doi.org/10.26434/chemrxiv.12649343.v1

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

Somesh Mohapatra, Nina Hartrampf, Mackenzie Poskus, Andrei Loas, Rafael Gomez-Bombarelli, Bradley L. Pentelute

Submitted date: 13/07/2020 • Posted date: 14/07/2020

Citation information: Mohapatra, Somesh; Hartrampf, Nina; Poskus, Mackenzie; Loas, Andrei; Gomez-Bombarelli, Rafael; Pentelute, Bradley L. (2020): Deep Learning for Prediction and Optimization of Fast-Flow Peptide Synthesis. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12649343.v1

Chemical synthesis of polypeptides involves stepwise formation of amide bonds on an immobilized solid support. The high yields required for efficient incorporation of each individual amino acid in the growing chain are often impacted by sequence-dependent events such as aggregation. Here we apply deep learning over ultraviolet-visible (UV-Vis) analytical data collected from 35,485 individual fluorenylmethyloxycarbonyl (Fmoc) deprotection reactions performed with an automated fast-flow peptide synthesizer. The integral, height and width of these time-resolved UV-Vis deprotection traces indirectly allow for analysis of the iterative amide coupling cycles on resin. The computational model maps structural representations of amino acids and peptide sequences to experimental synthesis parameters and predicts the outcome of deprotection reactions with less than 4% error. Our deep learning approach enables experimentally-aware computational design for prediction of Fmoc deprotection efficiency and minimization of aggregation events, building the foundation for real-time optimization of peptide synthesis in flow.

File list (2)

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]*

Affiliations:

[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]The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02142, 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

Abstract:
Chemical synthesis of polypeptides involves stepwise formation of amide bonds on an immobilized solid support. The high yields required for efficient incorporation of each individual amino acid in the growing chain are often impacted by sequence-dependent events such as aggregation. Here we apply deep learning over ultraviolet-visible (UV-Vis) analytical data collected from 35,485 individual fluorenylmethyloxycarbonyl (Fmoc) deprotection reactions performed with an automated fast-flow peptide synthesizer. The integral, height and width of these time-resolved UV-Vis deprotection traces indirectly allow for analysis of the iterative amide coupling cycles on resin. The computational model maps structural representations of amino acids and peptide sequences to experimental synthesis parameters and predicts the outcome of deprotection reactions with less than 4% error. Our deep learning approach enables experimentally-aware computational design for prediction of Fmoc deprotection efficiency and minimization of aggregation events, building the foundation for real-time optimization of peptide synthesis in flow.

Introduction:

Amide bonds play a central role in nature. They covalently link amino acids in the peptides and proteins involved in every aspect of life. In addition, amide bond formation is the most frequently used reaction in medicinal chemistry and its preponderance is still increasing.(*1*) It was used at least once in ~60% of the medicinal chemistry literature in 2014, and in ~7.2% of these reports amide bond formation occurred in the context of amino acid couplings in solid phase peptide synthesis (SPPS).(*1*) In SPPS, multiple iterations of amino acid couplings and deprotections on a solid support enable elongation of a polypeptide chain.(*2*) By contrast to recombinant expression, SPPS allows for the incorporation of a virtually unlimited number of non-canonical amino acids and site-directed mutations.(*3*) Synthetic peptides and proteins obtained with SPPS technology are therefore of great therapeutic interest, but low atom-economy and secondary events on resin, such as aggregation and aspartimide formation, limit their current application.(*4*, *5*) The availability or routine computational tools to predict and correct these events in real-time would be a major breakthrough in improving overall synthesis quality of polypeptides.

Method development and optimization of organic reactions is labor-intensive and requires multiple rounds of trial-and-error experimentation.(*6*) Flow chemistry offers the possibility to automate these processes and often improves reaction outcomes relative to batch methods due to increased heat and mass transfer. Automation of chemical reactions therefore leads to enhanced productivity and high reproducibility.(*7*, *8*) For example, a modular synthesis platform developed by Burke and co-workers allows for the rapid synthesis and purification of various small molecules using bifunctional *N*-methyliminodiacetic acid (MIDA) boronates as building blocks for Suzuki-Miyaura cross-couplings.(*9*, *10*) In addition to in-line purification, data collection from continuous flow systems is enabled by in-line analysis, which increases mechanistic understanding through real-time monitoring of intermediates and by-products in response to variation of synthesis parameters.(*11*) Building on similar concepts, Jamison and co-workers developed a compact, fully-integrated and easily reconfigurable, benchtop system that enables automated optimization of various chemical transformations using flow chemistry.(*6*) In addition, we recently demonstrated the advantages of automated fast-flow peptide synthesis (AFPS) over traditional SPPS techniques in terms of higher synthetic fidelity, increased length of the peptide chains accessible, and significant decrease in synthesis time.(*12*)

Advancements in computational methods allow for the investigation of large-scale problems and previously inaccessible correlations in organic reaction methodology. Improved algorithms can predict reactivity and plan retrosynthetic routes from data.(*13*–*16*) Furthermore, their combination with state-of-the-art automated experimental platforms can bring us closer to autonomous discovery. The Jensen and Jamison groups developed a robotic flow chemistry platform able to plan, execute and evaluate new reactions.(*17*) They demonstrated the capabilities of this setup by designing and conducting the synthesis of multiple drug-like molecules. Because training data on flow chemistry is scarce, this approach requires pre-processing batch synthesis data into equivalent flow parameters. To circumvent this issue and directly build upon batch chemistry-based literature, Cronin and co-workers developed the Chemputer, an automated synthesis platform that mimics batch synthesis.(*18*) Ada is another example of self-driving lab for accelerated development of thin-films, based on ChemOS,(*19*) a software package for autonomous discovery, and Phoenics,(*20*) a Bayesian optimization algorithm.(*21*) Additional efforts have utilized data-driven approaches to predict products and reaction types from reactants and reagents,(*13*) and optimize retrosynthetic routes using Monte Carlo tree search.(*22*) There have been attempts to optimize reaction conditions using reinforcement learning and machine learning.(*23*, *24*) Although these approaches are able to predict retrosynthesis routes and optimize the conditions of reactions one at a time, prediction and optimization of overall synthetic yield for arbitrary new reactions remains an open challenge.

Access to high-quality, interpretable and standardized datasets suitable for machine learning is a current bottleneck as the literature on chemical reactions is often unstructured, exists in multiple formats, sometime behind paywalls, and was collected on different reaction setups.(*25*) In addition, the published literature contains partially irreproducible data, which are difficult to identify a priori.(*26*) Learning based on data generated from automated experimental platforms could significantly improve predictions of synthesis outcomes, but these data sets are usually limited in size.

Here, we demonstrate that an in-line collected large set of high-quality peptide synthesis data can be leveraged to train effective deep learning approaches that predict reaction yield and in silico optimization of synthesis parameters. A better understanding of individual reactions on resin could further improve the synthetic process.(*27*) However, there are 400 possible binary couplings and 20^n such possible coupling steps for an n-amino acid polypeptide, considering only the canonical, proteinogenic amino acids. The growth of the peptide chain on resin is complicated by additional sequence-dependent events, such as aggregation.(*28–30*) Predictions of interactions that cause aggregation and strategies to prevent them are described in the literature,(*31–33*) but the molecular and structural factors affecting aggregation during synthesis on the solid support are not fully elucidated and therefore difficult to predict. Another layer of complexity is added by the incorporation of non-canonical amino acids or building blocks with uncommon protecting groups.

Automated fast-flow peptide synthesizer (AFPS) gives access to highly reproducible data. Peptide synthesis data was generated on a fully automated fast-flow peptide synthesizer (AFPS) developed in our laboratory which forms amide bonds orders of magnitude faster than commercial instruments (**Fig. 1A**).(*34*, *35*) With this machine, deprotection of fluorenylmethyloxycarbonyl (Fmoc) groups is generally quantitative and the resulting byproduct dibenzofulvene can be detected using an in-line UV-Vis detector (310 nm).(*36*, *37*) This data can be used to indirectly obtain information on the individual stepwise coupling cycles and the overall synthesis performance. In contrast to conventional peptide synthesizers, automated flow synthesis yields additional direct information on the Fmoc-deprotection steps by generating a time-dependent UV-Vis trace.(*38*, *39*) The integral and shape (width, height) of these signals can be used to identify mass transfer issues during deprotection, which are interpreted as aggregation on resin.

