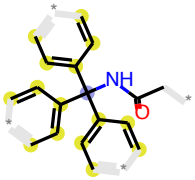
115



117



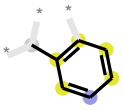
120



122

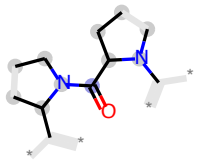
123

125

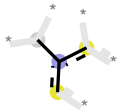


127

Proline



1

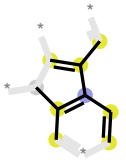


3

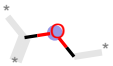
4

5

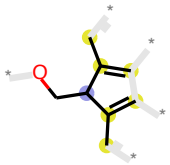
10



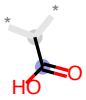
17



19

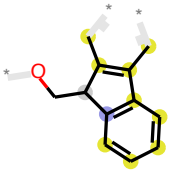


22

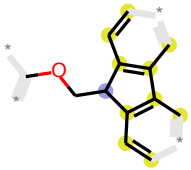


23

30

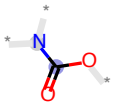


32

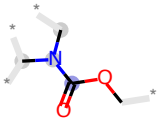


36

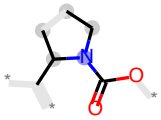
39



41



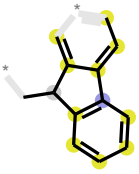
42



43

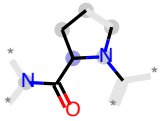
45

47



49

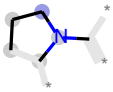
55



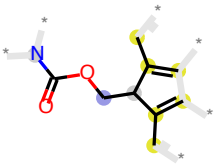
58

62

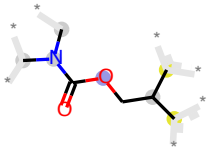
64



65

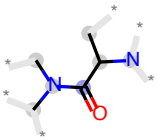


66



70

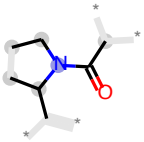
72



78

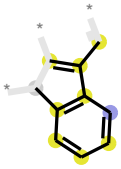
80

81

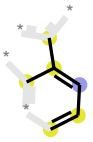


85

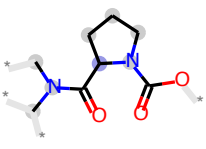
86



87

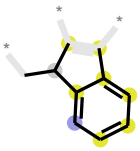


88

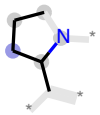


93

96

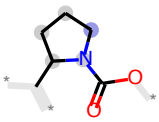


97

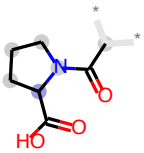


98

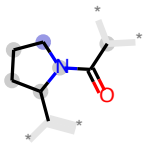
100



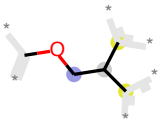
102



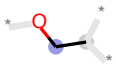
105



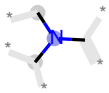
108



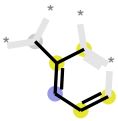
110



112



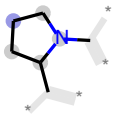
114



115

123

125



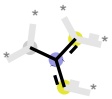
126

127

Glutamine

0

1

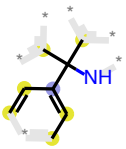


3

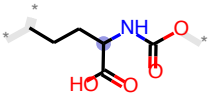
5

9

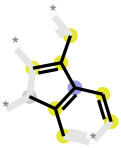
10



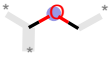
11



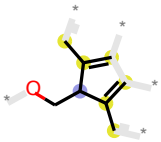
16



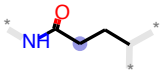
17



19

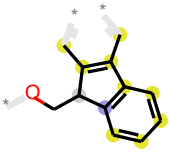


22

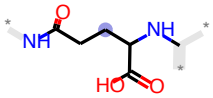


27

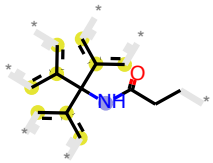
28



32

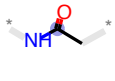


33

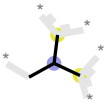


36

39

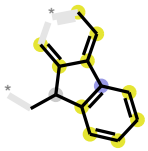


41



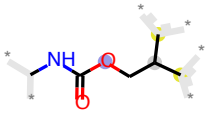
45

47



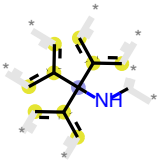
49

52

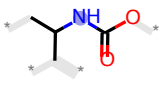


53

55

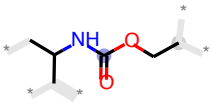


56

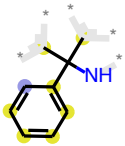


57

64



66

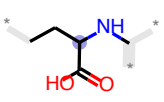


69

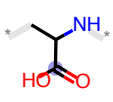
80

81

86



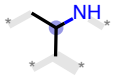
87



88

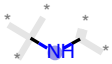
91

94

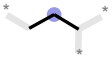


96

97

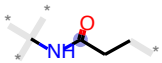


98



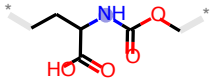
99

100

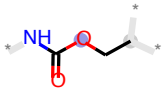


101

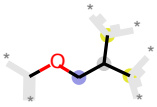
105



108



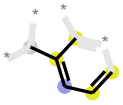
109



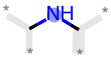
110

112

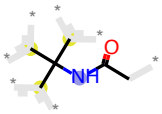
114



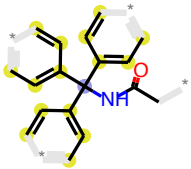
115



117



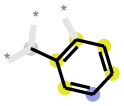
120



122

123

125

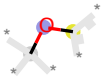


127

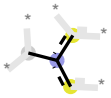
Arginine

0

1

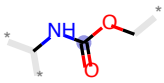


2



3

5



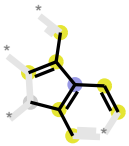
9

10

12

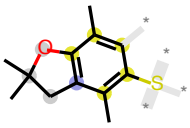
14

16

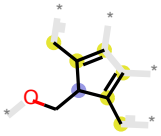


17

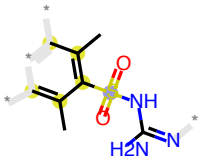
19



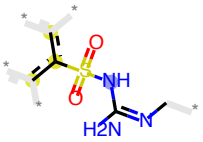
21



22

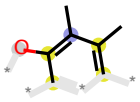


23



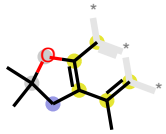
27

30



32

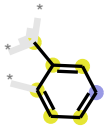
33



34

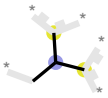
36

39

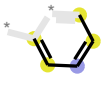


41

44



45

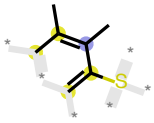


47

49

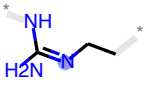
51

52

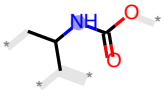


53

55

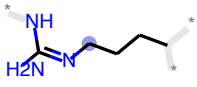


56

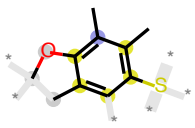


57

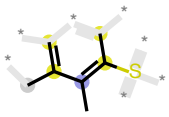
58



61

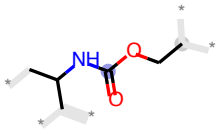


62



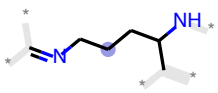
63

64

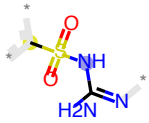


66

68



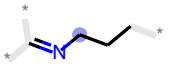
70



74

80

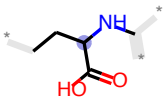
81



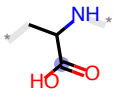
84

85

86



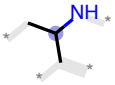
87



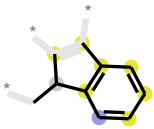
88

93

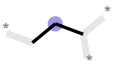
94



96

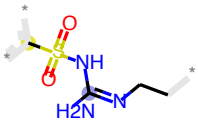


97



99