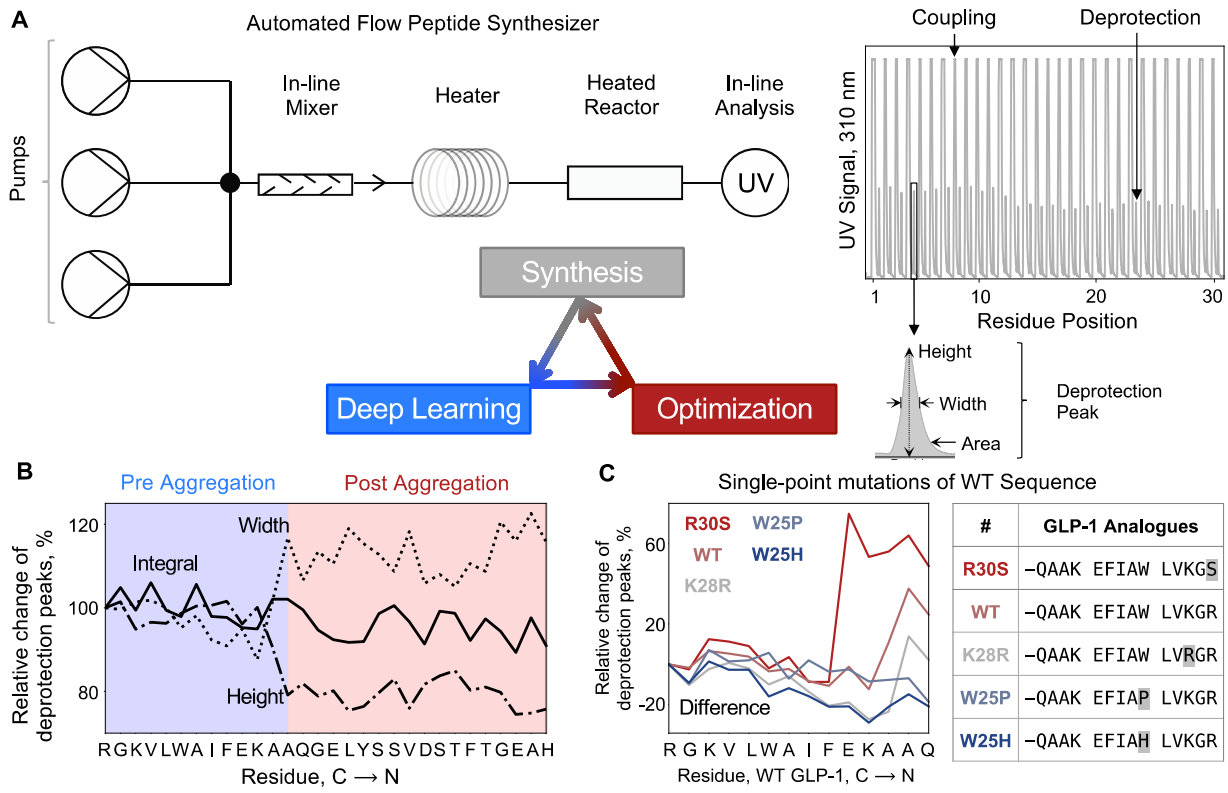


Figure 1. Deep learning enables prediction and optimization of fast-flow peptide synthesis. A. An automated fast-flow peptide synthesizer is used for the synthesis of peptides. Each synthesis run delivers UV-Vis traces for all coupling and deprotection chemical steps. **B.** Deep learning is done over parameters – integral, width, height and difference between width and height – calculated from the deprotection steps in the experimental data. The model predicts the relative change of deprotection peaks, as a proxy for synthesis success, and aggregation events, based on the difference between width and height. The difference is calculated by subtracting the percentage values of normalized height from normalized width. **C.** The model is used to predict relative change in deprotection peaks and aggregation for all single-point mutations of the wild-type sequence. Mutants predicted to be less aggregating and more aggregating than the wild-type sequence are experimentally synthesized and validated.

Over the past years we have systematically improved synthesis parameters and developed an amino acid-specific recipe.(*27*) First, we screened various solvents, synthesis temperatures, coupling and deprotection bases, coupling agents and flow rates. We then identified amino acids with low coupling efficiency and optimized coupling times and reagents. Using this approach, we defined a recipe which now allows for the routine synthesis of polypeptides of length corresponding to single domain proteins (up to 164 amino acids in length).(*27*) We envisaged that automated flow peptide synthesis could be improved even further if we had a better understanding of sequence-dependent events, e.g., aggregation, that occur during the process.

The dataset obtained from our optimization experiments contains 35,427 individual, highly reproducible deprotection steps. Each reaction step is defined by the pre-synthesized sequence on the resin (termed 'pre-chain'), the features of the incoming amino acid and a set of synthesis parameters. There are 17,459 unique reaction steps, after removing outliers and averaging over duplicates (**SI Fig. 1, SI Section 3.1**). The integral, height and width of deprotection traces were normalized to the first coupling step. Across all unique Fmoc-deprotection steps, the average relative integral was 89% and the reproducibility was within 10%. From the statistical analysis of this high-fidelity data, we identified particularly challenging binary coupling steps and looked for solutions to address them. The influence of sequence-specific interactions

on peptide synthesis cannot be addressed through human intuition alone due to the large combinatorial design space and overwhelming dataset size. In order to understand and predict how peptide sequence affects synthetic performance we turned to deep-learning algorithms (**Fig. 1B**).

Deep learning on high-quality synthesis data allows for prediction of UV-Vis deprotection traces.

Monomers in the pre-chain and incoming amino acids were represented using Extended-Connectivity Fingerprints (ECFP, **Fig. 2A**).(*40*) This topological representation encodes the molecular graph into a bit-vector of desired length where every feature represents one or more particular substructures. Common substructures such as the amide backbone, C-terminal carboxy groups, and N-terminal amines appear in most bit-vectors, while unique substructures in the side-chains distinguish the amino acid bit-vectors from one another (**SI Appendix 1**).

All amino acids were represented with explicit protecting groups, since these can influence their reactivity and physicochemical properties such as polarity. In the case of incoming amino acid, fingerprints were generated from molecules with Fmoc protecting groups. The pre-chain was featurized as a row matrix of ECFP bit-vectors with free amine groups (**Fig. 2B**). The peptide primary structure is thus captured by the sequence of fingerprints, and each monomer chemistry by the ECFP bit-vector.

A deep neural network model was trained over the peptide representation and the synthesis parameters to predict the integral, height and width of UV-Vis Fmoc deprotection traces normalized to the first coupling in the peptide synthesis. These variables quantify the success of each reaction step (**Fig. 2C**). The reactive structures are represented by the pre-chain row matrix and the incoming amino acid bit-vector. The synthesis parameters include categorical and numerical features: reactor temperature, flow rate and coupling-deprotection variables — coupling agent, pre-activation loop temperature, coupling and deprotection strokes. The model architecture first processes individual variables, then concatenates the outputs of the individual representation-learning layers, followed by fully-connected layers. This allows the model to process and transform every variable in an optimal way before combining them.

The model was trained and validated on a random 70:30 split of the available data. For integral, height and width, the prediction errors on held out test data are under 0.1 RMSE (6% relative to the range of the training data, **Fig. 2D, SI Fig. 2, SI Table 1**). For GLP-1 and other test sequences held out from the training dataset, the UV-Vis traces predicted using the deep learning model match the experimentally obtained traces within said uncertainty (**Fig. 2E**, **SI Fig. 2**).

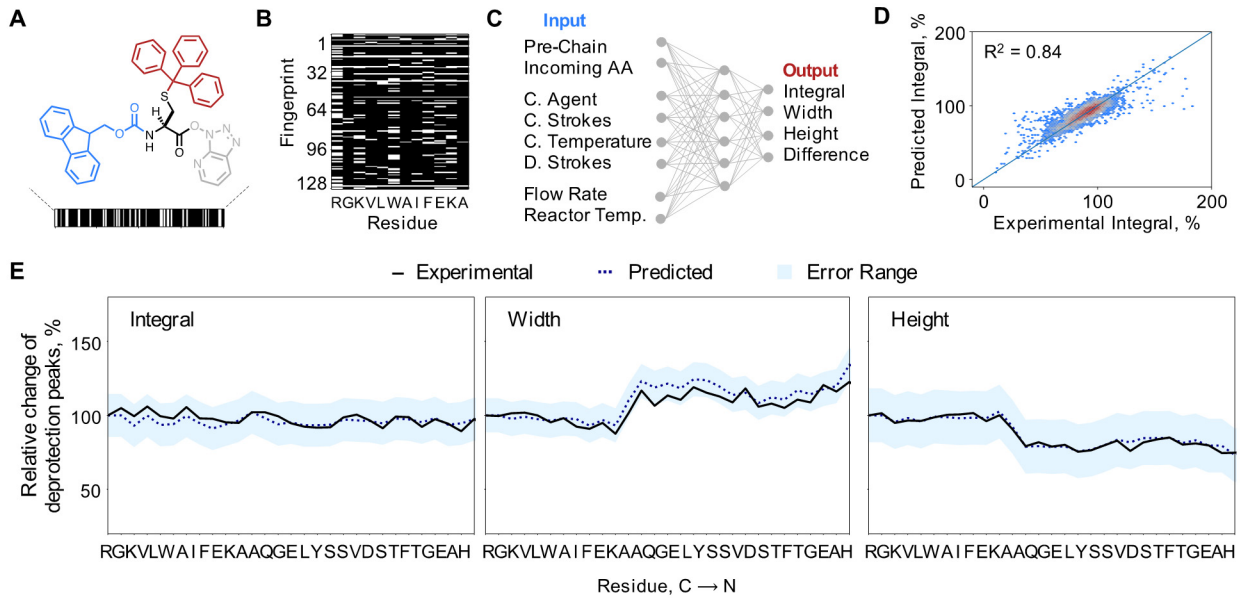


Figure 2. Deep learning predicts near-accurate UV-Vis Fmoc deprotection traces. A. Amino acids are represented using topological fingerprints. Fmoc- and side-chain protected representations are used for incoming amino acids, while amino acids in the pre-chain are represented with only side-chain protection. Amino acid = black, Fmoc = blue, active ester = grey, side-chain protecting group = red. **B.** The sequence on the resin is represented as a matrix of side-chain protected amino acid fingerprints. The order of amino acids in the matrix is same as the order in the sequence. **C.** Schematic of the machine learning model shows the multiple input and output variables. In the input, pre-chain and incoming amino acid representations featurize the chemistry of the reaction, while other variables featurize the synthesis parameters – coupling agent, number of coupling strokes, temperature of coupling, number of deprotection strokes, flow rate and temperature of reactor. In the output, integral of the Fmoc deprotection bands, and their height, width and difference are used to train the model. The model was trained on 70% of the dataset and its performance was evaluated on the remaining 30% of the dataset. **D.** The model predicts the integral for a particular reaction step with error under 13% of the data range on the validation dataset. **E.** Integral, height and width obtained from the model and experimental UV-Vis deprotection traces are overlaid for GLP-1 synthesis. The predictions from the model match the experimental values within the error range. GLP-1 was not part of the training dataset.