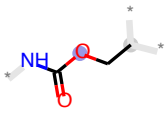
100



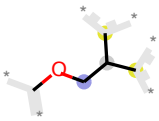
103

105

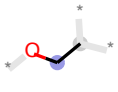
108



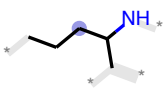
109



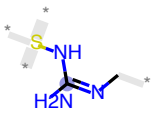
110



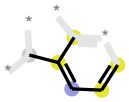
112



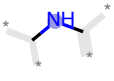
113



114

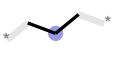


115

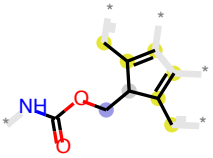


117

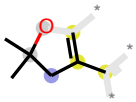
118



119



120

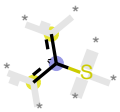


121

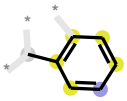
122

123

124



125

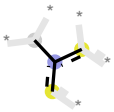


127

Serine

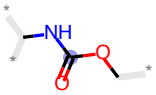
0

1



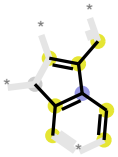
3

8



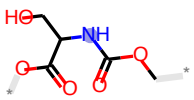
9

10

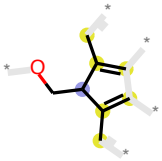


17

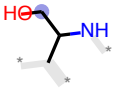
19



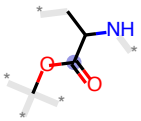
20



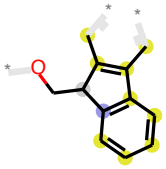
22



23

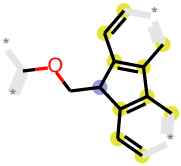


28



32

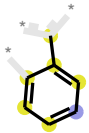
33



36

38

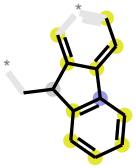
39



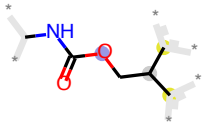
41

45

47

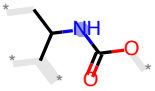


49



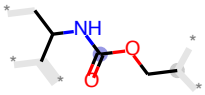
53

55



57

64



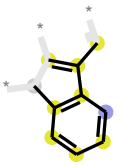
66

67

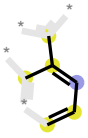
80

81

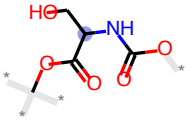
86



87

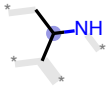


88

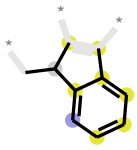


91

94



96

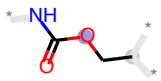


97

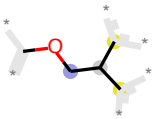
98

100

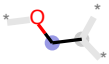
103



109

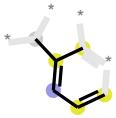


110

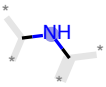


112

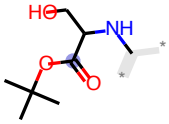
114



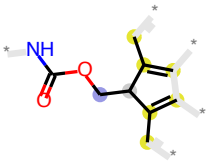
115



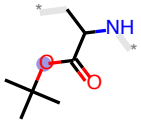
117



118



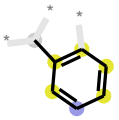
120



122

123

125

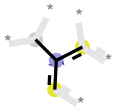


127

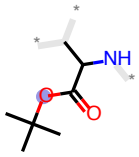
Threonine

0

1

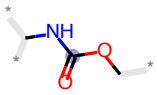


3



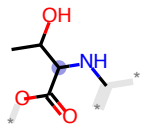
6

8

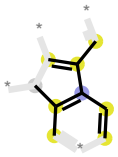


9

10

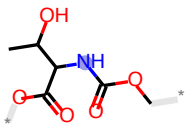


13

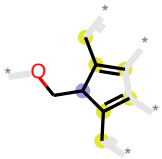


17

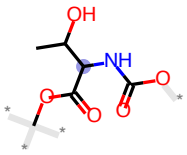
19



21



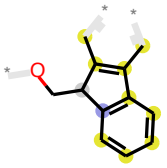
22



24

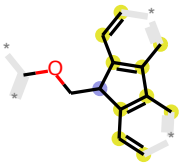
27

29



32

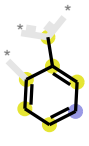
33



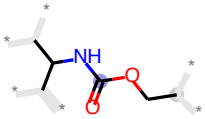
36

38

39



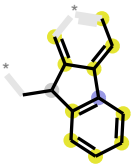
41



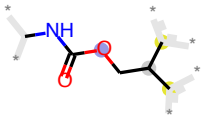
43

45

47

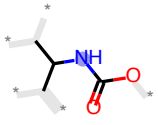


49

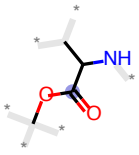


53

55

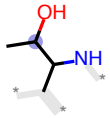


59



62

64



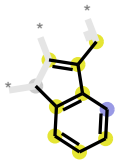
65

67

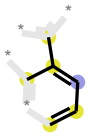
80

81

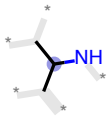
86



87

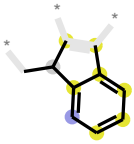


88



94

96

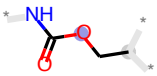


97

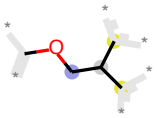
99

100

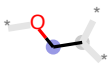
103



109

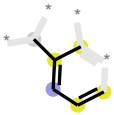


110

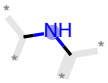


112

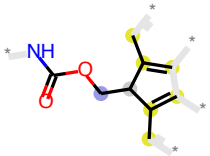
114



115



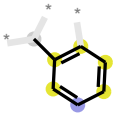
117



120

123

125



127

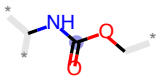
Valine

0

1

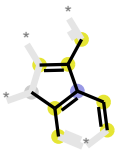
3

5



9

10

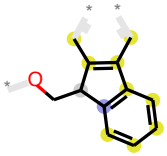


17

19

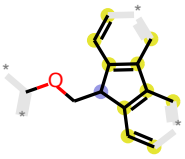
22

27



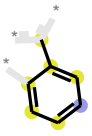
32

33

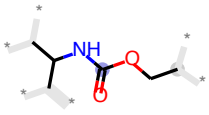


36

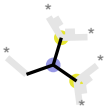
39



41

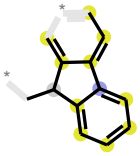


43



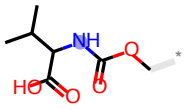
45

47



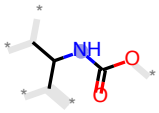
49

52

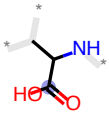


53

55

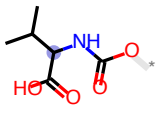


59



63

64

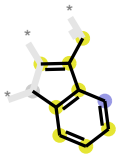


65

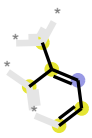
80

81

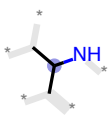
86



87

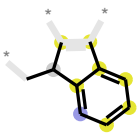


88



94

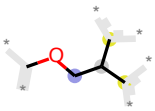
96



97

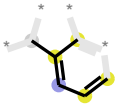
100

109

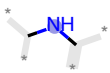


110

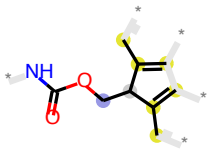
112



115



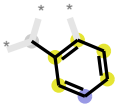
117



120

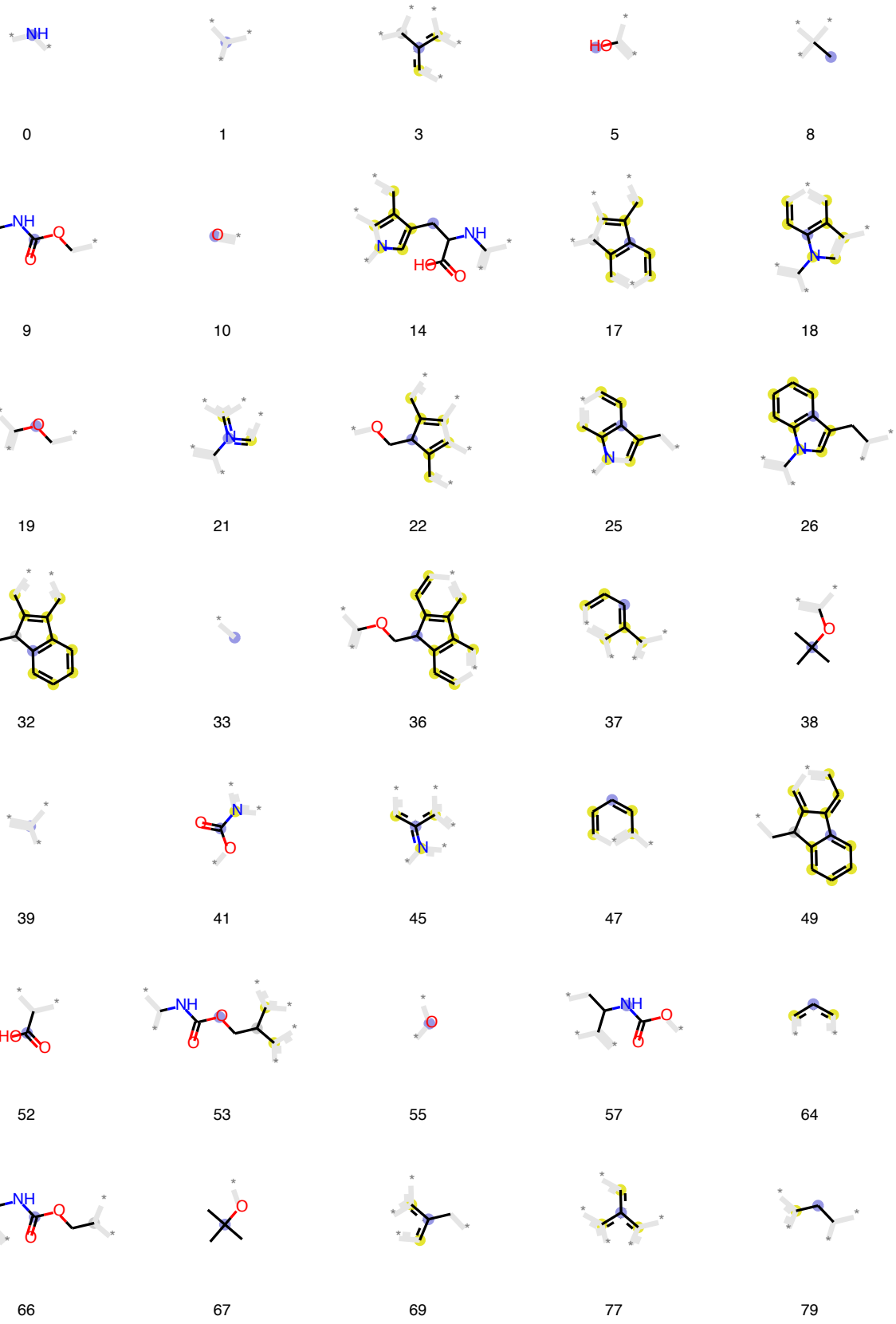
123

125



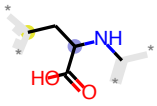
127

Tryptophan


0
1
3
5
8
9
10
14
17
18
19
21
22
25
26
32
33
36
37
38
39
41
45
47
49
52
53
55
57
64
66
67
69
77
79

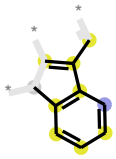
80

81

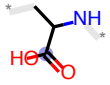


83

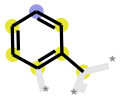
86



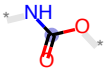
87



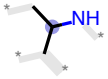
88



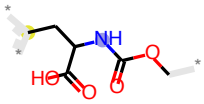
93



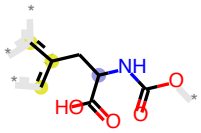
94



96

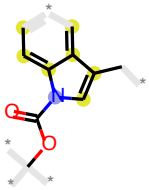


97

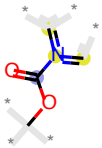


98

100

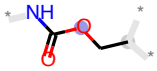


101

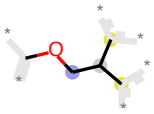


102

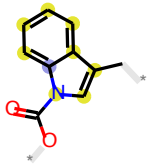
103



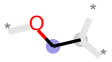
109



110

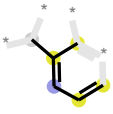


111

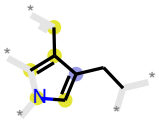


112

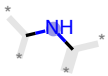
114



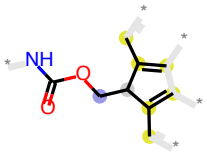
115



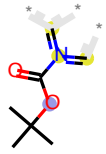
116



117



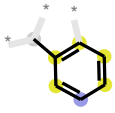
120



121

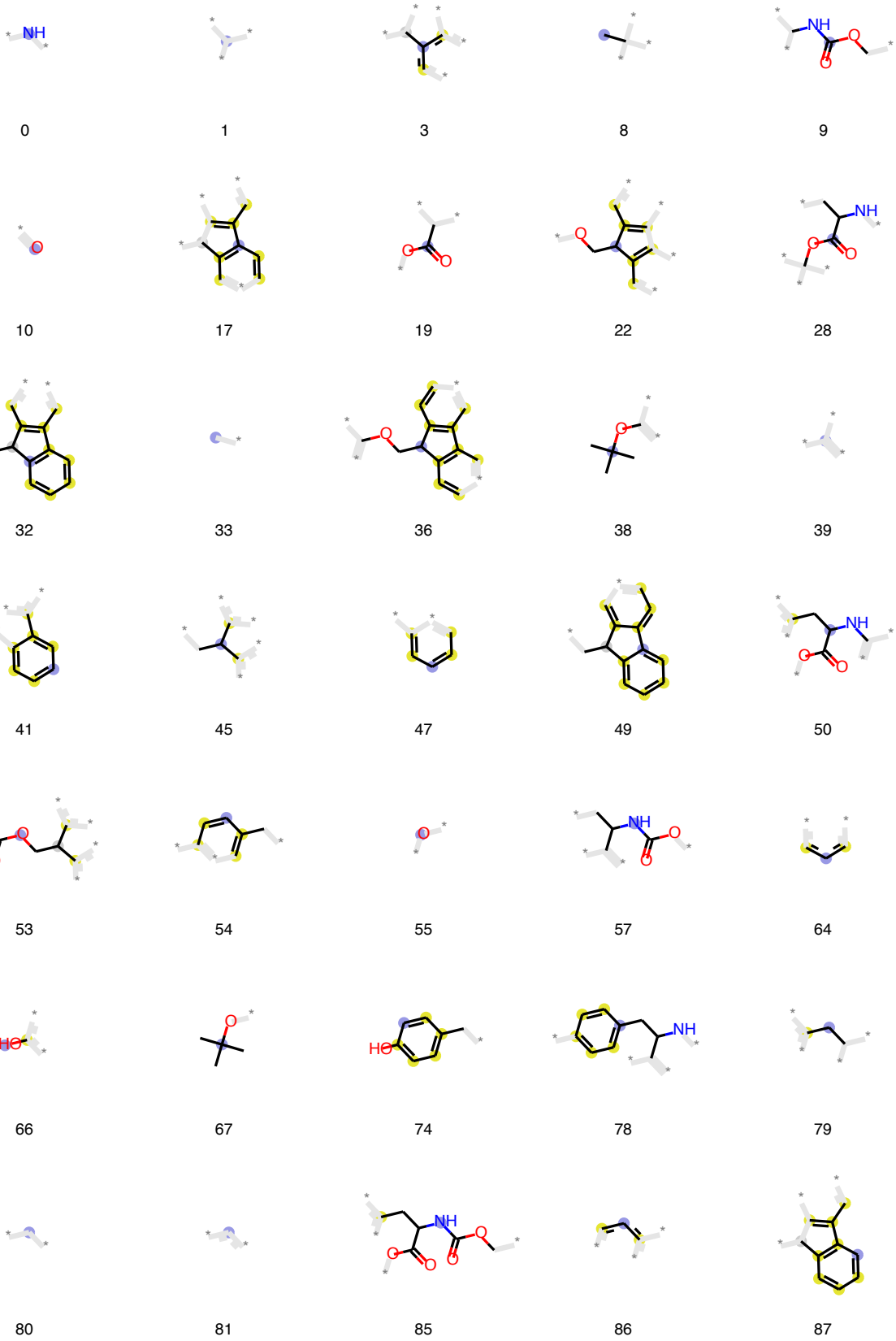
123

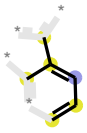
125



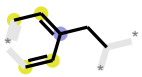
127

Tyrosine


0
1
3
8
9
10
17
19
22
28
32
33
36
38
39
41
45
47
49
50
53
54
55
57
64
66
67
74
78
79
80
81
85
86
87

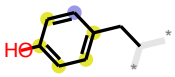


88

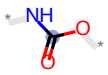


89

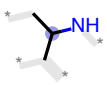
90



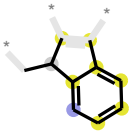
93



94



96

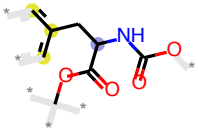


97

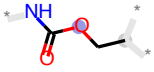
100

103

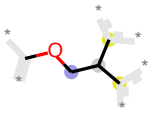
105



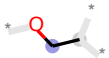
107



109

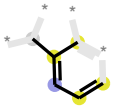


110

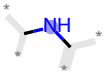


112

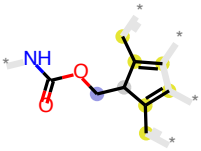
114



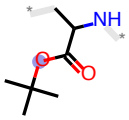
115



117



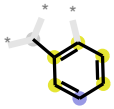
120



122

123

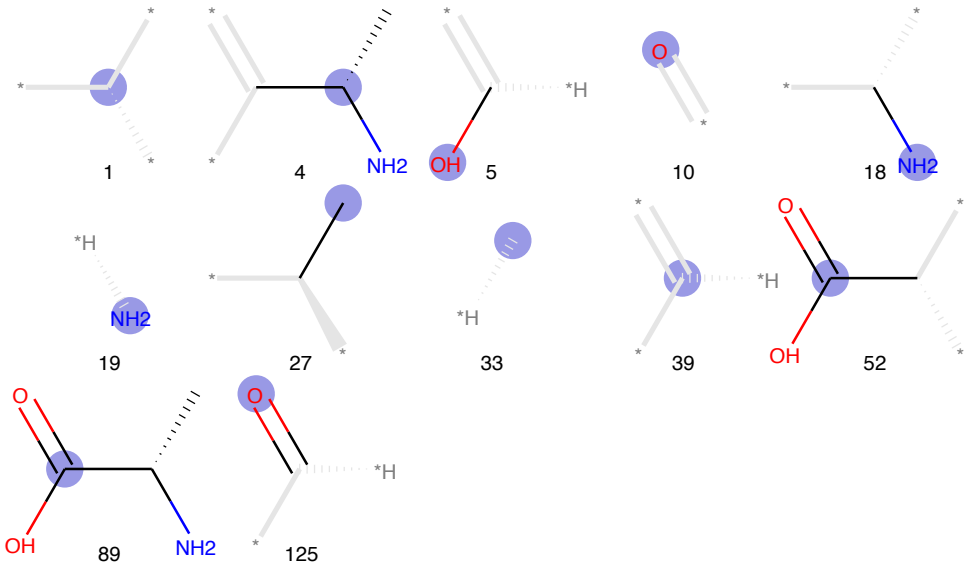
125



127

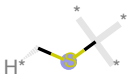
8.2 Substructures for pre-chain residues