Deep learning predicts and enables interpretation of aggregation.

We predicted sequence-dependent aggregation using our model. The analysis of previously collected experimental data (*27*) suggests that certain sequence-dependent events, which are commonly defined as aggregation, result in poor synthetic outcome. These are characterized with mass transfer issues and slow reaction kinetics that are reflected in flattened, wider UV-Vis deprotection peaks. We use the difference of normalized width minus normalized height (W-H) to quantify such events, and define aggregation to have occurred when this difference is greater than 0.2 for a reaction step. We used the model trained above to predict W-H difference directly. The model was able to predict W-H difference on held out data with a RMSE of 0.13 (5.4% relative to the data range) (**SI Fig. 2C**) which allows to identify aggregation events. For GLP-1, which was not a part of the training dataset, the model is accurately able to identify the aggregating step, i.e., the addition of Ala18 (A18) (**Fig. 3A**).

In order to interpret the decision-making process of the neural network, we trained a minimal model. (*41*) This model was limited to pre-chain and incoming amino acid as input and difference between normalized width and height as output. By taking the normalized gradient of the neural-network predictions with respect to each bit-vector index of the input matrix, it is possible to quantify the contribution of the particular index towards aggregation. Representing these values as a heatmap allows visualization of the

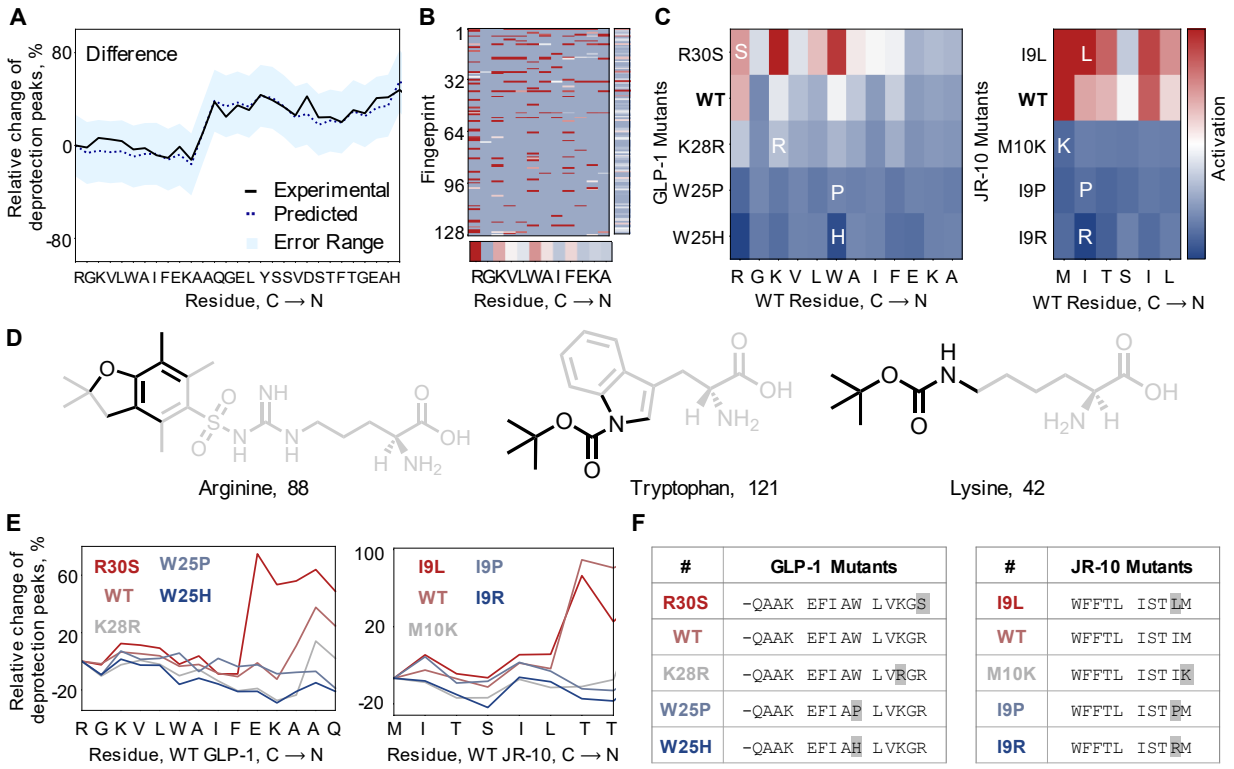
decision-making process of the model, and enables identification of features in the input representation which are responsible for aggregation.

We visualized the gradient activation map at the onset of aggregation for GLP-1 (**Fig. 3B**). The substructures by amino acid are ranked from the ones contributing most (red) to least (blue) towards aggregation. Averaged over the fingerprint indices, the model predicts Arg30 (R30) is the amino acid contributing the most to aggregation, followed by Trp25 (W25) and Lys28 (K28) (**Fig. 3C**). Noteworthy, the amino acids that impacted aggregation the most were far removed from the point of aggregation. Bulkier side-chain protecting groups such as the aromatic moieties in arginine (Arg) and tryptophan (Trp), and the *tert*-butyl protecting group in lysine (Lys) are the most activated substructures by amino acid respectively (**Fig. 3D**). Substructures common to all amino acids are always present in the fingerprints and were excluded from the substructure activation analysis.

To gain further insight on how the model learns aggregation, we interrogated the predictions of the aggregation model using a reference dataset of 8,441 natural proteins with 50 amino acids or fewer from the Protein Data Bank (accessed on April 17, 2020).(*42*) Similar trends were obtained in the activation analysis of aggregation (**SI Fig. 7**). 45% of the sequences were predicted to be aggregating. Amino acids in the pre-chain with aryl groups and bulkier side-chain protecting groups were found to be most activated for aggregation (**SI Fig. 10**, **SI Table 4**). On average, amino acids closest to the C-terminus are predicted to contribute the most towards aggregation (**SI Fig. 11**). The relative contribution from subsequent amino acids decreases the further their position is in the chain. The results delivered by our model suggest that aromatic and bulky side-chain protecting groups are a main pre-chain structural determinant of aggregation.

Deep learning model allows for sequence optimization of 'difficult peptides' using single-point mutations.

Single-site mutagenesis coupled with interpretation of gradient activation maps enable optimization of synthesis performance (**Fig. 3C**). All possible single-point mutants of wild-type GLP-1 and JR-10 were computationally enumerated and ranked by the aggregation model. The selection of least aggregating sequences was based on predicted aggregation and gradient activation maps. We observed that in most cases the mutations of amino acids which were most activated for aggregation (**Fig. 3C**) led to a decrease in the predicted aggregation.



#	GLP-1 Mutants
R30S	-QAAK EFIAW LVKGS
WT	-QAAK EFIAW LVKGR
K28R	-QAAK EFIAW LVRGR
W25P	-QAAK EFIAP LVKGR
W25H	-QAAK EFIAH LVKGR

#	JR-10 Mutants
I9L	WFFTL ISTLM
WT	WFFTL ISTIM
M10K	WFFTL ISTIK
I9P	WFFTL ISTPM
I9R	WFFTL ISTRM

Figure 3. The deep learning model predicts, interprets and optimizes aggregation. A. Predicted difference (width-height) is overlaid on the calculated difference from the experimentally obtained UV-Vis deprotection trace for GLP-1. The predicted difference is within the error for the experimentally observed difference. Aggregation is defined as the step where the difference between width and height is greater than 20%. **B.** Positive activation gradient map for GLP-1 pre-chain prior to the addition of third Ala (A18). The mean activation values for individual amino acids and bit-vectors are shown along respective axes. **C.** Positive activation gradient maps averaged over fingerprint indices for GLP-1 and JR-10 mutants show a sharp decrease in aggregation from the negative control (GLP-1: R30S, JR-10: I2L) to the wild-type and the other mutants. The pre-chains considered in the analysis are for the known aggregating regions in GLP-1 (addition of third Ala, A18) and JR-10 (addition of second Thr, T4). The most activated amino acids are Arg, Trp and Lys in WT GLP-1, and Met and Ile in WT JR-10. **D.** Most activated substructures by amino acid for GLP-1 are shown. Amino acids with aryl groups and bulkier side-chain protecting groups are found to be most activated. The analysis excluded substructures in the amino acid scaffold, both amide backbone and side chain native to the respective amino acid. The red dot is the node atom and the black bonds/atoms represent the topological exploration of n-nearest neighbors. **E.** Calculated difference from the experimental synthesis run for predicted sequence analogues of WT GLP-1 and WT JR-10. The analogues are predicted single point mutations of the sequence – K28R, W25P and W25H for GLP-1, and M10K, I9P and I9R for JR-10. The predicted negative controls are R30S for GLP-1 and I9L for JR-10. The predicted sequence analogues, except negative controls, are less aggregating at the respective step. Negative control for GLP-1 is more aggregating than GLP-1 itself. Negative control for JR-10 is less aggregating than JR-10, but more aggregating than the other analogues. **F.** Predicted GLP-1 and JR-10 mutants which were experimentally validated are listed. All mutants predicted using the model contain the mutation before the aggregating step, i.e., addition of third Ala for GLP-1, and addition of second Thr for JR-10. The in silico generation of mutants had no such constraints.