Alanine



Cysteine

1



2

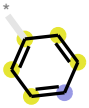
5

9

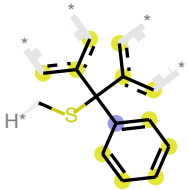
10

18

19

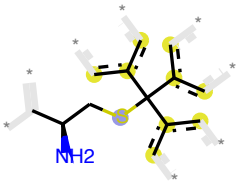


27

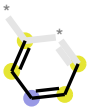


37

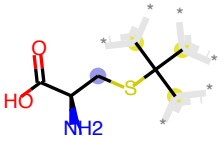
39



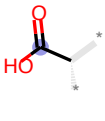
45



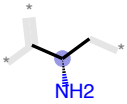
47



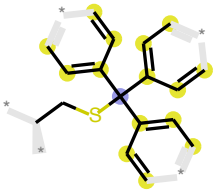
49



52

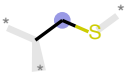


57



61

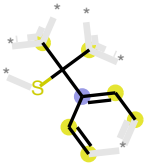
64



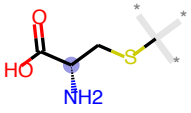
69

80

81

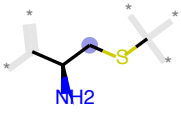


83

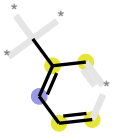


85

86

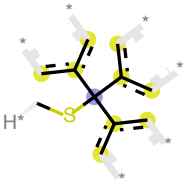


88



97

100



112

114

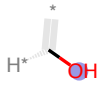
116

124

125

Aspartic acid

1



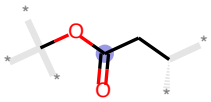
5

8

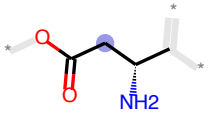
10

18

19

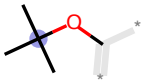


22



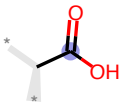
26

33



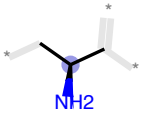
38

39

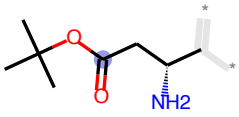


52

55

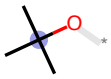


57



59

60



67

80

103

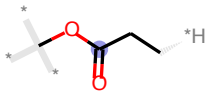
106

114

125

Glutamic acid

1



4

5

8

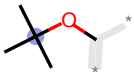
10

18

19

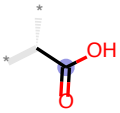
28

33



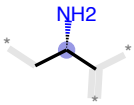
38

39

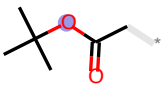


52

55

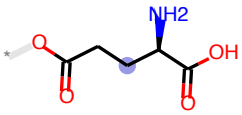


57



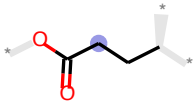
60

67

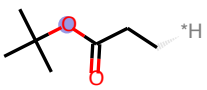


75

80



88

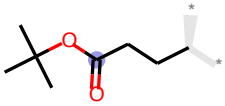


90

99

103

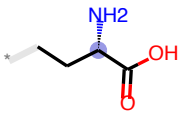
106



109

114

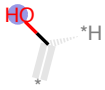
125



127

Phenylalanine

1

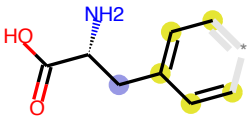


5

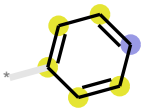
10

18

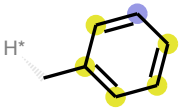
19



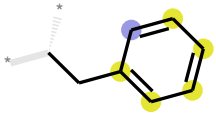
22



27

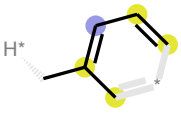


32



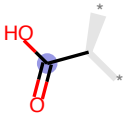
36

39

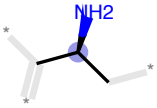


42

47

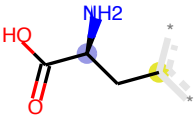


52

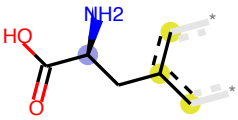


57

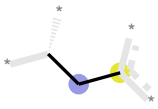
64



69



73

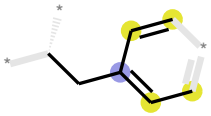


79

80

81

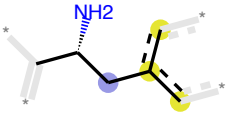
86



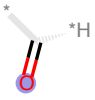
89

90

100

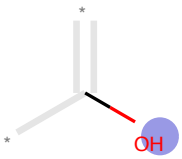


121

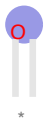


125

Glycine

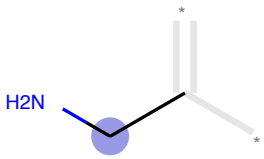


5

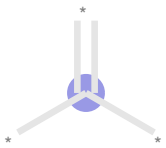


10

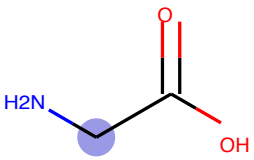
19



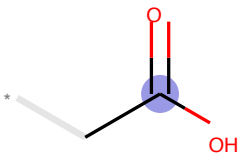
27



39



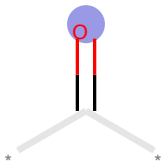
70



73

80

85



125

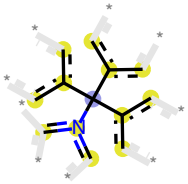
Histidine

1

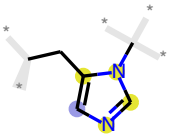
5

9

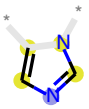
10



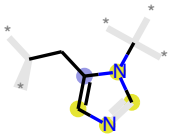
12



13



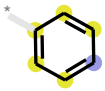
14



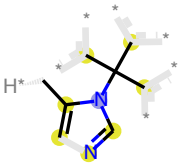
17

18

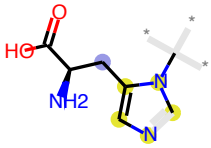
19



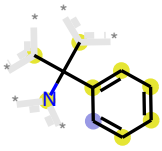
27



29



30

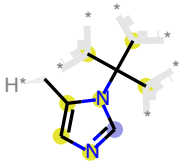


32

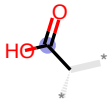
36

39

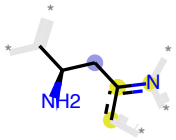
47



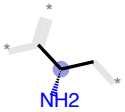
49



52



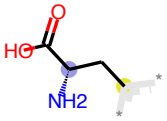
54



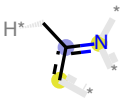