From the list of mutants, we selected four sequences predicted to be less aggregating and one sequence predicted to be more aggregating than the wild type sequence to evaluate our predictions experimentally (**Fig. 3E, F**). The experimental traces for the difference between normalized width and height for the mutants, including the negative control, matched the predictions of the model within 5% error (RMSE: 0.13). This outcome validates the accuracy of the model in minimizing aggregation and its robustness in predicting negative controls.

The model was trained on a representation that is transferable across chemical structures and we therefore determined if it would be able to predict synthesis outcome for unseen building blocks. We therefore synthesized GLP-1 with backbone-modified glycine and pseudoprolines, both types of building blocks are commonly used to avoid aggregation (**SI Table 3**). For the pseudoproline building blocks Fmoc-Ser(*t*-Bu)-Ser($\Psi^{Me,Me}$pro)-OH and Fmoc-Phe-Thr($\Psi^{Me,Me}$pro)-OH, the synthesis outcome was predicted with high accuracy, whereas prediction for Fmoc-(DMB)Gly-OH building blocks was less accurate. These experiments show the potential but also the limitations of the model, as training on more diverse building blocks will likely improve the ability to predict synthesis outcome for completely new building blocks in the future.

Statistical analysis of AFPS and PDB data sets

Statistical analysis over the entire AFPS dataset can inform future optimization of fast-flow peptide synthesis (**Fig. 4A-C**). When we compared different synthesis parameters for all amino acid couplings combined, we noticed that PyAOP shows improved synthesis outcomes when compared to the related coupling agent HATU. In addition, extended coupling times also had a positive effect on the synthesis. The overall differences for the coupling parameters are small, but these minor effects add up to have a potentially major detrimental impact in the synthesis of long peptides, where >99% coupling efficiency per incorporated amino acid is crucial.

Amino acids coupled under identical coupling conditions (single coupling with HATU) show diverse histogram profiles for their relative change in deprotection peak integrals (**SI Fig. 3**). Some residues, such as glycine, leucine and lysine show narrow distributions around 100%, whereas alanine, cysteine, histidine, asparagine, glutamine, arginine, serine, valine, tryptophan and threonine show broader distributions. The latter set of residues in comparison to the former set is more prone to reduction in deprotection yield, and are generally responsible for the overall decrease in synthesis quality. In our optimized recipe file, all of these residues —except for tryptophan— are already coupled under modified conditions (**SI Fig. 4**). To identify additional areas for optimization, we analyzed average coupling efficiencies for our optimized synthesis recipe (**Fig. 4D, SI Fig. 5**). It was found that all amino acids couple with high yields, however, tryptophan, cysteine, isoleucine and lysine present opportunities for improvement.

In addition, we found that aggregation is likely to occur at any position of the peptide chain more than 4 residues from the C-terminus, with an increased probability around positions 8 to 15 from the C-terminus for both experimental AFPS and predicted PDB data sets (**SI Fig. 6, 8**). For this analysis we compared all aggregating peptide sequences >20, >25 and >30 amino acids in length to obtain statistical information. In addition, we also validated that aggregation is generally position-independent, except for the very first amino acids that are coupled to the solid support (**Fig. 4E**). Further, the relative distributions of amino acids in non-aggregating sequences and pre-chains of aggregating sequences were found to be similar (**SI Fig. 9**). We therefore conclude that amide bond formation in flow is amino acid- and sequence-dependent but generally independent of the position of specific amino acids in the peptide.

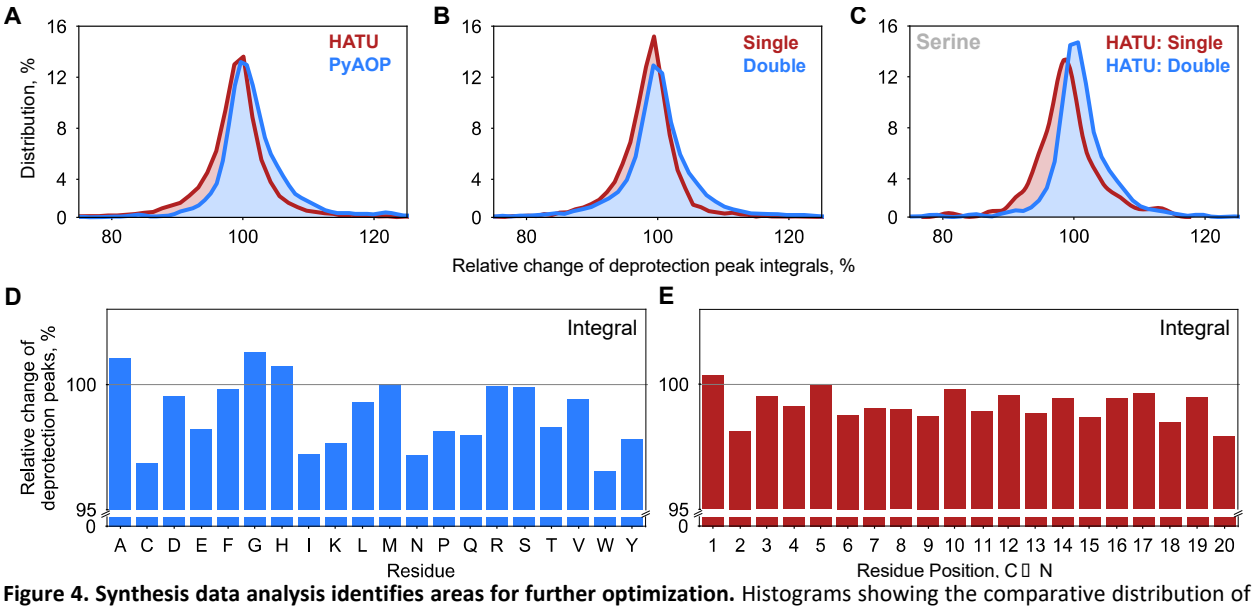


Figure 4. Synthesis data analysis identifies areas for further optimization. Histograms showing the comparative distribution of the relative change of deprotection peak integrals filtered across the entire reaction step dataset by **A.** coupling agent (HATU, PyAOP), **B.** coupling strokes (Single, Double) and **C.** coupling of serine from HATU subset with single coupling stroke to HATU subset with double coupling strokes. The mean value and the distribution as a whole move towards the ideal relative change of 100% in all the latter cases. The integrals were normalized to the integral of the preceding reaction step. An integral greater than 100% only indicates that the present reaction was better than the preceding step. **D.** Mean values of amino acid-specific deprotection peak integrals are shown. The integrals are normalized relative to the previous deprotection peak integrals. **E.** Mean values of deprotection peak integrals at different positions are shown. Residue position pre-aggregation is same as the position of the amino acid in the synthesis step. Residue position post-aggregation is the nth synthesis step after the aggregating step. The analysis is based on the optimized recipe for each individual amino acid, except Trp which needs to be optimized further.

Discussion

Deep learning on an automatically collected analytical data set from an AFPS setup can be used to predict peptide synthesis and sequence-specific events. Predicting sequence-dependent SPPS events is crucial for developing more efficient synthesis protocols. Here, we make a first step towards this goal by using analytical data from 35,427 individual, highly reproducible deprotection steps. Our model is able to predict the synthesis outcome for sequences which are not part of the training dataset. In addition, the sequences of aggregation-prone peptides were optimized for minimum aggregation using deep learning. As a first demonstration, we analyzed the synthesis of GLP-1 and JR-10. We predicted single-point mutations and experimentally validated improved synthesis outcome as a result of reduced aggregation. In the future, we intend to extend this to optimize synthetic accessibility and functionality together.

Computational analysis and interpretable deep learning can be used to extract non-obvious or previously hidden information from a large and complex data set. The general effect of changing key parameters in the recipe (e.g., coupling agent, coupling strokes, temperature) was obtained from statistical analysis of the entire data set and areas for additional improvement were identified. Regions prone to aggregation, which are the source of many deleterious side-reactions, were predicted with high confidence. Statistical analysis of the experimental AFPS data set and predicted PDB data set furthers the hypothesis that aggregation occurs with increased probability around the tenth position from the C-terminus,(*28*) although we also found aggregation at every other position of the protected peptide chain. We determined that aggregation does not depend on the position of specific amino acids in the sequence. We had already observed previously that the onset of aggregation can be shifted by increasing the synthesis temperature,(*27*) and here we also demonstrate how a single point mutation far from the actual predicted location of aggregation onset can obviate aggregation completely.

Intrigued by these results we strived to decode main contributors to aggregation by understanding how the model predicts these events. Using gradient activation on the deep learning model we determined that residue-specific “activators for aggregation” are often found at a location in the peptide chain far from the actual point of aggregation and close to the C-terminus of the peptide. The latter observation may be a consequence of SPPS proceeding in all cases from the peptide C-terminus. Interrogation of the activation maps revealed sequence-specific amino acids or substructures thereof that are most likely to cause aggregation. We found that aromatic, hydrophobic side-chains and protecting groups such as 2,2,4,6,7-pentamethyldihydrobenzofuran-5-sulfonyl (Pbf) and trityl (Trt) increased the probability for aggregation and Arg(Pbf), Trp(Boc), His(Trt), Asn(Trt) and Cys(Trt) were the main contributors (in decreasing order of relative contributions). This analysis is in line with reports in the literature stating that hydrophobic amino acids lead to aggregation.(*28*, *31*) However, we found aryl-containing residues and protecting groups to be more activating than *t*-Bu groups or aliphatic amino acids.