57

64

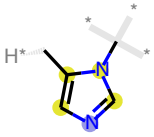
67



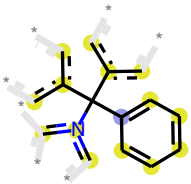
69



70



72



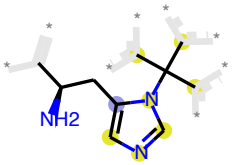
73

79

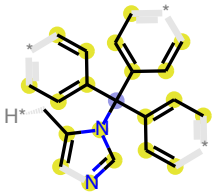
80

81

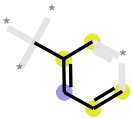
86



87

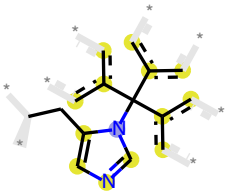


91

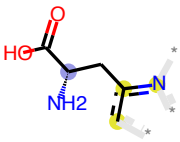


97

100

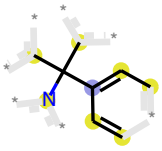


105



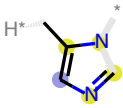
107

111



112

114



115

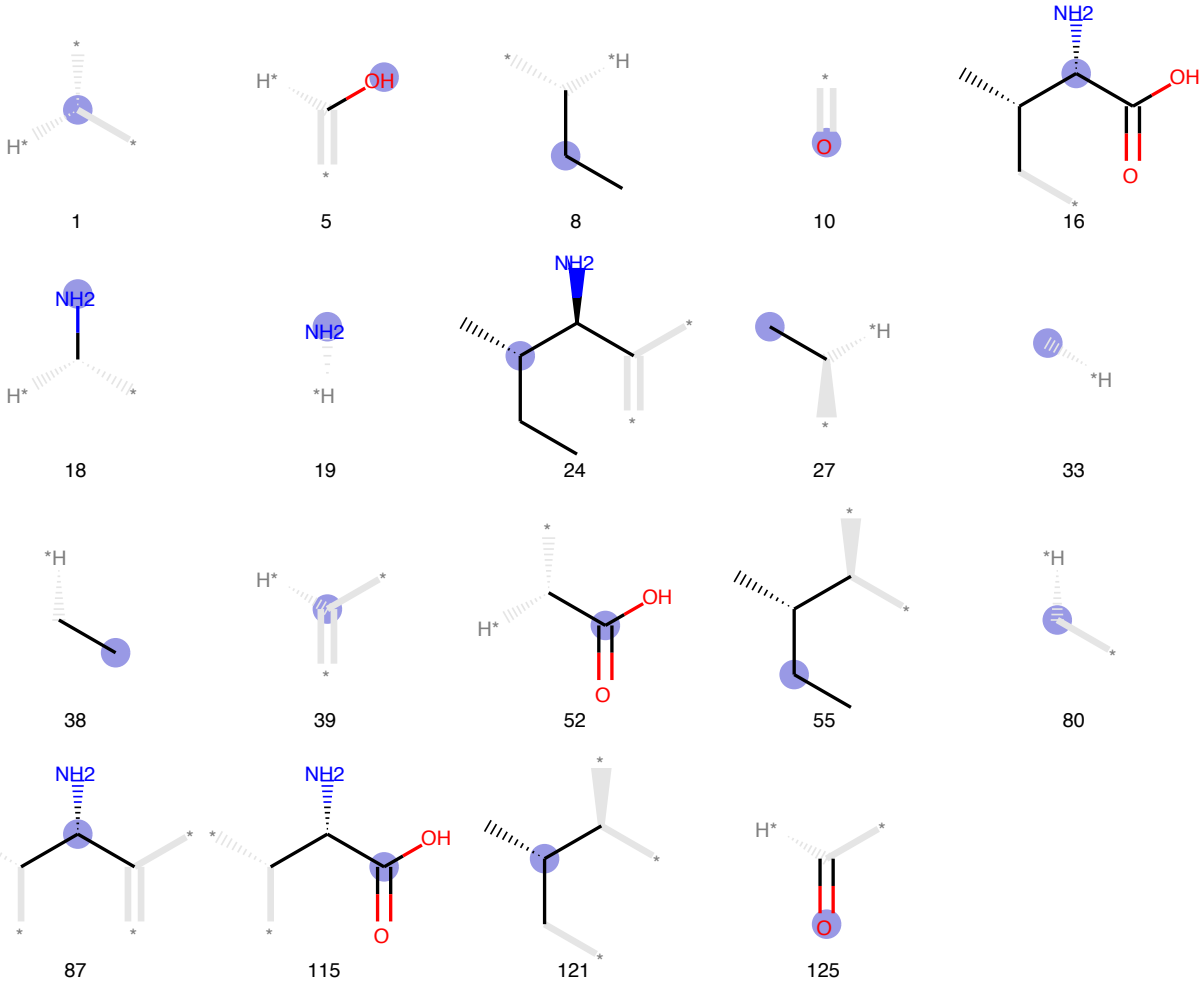
122

123

125

127

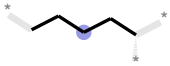
Isoleucine



Lysine

0

1

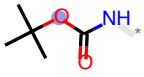


3

5

8

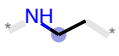
10



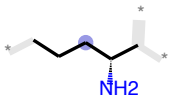
16

18

19



27

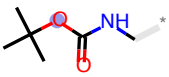


28

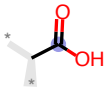
33

38

39

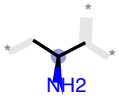


42



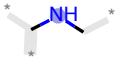
52

55



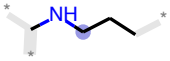
57

67

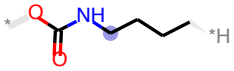


69

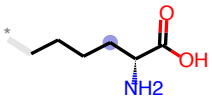
80



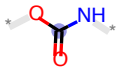
84



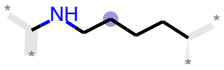
85



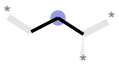
86



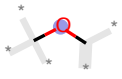
94



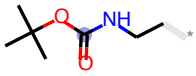
98



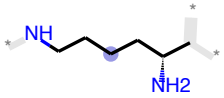
99



103



105

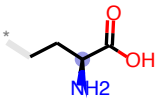


106

114

119

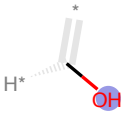
125



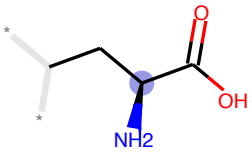
127

Leucine

1

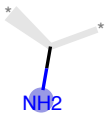


5



8

10

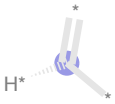


18

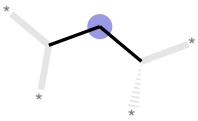
19

27

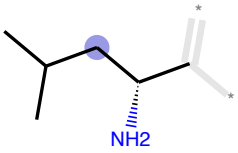
33



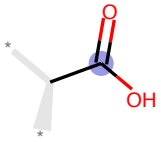
39



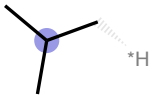
40



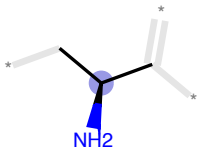
41



52

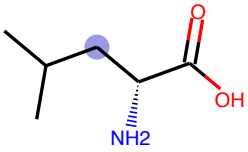


55

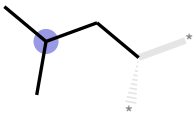


57

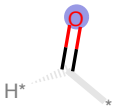
80



110



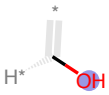
119



125

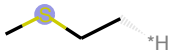
Methionine

1



5

10



11

18

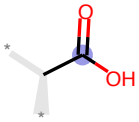
19

22

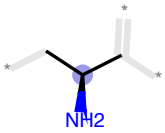
30

33

39



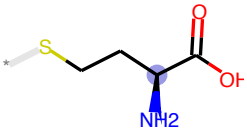
52



57

80

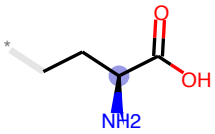
99



112

116

125

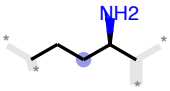


127

Asparagine

0

1

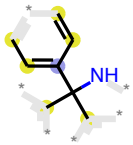


4

5

9

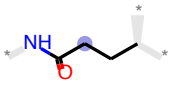
10



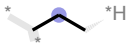
11

18

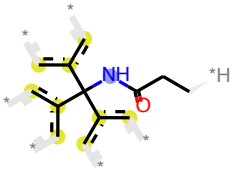
19



27

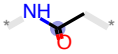


28



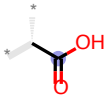
36

39

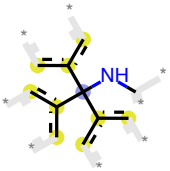


41

47



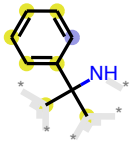
52



56

57

64

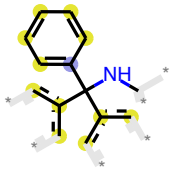


69

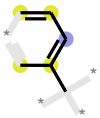
80

81

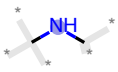
86



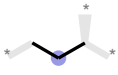
91



97

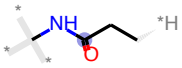


98

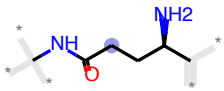


99

100

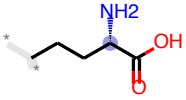


101



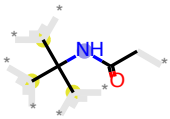
104

105

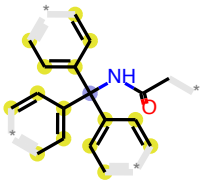


106

114

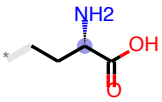


120



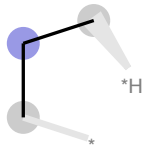
122

125

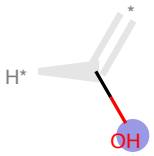


127

Proline

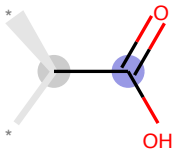


4



5

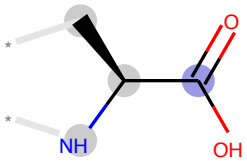
10



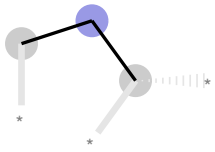
23

30

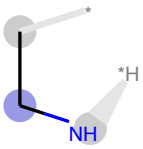
39



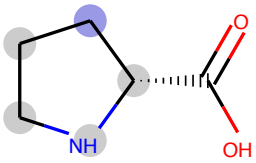
41



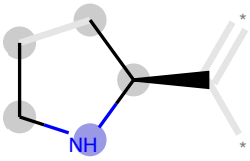
45



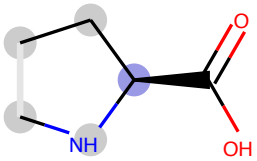
49



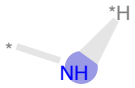
63



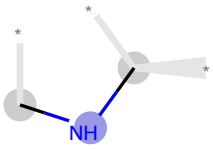
72



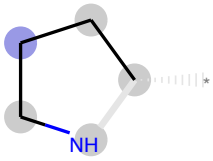
74



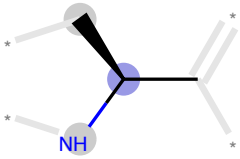
90



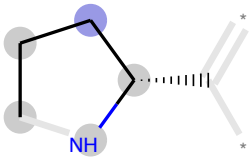
99



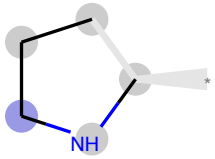
102



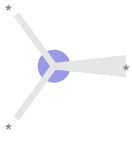
106



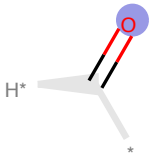
118



120



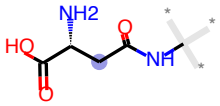
123



125

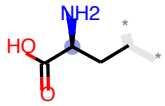
0

1



4

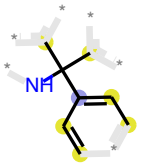
5



8

9

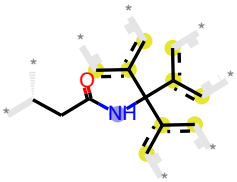
10



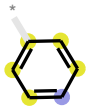
11

18

19

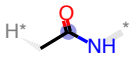


24



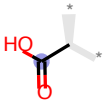
27

39

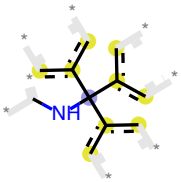


41

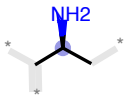
47



52

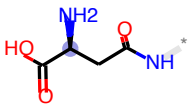


56

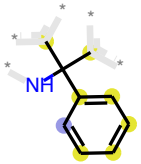


57

64



68

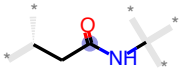


69

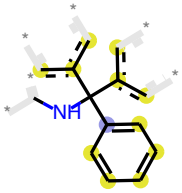
80

81

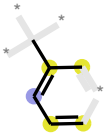
86



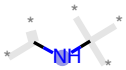
89



91



97

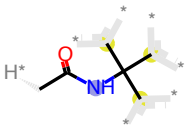


98

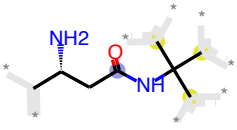
100

105

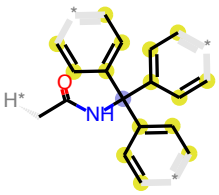
114



120



121



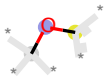
122

125

Arginine

0

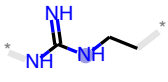
1



2

5

10



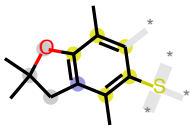
12

14

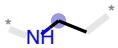
16

18

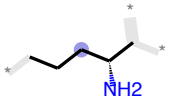
19



21

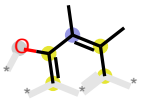


27



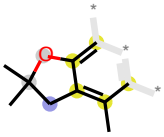
28

30



32

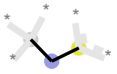
33



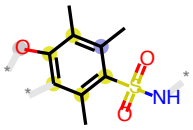
34

36

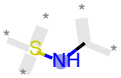
39



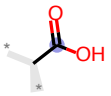
44



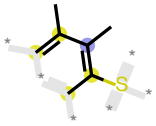
49



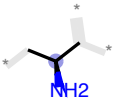
51



52

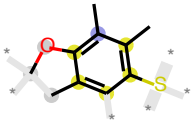


53

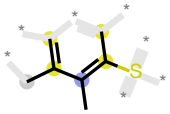


57

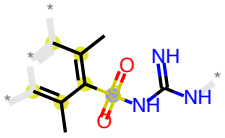
58



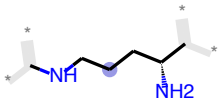
62



63

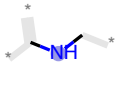


66

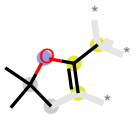


67

68

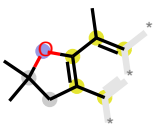


69

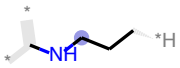


74

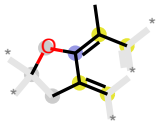
80



81



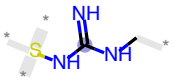
84



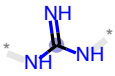
88

93

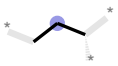
94



95

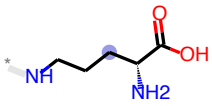


98



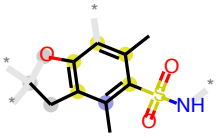
99

100



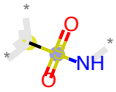
103

105



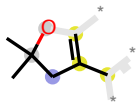
109

111



118

119

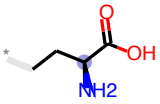


121

122

124

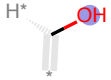
125



127

Serine

1



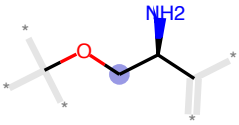
5

8

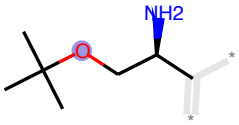
10

18

19



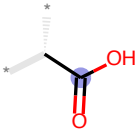
21



29

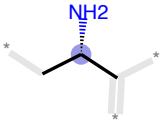
33

39

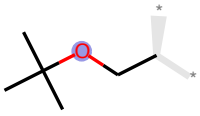


52

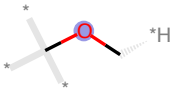