The tools we developed here are valuable for de novo computational design and optimization of peptide sequences, e.g., for personalized medicine. Artificial peptide and protein sequences are designed de novo to address challenges in medicine and nanotechnology.(*43*) Most of these biopolymers, however, are currently produced by recombinant methods. AFPS can significantly expedite and improve the synthesis quality of these structures, as already demonstrated in the context of tumor neoantigen peptides for personalized immunotherapy and cell-penetrating peptides.(*44–46*) In the event of an aggregating sequence, we demonstrated that single-point mutations can avoid aggregation during synthesis. Introducing point mutations into peptide and protein sequences is common practice in biology, often to interrogate function of a specific amino acid. Although it needs to be evaluated if the mutated sequences retain their biological function, we demonstrate how this approach can be valuable for improving the quality of peptide synthesis. In future developments, homology search can be integrated in the optimization process of bioactive peptide and protein chains to inform on mutational tolerance of the sequence.

This method demonstrates how deep learning can be used to predict and optimize chemical reactions using automated flow synthesis platforms. The model framework is agnostic of the experimental instrumentation and can be used in principle for any flow chemistry reaction setup with capability for in-line analysis. For polymer addition reactions, such as the synthesis of polyglycans or antisense oligonucleotides, the pre-chain and incoming monomer may be based on the current featurization framework with appropriate synthesis parameters, and trained on in-line monitoring parameters such as those obtained using various analytical methods. The model's predictive power is intrinsically linked to the availability of reproducible, standardized high-quality synthesis data for training. As our data set continues to grow with every biopolymer that is synthesized on our AFPS systems, we intend to expand the applicability of our model to additional reactions and building blocks, e.g., non-canonical amino acids or backbone modifications, as already demonstrated for some new building blocks. In the future, we hope to make the transition from an amino acid-based recipe to a sequence-dependent recipe wherein each amino acid is coupled according to its nature and position in the peptide chain. We envision this approach will ultimately lead to real-time in-line suggestion of synthesis parameters – a principle envisioned by Erickson as early as 1981.(*36*)

References

1. D. G. Brown, J. Boström, Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **59,** 4443–4458 (2016).

2. R. B. Merrifield, Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. *Journal of the American Chemical Society*. **85,** 2149–2154 (1963).

3. S. Bondalapati, M. Jbara, A. Brik, Expanding the chemical toolbox for the synthesis of large and uniquely modified proteins. *Nature Chemistry*. **8,** 407–418 (2016).

4. S. B. H. Kent, Total chemical synthesis of proteins. *Chem. Soc. Rev.* **38,** 338–351 (2009).

5. A. A. Zompra, A. S. Galanis, O. Werbitzky, F. Albericio, Manufacturing peptides as active pharmaceutical ingredients. *Future medicinal chemistry*. **1,** 361–377 (2009).

6. A.-C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen, T. F. Jamison, Reconfigurable system for automated optimization of diverse chemical reactions. *Science*. **361,** 1220 (2018).

7. D. Waltz, B. G. Buchanan, Automating Science. *Science*. **324,** 43 (2009).

8. M. Trobe, M. D. Burke, The Molecular Industrial Revolution: Automated Synthesis of Small Molecules. *Angewandte Chemie International Edition*. **57,** 4192–4214 (2018).

9. J. Li, S. G. Ballmer, E. P. Gillis, S. Fujii, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse, M. D. Burke, Synthesis of many different types of organic small molecules using one automated process. *Science*. **347,** 1221 (2015).

10. E. M. Woerly, J. Roy, M. D. Burke, Synthesis of most polyene natural product motifs using just 12 building blocks and one coupling reaction. *Nature Chemistry*. **6,** 484–491 (2014).

11. B. J. Reizman, K. F. Jensen, Feedback in Flow for Accelerated Reaction Development. *Acc. Chem. Res.* **49,** 1786–1796 (2016).

12. N. Hartrampf, A. Saebi, M. Poskus, Z. P. Gates, A. J. Callahan, A. E. Cowfer, S. Hanna, S. Antilla, C. K. Schissel, A. J. Quartararo, X. Ye, A. J. Mijalis, M. D. Simon, A. Loas, S. Liu, C. Jessen, T. E. Nielsen, B. L. Pentelute, Synthesis of proteins by automated flow chemistry. *Science*. **368,** 980 (2020).

13. J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2,** 725–732 (2016).

14. M. H. S. Segler, M. Preuss, M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. **555,** 604–610 (2018).

15. T. Badowski, E. P. Gajewska, K. Molga, B. A. Grzybowski, Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angewandte Chemie International Edition*. **59,** 725–730 (2020).

16. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **3,** 434–443 (2017).

17. C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison, K. F. Jensen, A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*. **365**, eaax1566 (2019).

18. S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone, L. Cronin, Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*. **363**, eaav2211 (2019).

19. L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, A. Aspuru-Guzik, ChemOS: Orchestrating autonomous experimentation. *Sci. Robotics*. **3**, eaat5559 (2018).

20. F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).

21. B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, C. P. Berlinguette, Self-driving laboratory for accelerated discovery of thin-film materials. *Sci Adv*. **6**, eaaz8867 (2020).

22. C. W. Coley, W. H. Green, K. F. Jensen, Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).

23. Z. Zhou, X. Li, R. N. Zare, Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **3**, 1337–1344 (2017).

24. H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).

25. C. W. Coley, N. S. Eyke, K. F. Jensen, Autonomous discovery in the chemical sciences part II: Outlook. *Angewandte Chemie International Edition*. **0** (2019), doi:10.1002/anie.201909989.

26. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nature News*. **533**, 452 (2016).

27. N. Hartrampf, A. Saebi, M. Poskus, Z. P. Gates, A. J. Callahan, A. E. Cowfer, S. Hanna, S. Antilla, C. K. Schissel, A. J. Quartararo, X. Ye, A. J. Mijalis, M. D. Simon, A. Loas, S. Liu, C. Jessen, T. E. Nielsen, B. L. Pentelute, Synthesis of Proteins by Automated Flow Chemistry (2020), doi:10.26434/chemrxiv.11833503.v1.

28. E. Atherton, V. Woolley, R. C. Sheppard, Internal association in solid phase peptide synthesis. Synthesis of cytochrome C residues 66–104 on polyamide supports. *J. Chem. Soc., Chem. Commun.*, 970–971 (1980).

29. S. Kent, *ChemInform*, in press.

30. V. K. Sarin, S. B. H. Kent, R. B. Merrifield, Properties of swollen polymer networks. Solvation and swelling of peptide-containing resins in solid-phase peptide synthesis. *J. Am. Chem. Soc.* **102**, 5463–5470 (1980).

31. R. C. de L. Milton, S. C. F. Milton, P. A. Adams, Prediction of difficult sequences in solid-phase peptide synthesis. *J. Am. Chem. Soc.* **112**, 6039–6046 (1990).

32. W. J. Van Woerkom, J. W. Van Nispen, Difficult couplings in stepwise solid phase peptide synthesis: predictable or just a guess? *International Journal of Peptide and Protein Research*. **38**, 103–113 (1991).

33. J. Bedford, C. Hyde, T. Johnson, W. Jun, D. Owen, M. Quibell, R. Sheppard, Amino acid structure and "difficult sequences" in solid phase peptide synthesis. *International journal of peptide and protein research*. **40**, 300–307 (1992).

34. A. J. Mijalis, D. A. Thomas III, M. D. Simon, A. Adamo, R. Beaumont, K. F. Jensen, B. L. Pentelute, A fully automated flow-based approach for accelerated peptide synthesis. *Nature Chemical Biology*. **13**, 464 (2017).

35. M. D. Simon, thesis, Massachusetts Institute of Technology, Cambridge, MA, USA (2017).

36. T. J. Lukas, M. B. Prystowsky, B. W. Erickson, Solid-phase peptide synthesis under continuous-flow conditions. *Proc Natl Acad Sci USA*. **78**, 2791 (1981).

37. E. Atherton, E. Brown, R. C. Sheppard, A. Rosevear, A physically supported gel polymer for low pressure, continuous flow solid phase reactions. Application to solid phase peptide synthesis. *J. Chem. Soc., Chem. Commun.*, 1151–1152 (1981).

38. L. R. Cameron, J. L. Holder, M. Meldal, R. C. Sheppard, Peptide synthesis. Part 13. Feedback control in solid phase synthesis. Use of fluorenylmethoxycarbonyl amino acid 3,4-dihydro-4-oxo-1,2,3-benzotriazin-3-yl esters in a fully automated system. *J. Chem. Soc., Perkin Trans. 1*, 2895–2901 (1988).

39. A. Dryland, R. C. Sheppard, Peptide synthesis. Part 8. A system for solid-phase synthesis under low pressure continuous flow conditions. *J. Chem. Soc., Perkin Trans. 1*, 125–137 (1986).

40. D. Rogers, M. Hahn, Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

41. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*. **128**, 336–359 (2020).

42. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Research*. **28**, 235–242 (2000).

43. P.-S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature*. **537**, 320–327 (2016).

44. N. L. Truex, R. L. Holden, B.-Y. Wang, P.-G. Chen, S. Hanna, Z. Hu, K. Shetty, O. Olive, D. Neuberg, N. Hacohen, D. B. Keskin, P. A. Ott, C. J. Wu, B. L. Pentelute, Automated Flow Synthesis of Tumor Neoantigen Peptides for Personalized Immunotherapy. *Scientific Reports*. **10**, 723 (2020).

45. J. M. Wolfe, C. M. Fadzen, Z.-N. Choo, R. L. Holden, M. Yao, G. J. Hanson, B. L. Pentelute, Machine Learning To Predict Cell-Penetrating Peptides for Antisense Delivery. *ACS Cent. Sci.* **4**, 512–520 (2018).

46. C. K. Schissel, S. Mohapatra, J. M. Wolfe, C. M. Fadzen, K. Bellovoda, C.-L. Wu, J. A. Wood, A. B. Malmberg, A. Loas, R. Gomez-Bombarelli, B. L. Pentelute, *bioRxiv*, in press, doi:10.1101/2020.04.10.036566.

47. H. L. Morgan, The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*. **5**, 107–113 (1965).

48. G. Landrum, RDKit: Open-source cheminformatics. *RDKit: Open-source cheminformatics* (2006).

49. S. Clark, P. Hayes, SigOpt WebPage. *SigOpt Web page* (2019).

Methods

Automated flow peptide synthesis and UV-Vis data collection. All peptides were synthesized on three automated-flow systems, which were built in the Pentelute lab and were described in detail in previous publications.(*27*, *34*, *35*) The automated set up records amino acid sequence, stock solution type, pump strokes, flow rate, temperatures in heating loops and at the entrance and exit of the reactor, backpressure and in line UV-Vis data for every synthesis.

For test syntheses in this manuscript, synthesis conditions detailed in **SI Table 2** were used. Capitalized letters refer to L-amino acids, uncapitalized letters refer to D-amino acids or uncommon building blocks, which are defined in the SI.

Unless otherwise noted, the following stock solutions were used for peptide synthesis: Fmoc-protected amino acids (Fmoc-Ala-OHxH2O, 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) as a 0.40 M stock solution in DMF, activating agents (HATU and PyAOP) as a 0.38 M stock solution in DMF, DIEA (undiluted), deprotection stock solution (40% piperidine, 2% formic acid, 58% DMF). DMF was pre-treated with AldraAmine trapping agents >24 h before synthesis. 50–200 mg H-Rink Amide (0.49 mmol/g and 0.18 mmol/g loading) and HMPB ChemMatrix polyethylene glycol (0.45 mmol/g loading) resin was used in all experiments in the data set, details on resin and scale are given for synthesis examples in this manuscript SI.

Unless otherwise noted, a flow-rate of 40 mL/min and a temperature of 90 °C in the loop and 85–90 °C in the reactor was used. Briefly,two large pumps (50 mL/min pump head) delivers 400 µL of solution per pump stroke and a small pump (5 mL/min pump head) delivers 40 µL of solution per pump stroke. A standard synthesis cycle involves (a) prewashing of the resin, (b) iterative coupling, washing, deprotection and washing steps per amino acid building block. In the prewashing step the resin in swollen at elevated temperatures for 60 s at 40 mL/min. The iterative synthesis cycles start with a coupling step where three HPLC pumps are used: a large pump delivers the activating agent stock solution, a second large pump delivers the amino acid stock solution and a small pump delivers DIEA. It is important to make sure that all solutions reach the mixer in the flow set up at the same time to avoid by-product formation. The first two pumps are delivering stock solutions for 8 pumping strokes in order to prime the coupling agent and amino acid lines before the DIEA pump is started. The three pumps are then delivering sock solutions together for a period of 7 pumping strokes. Afterwards, the activating agent pump and the amino acid pump are changed using a rotary valve to select DMF. The three pumps are pumping together for a final 8 pumping strokes. For the consecutive washing step, the DIEA pump is stopped and the other two pumps continue delivering DMF for another 40 pump strokes.

In the deprotection step, the two large pumps are used, one delivering DMF and one delivering the deprotection solution in a 1:1 ratio. The pumps are activated for 13 pump strokes. Next, the rotary valves select DMF for both pumps, and the resin is washed for an additional 40 pump strokes. The coupling–deprotection cycle is repeated for every additional amino acid.

UV-Vis in line analysis is recorded past the reactor and prior to waste collection. The UV synthesis data at a wavelength of 310 nm was collected from 35,427 individual deprotection steps from 1523 unique peptide synthesis experiments on three AFPS systems. Sequences with canonical amino acids and with length between 5 and 50 amino acids only were considered in the making of the data set. The recipe file and AFPS raw file were analyzed to collect information about the coupling agent, coupling strokes, coupling temperature, deprotection strokes, flow rate and reactor temperature. Integral, width and height of the time-resolved traces were obtained using a modified version of the earlier published code.(*12*)

Deep learning and optimization

Data pre-processing. The data set obtained from the AFPS was pre-processed before analysis (**SI Section 3.1**). Two individual sets of normalization, by first and previous deprotection step, were performed. Difference of width and height was calculated from the normalized traces. With 4 parameters and 2 different types of normalization, each deprotection step was quantified in terms of 8 variables. Out of these variables, normalization-specific sets of 4 parameters were used for different tasks. The analysis was performed on the parameters normalized by the previous deprotection step, and machine learning model was trained on parameters normalized by first deprotection step.

The data set was trimmed to 28642 deprotection steps after removing the outliers. For parameters from UV traces, a cut-off of 2 standard deviation for integral, width and height, and 1.5 for difference was used to filter the data set. Deprotection steps with HATU and PyAOP as coupling agents; 8 and 21 as coupling strokes; 9, 13, 20 and 26 as deprotection strokes; and flow rates of 40 and 80 mL/min were considered in the data set. After averaging over the traces based on the pre-chain, incoming amino acid and synthesis parameters, a total of 17459 unique deprotection steps.

Featurization. The pre-chain and incoming amino acid were featurized using 128-bit Morgan fingerprint bit-vectors generated using RDKit (**Appendix 1, 2**).(*47*, *48*) Coupling agent (HATU, PyAOP), coupling strokes (8 - single, 21 - double), deprotection strokes (9, 13, 20, 26) and flow rates (40, 80 mL/min) were treated as one-hot encoding representations. A machine variable (AFPS00, AFPS01, AFPS02) representing the particular set-up in the lab on which the sequence was synthesized was added as a one-hot encoding. The coupling temperature and reactor temperature were treated as continuous parameters. All parameters were normalized to mean 0 and standard deviation 1 before training.

Model Training. The deep learning model was based on a multi-modal convolutional neural network architecture. The input parameters included pre-chain, incoming amino acid, coupling agent, coupling strokes, deprotection strokes, coupling temperature, reactor temperature, flow rate and machine variable. Different sets of output parameters with individual and multiple combinations of normalization-specific parameters were tried. The best performance was obtained using integral, width, height and difference normalized by the first deprotection step. All hyperparameters were optimized using SigOpt.(*49*) A train-validation split of 70-30 was used for the training. The model has a RMSE validation loss of 0.52, 0.56, 0.47 and 0.48 for normalized integral, width, height and difference respectively.

Interpretability using gradient activation. Gradient activation analysis, based on our earlier work, was used to interpret the decision-making process of the model. A model with pre-chain and incoming amino acid features was used for the analysis. The pre-chain gradient map was used for analyses of average of activated bit-vectors and amino acids. The map obtained from averaging over bit-vectors was used for interpretation of aggregating positions and optimization of synthesis success by single point mutations.

Generation of mutants for optimization of aggregation. A brute-force approach was used to explore all possible single point mutations of the seed sequence. Given the small sequence space for optimization, less than 1000 for sequences with 50 or less amino acids, this approach exhaustively explored the combinatorial space. The predicted trace and activation map for each mutant were obtained. The lowest aggregating sequences and the most aggregating sequence (as negative control) were selected for experimental validation.

Data availability

The dataset, excluding proprietary sequences, used in the training and analysis of the model has been provided in the online repository.

Code availability

All code used for training and optimization of the model is available at https://github.com/learningmatter-mit/peptimizer.

Acknowledgments

This research was funded by Novo Nordisk, the MIT-SenseTime Alliance on Artificial Intelligence, and by an award from the Abdul Latif Jameel Clinic for Machine Learning in Health (J-Clinic). We would like to thank Dr. Z. P. Gates, Dr. E. D. Evans, Dr. A. J. Mijalis, Prof. Dr. T. E. Nielsen, Prof. T. F. Jamison, Dr. C. Jessen, Dr. H. U. Stilz, Dr. L. F. Iversen, and Dr. K. Little for helpful discussions.

Author contributions

N.H., R.G.-B. and B.L.P. conceptualized the research; N.H. and M.P. curated AFPS data set and synthesized, purified and analyzed test peptides reported in this manuscript; S.M. analyzed the AFPS data set and wrote the code for the deep learning model, S.M. and N.H analyzed the output of the model, S.M, N.H., A.L., R.G.B. and B.L.P wrote the manuscript with input of all co-authors.

Competing interests

B.L.P. is a co-founder of Amide Technologies and Resolute Bio. Both companies focus on the development of protein and peptide therapeutics. B.L.P is co-inventor on U.S. Pat. Appl. 20170081358A1 (March 23, 2017) describing methods and systems for solid phase peptide synthesis and on U.S. Pat. 9,868,759 (January 16, 2018), U.S. Pat. 9,695,214 (July 4, 2017), and U.S. Pat. 9,169,287 (October 27, 2015) describing solid phase peptide synthesis processes and associated systems.