55



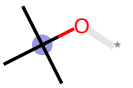
57



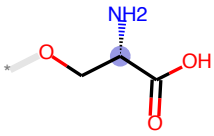
60



64

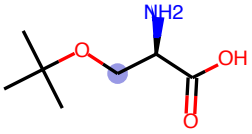


67



75

80



98

114

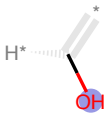
118

124

125

Threonine

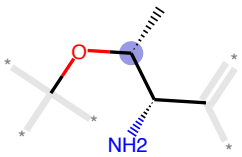
1



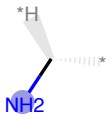
5

8

10

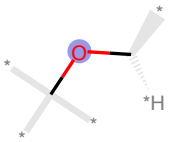


17

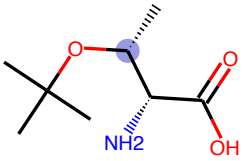


18

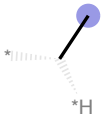
19



22



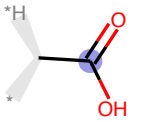
24



27

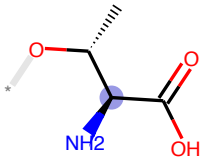
33

39

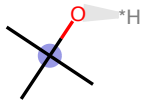


52

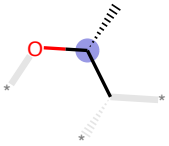
55



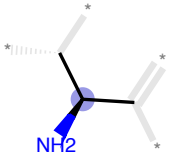
59



67

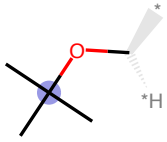


82



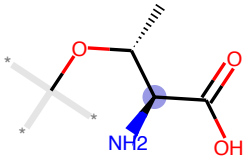
87

114



115

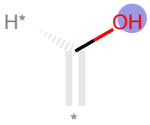
125



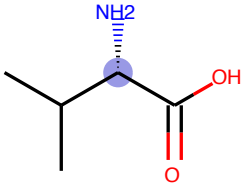
126

Valine

1

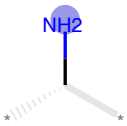


5



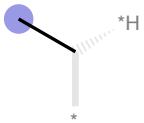
7

10



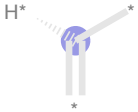
18

19

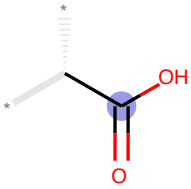


27

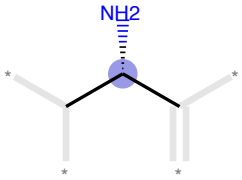
33



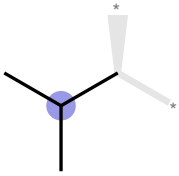
39



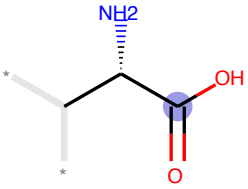
52



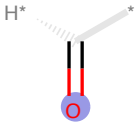
87



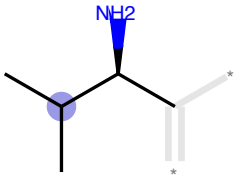
109



115



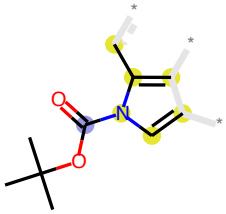
125



126

Tryptophan

1



3

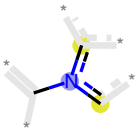
5

8

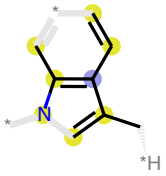
10

18

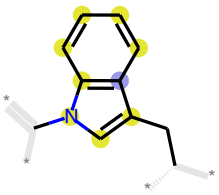
19



21

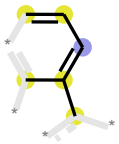


25



26

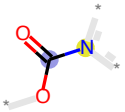
33



37

38

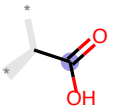
39



41

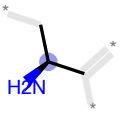
45

47



52

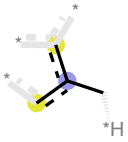
55



57

64

67



69

77

79

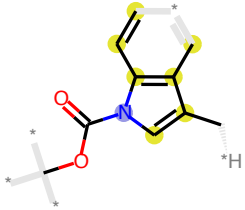
80

81

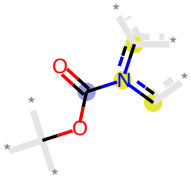
86

93

100

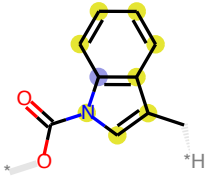


101

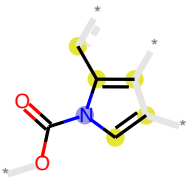


102

103



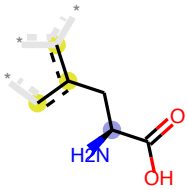
111



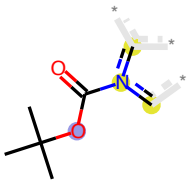
112

114

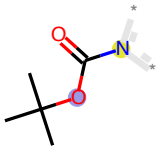
116



120



121



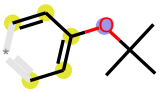
123

125

Tyrosine

1

3



4

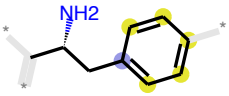
5

8

10

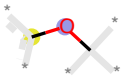
18

19



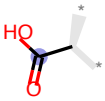
22

33

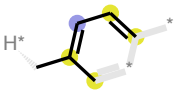


36

39

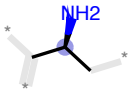


52

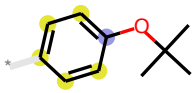


54

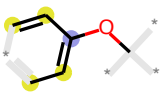
55



57



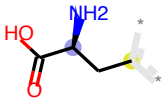
60



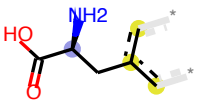
61

66

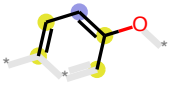
67



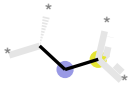
69



73



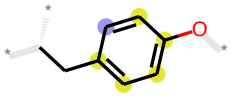
78



79

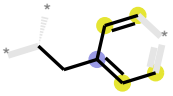
80

81



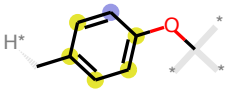
84

86



89

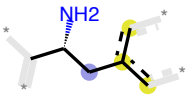
90



91

100

114

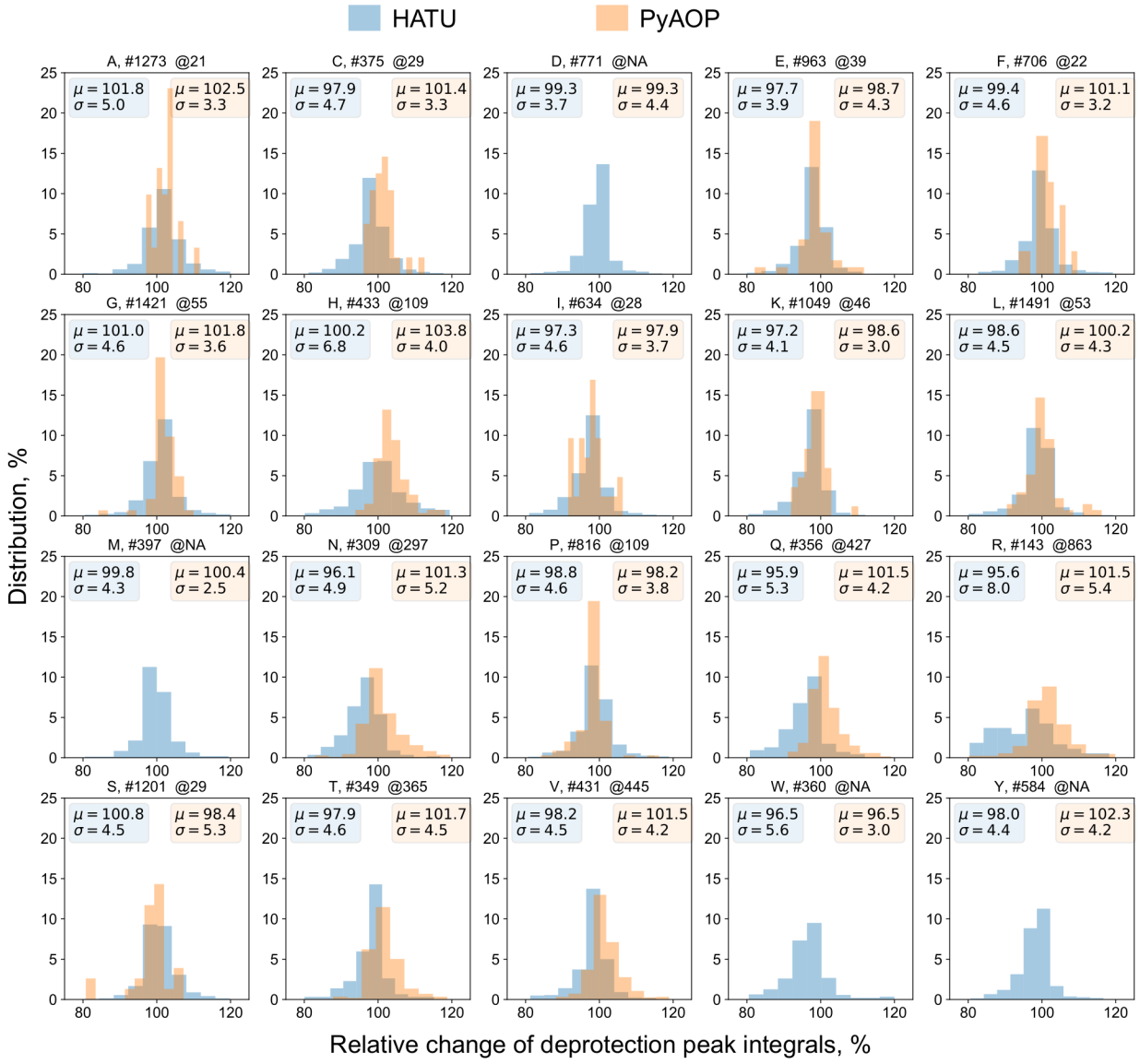


121

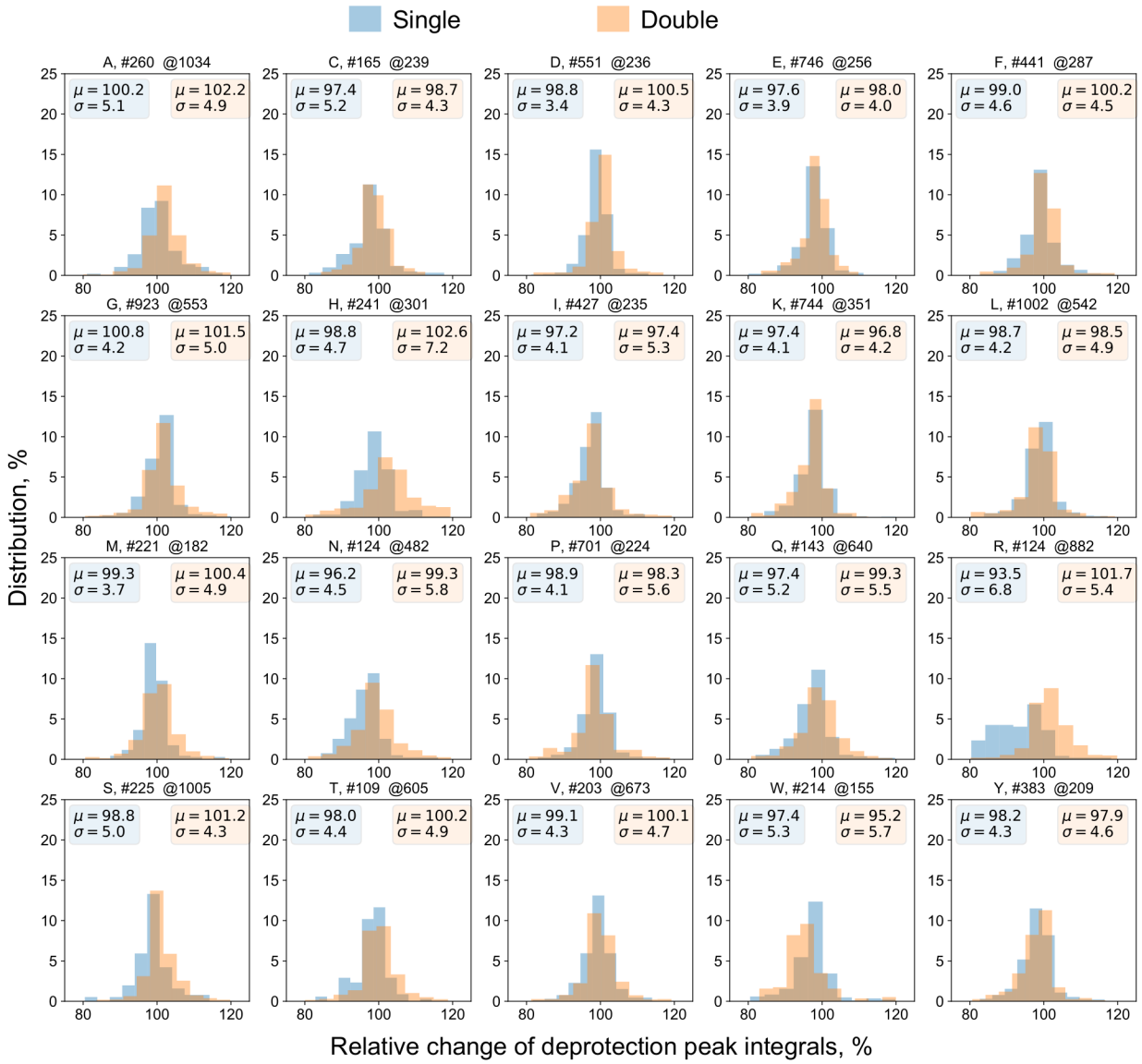
125

9 Appendix 2

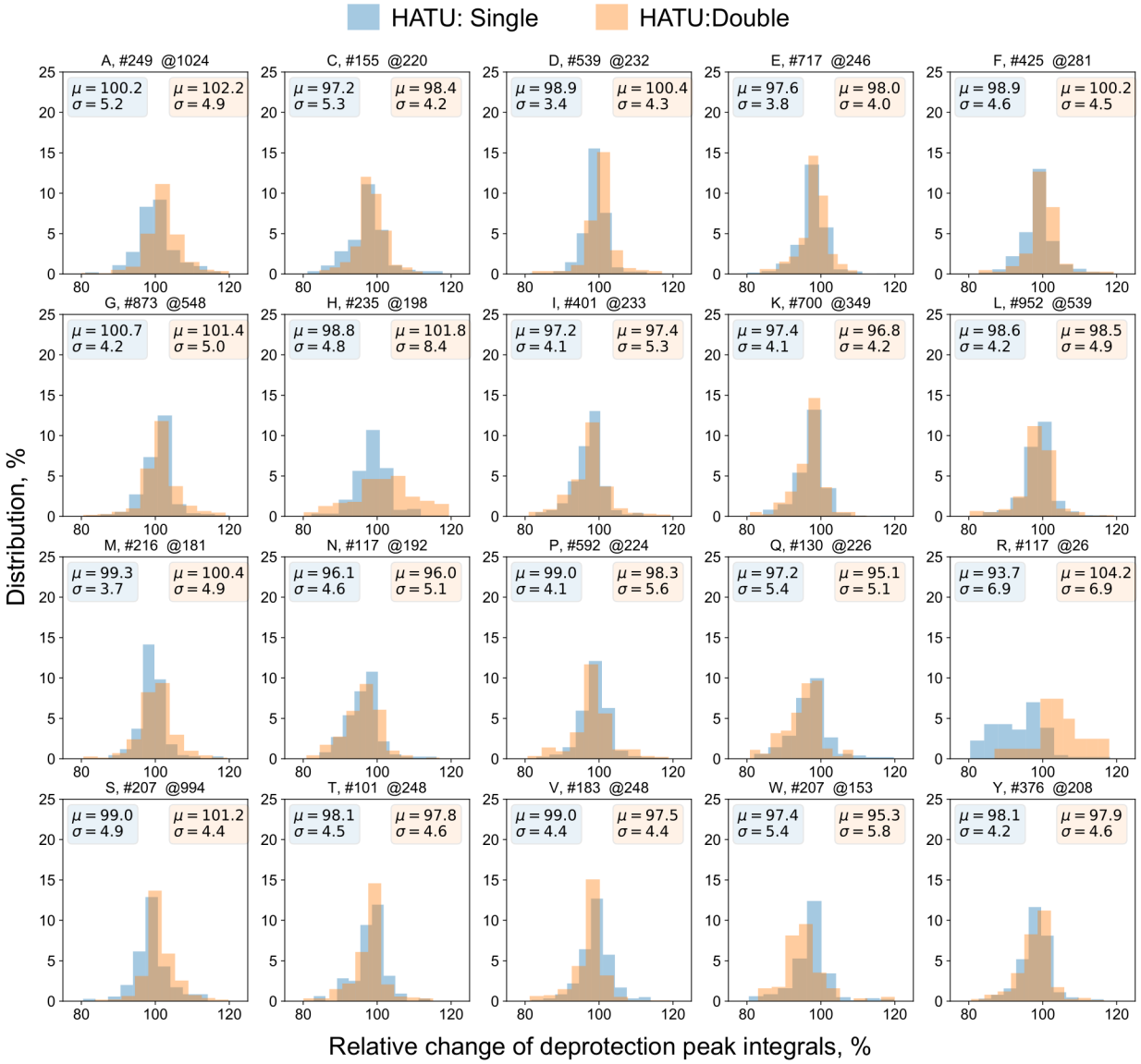
Histogram of relative change of deprotection peak integrals by amino acid, for different coupling agents – HATU (■) and PyAOP (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable). The mean and standard deviation for respective distributions are noted.