Supplementary Materials

Materials and General Methods
Supplementary Text
SI Figures 1-11
SI Tables 1-4
Appendices 1, 2

Supplementary Materials for

Deep Learning for Prediction and Optimization of Rapid 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,*]

Affiliations:

[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, Wintherthurerstrasse 190, 8057 Zurich, Switzerland

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

This file includes –

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-OHxH_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.

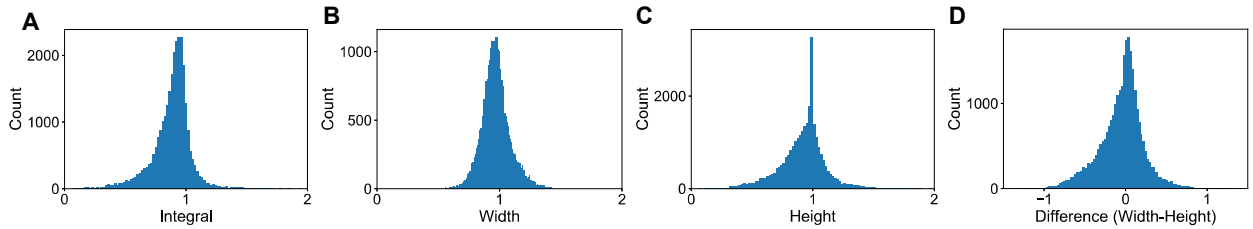
3 Deep learning and optimization

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

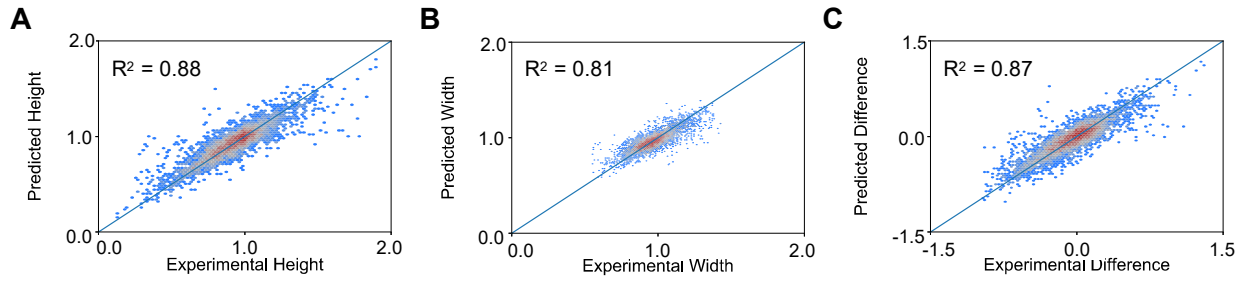
For the processed data set, where all parameters are normalized by the first deprotection step, we see a Gaussian-like distribution for each of the parameters (**SI Figure 1**).



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

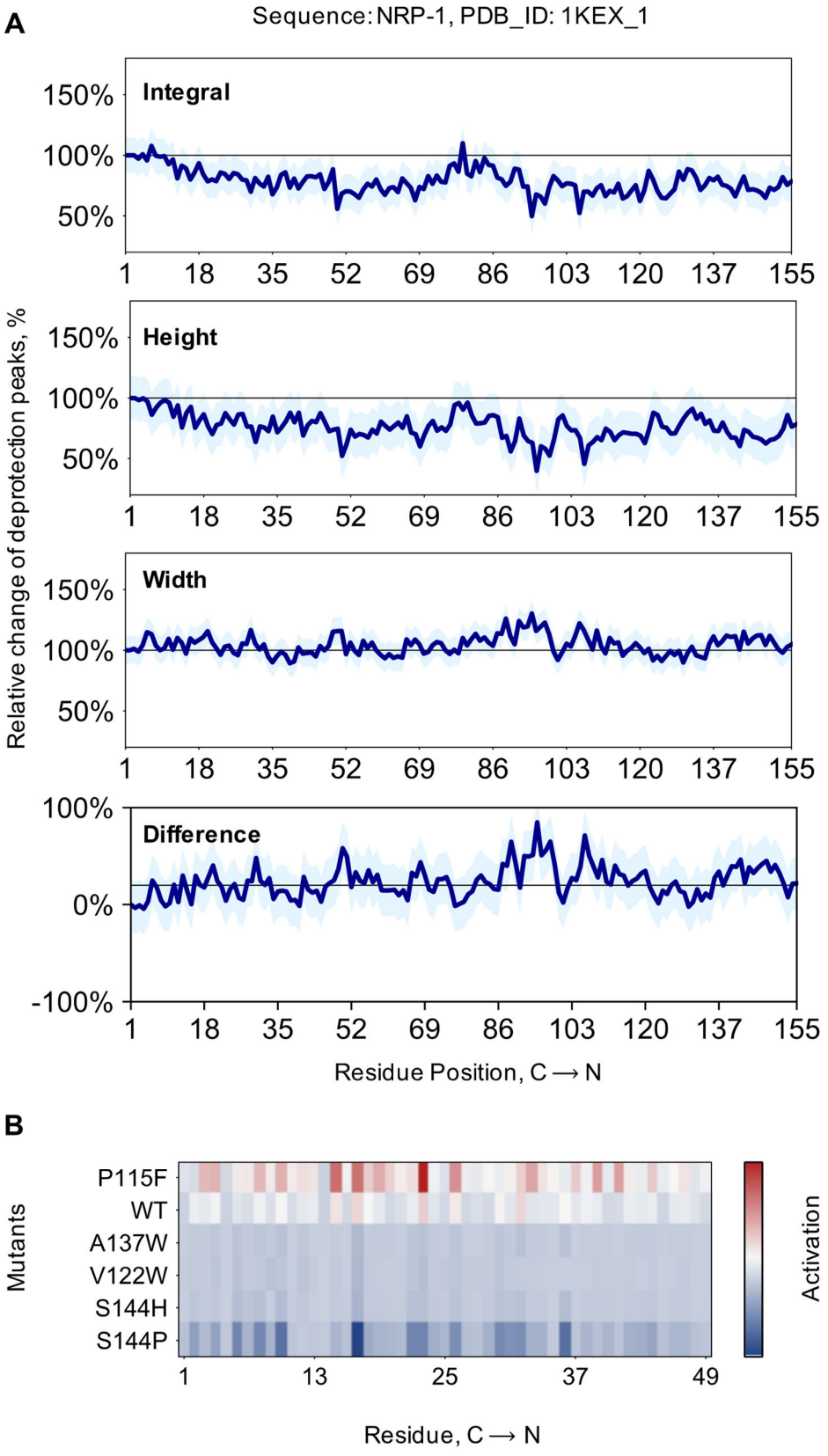
3.3 Automated pipeline for optimization of difficult-to-synthesize sequences

An automated pipeline was set-up for prediction of traces for classical difficult-to-synthesize sequences.(*8*, *9*) 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**).

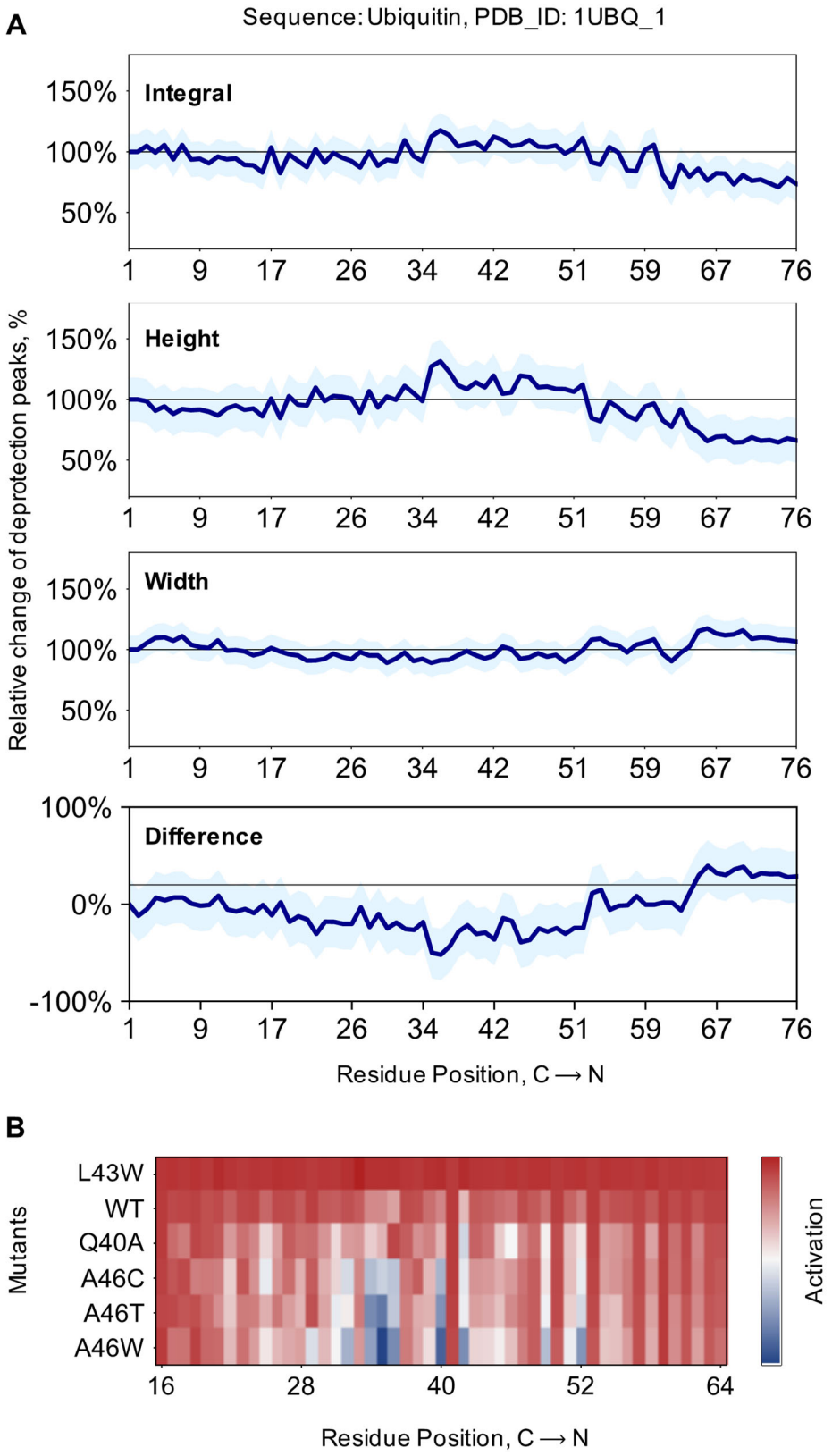
3.4 Prediction of traces and mutants for difficult-to-synthesize sequences