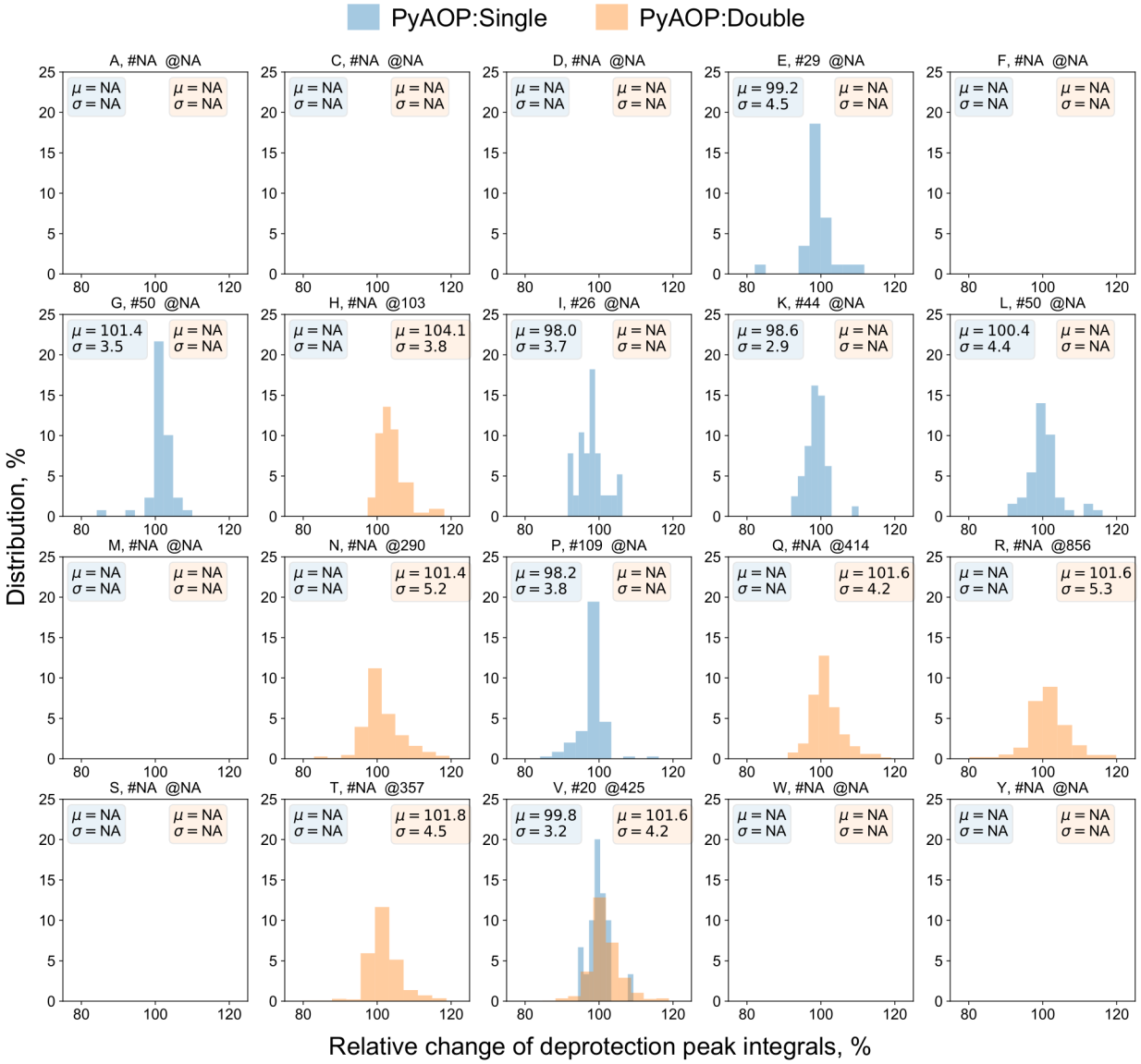
Histogram of relative change of deprotection peak integrals by amino acid, for different coupling strokes – Single (■) and Double (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable). The mean and standard deviation for respective distributions are noted.



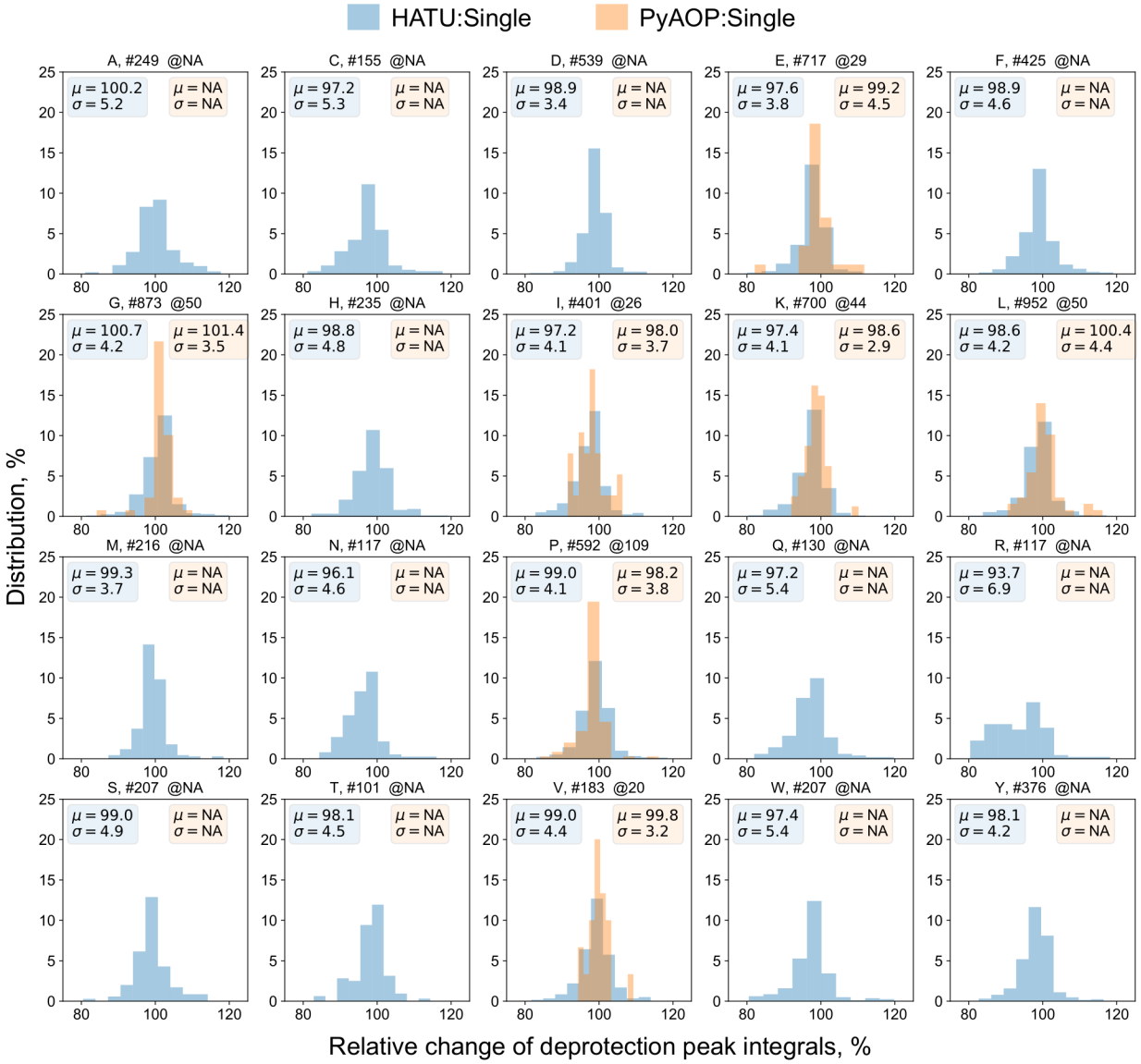
Histogram of relative change of deprotection peak integrals by amino acid, for different combinations of coupling agent:coupling strokes – HATU:Single (■) and HATU:Double (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable). The mean and standard deviation for respective distributions are noted.



Histogram of relative change of deprotection peak integrals by amino acid, for different combinations of coupling agent:coupling strokes – PyAOP:Single (■) and PyAOP:Double (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable). The mean and standard deviation for respective distributions are noted.).



Histogram of relative change of deprotection peak integrals by amino acid, for different combinations of coupling agent:coupling strokes – HATU:Single (■) and PyAOP:Single(■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable). The mean and standard deviation for respective distributions are noted.



Histogram of relative change of deprotection peak integrals by amino acid, for different combinations of coupling agent:coupling strokes – HATU:Double (■) and PyAOP:Double (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number of data points is less than 20 are not visualized and the number is not noted as NA (not applicable). The mean and standard deviation for respective distributions are noted.