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.

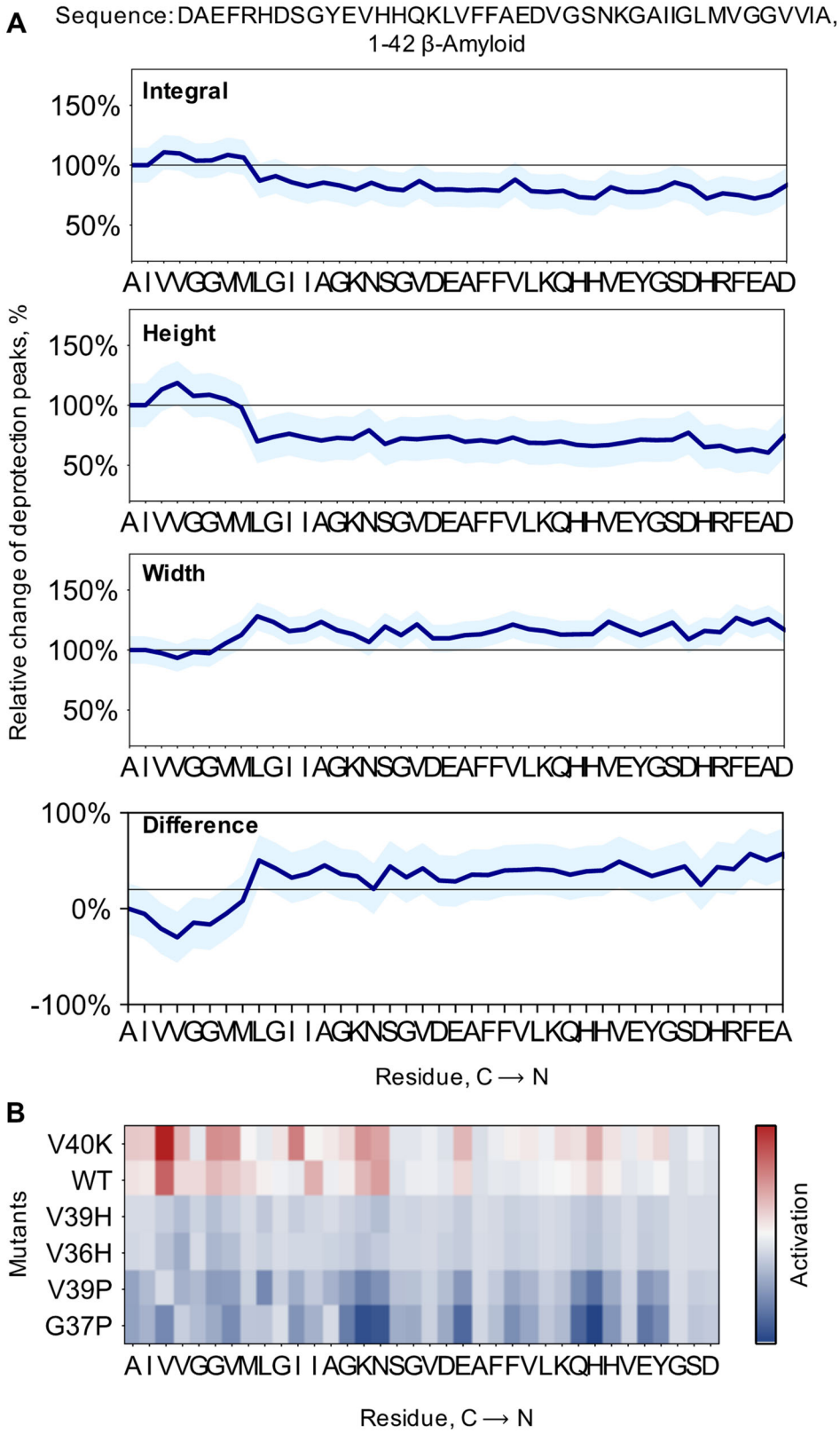
3.4.1 NRP-1, PDB_ID: 1KEX_1



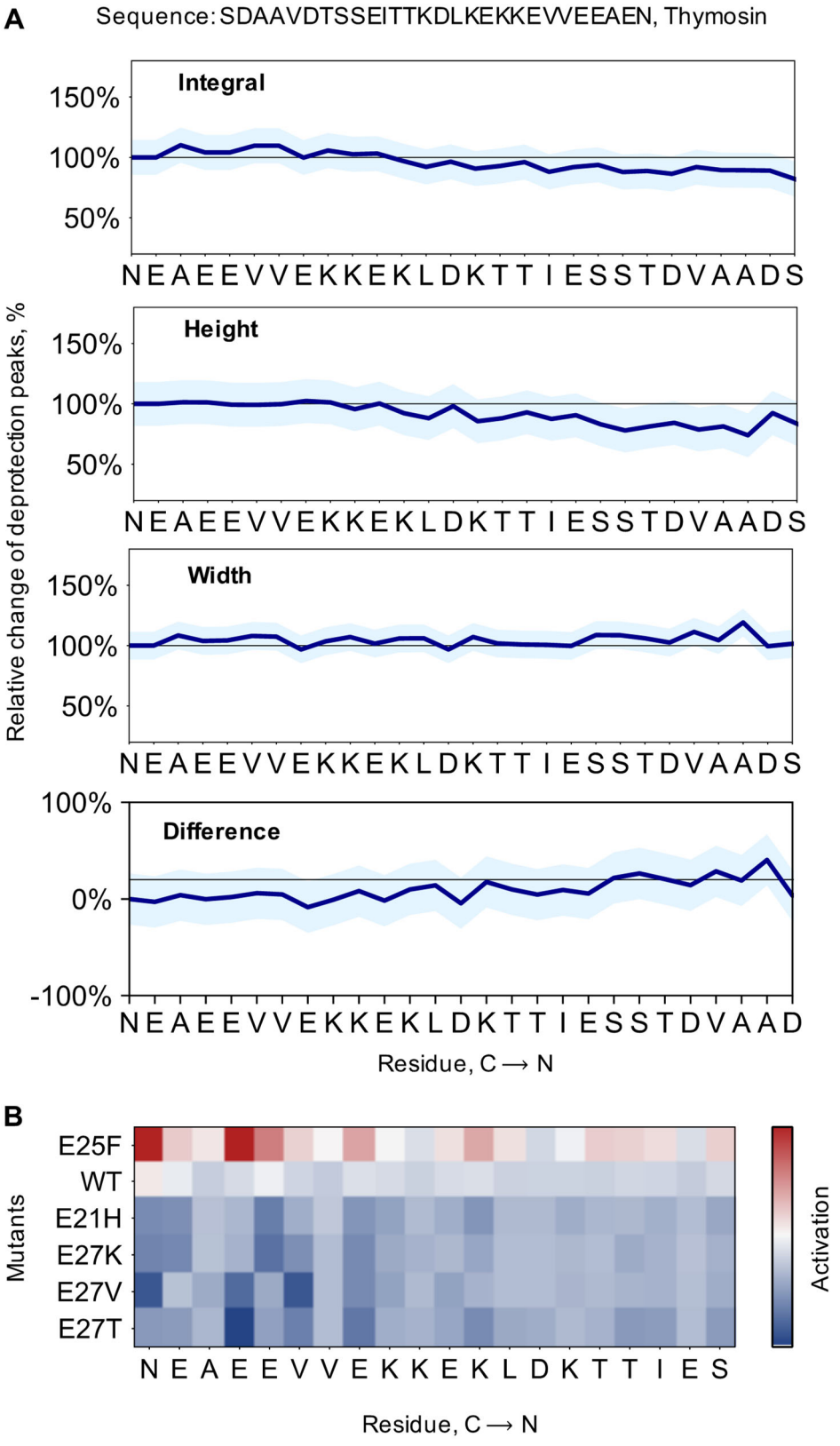
3.4.2 Ubiquitin, PDB_ID: 1UBQ_1



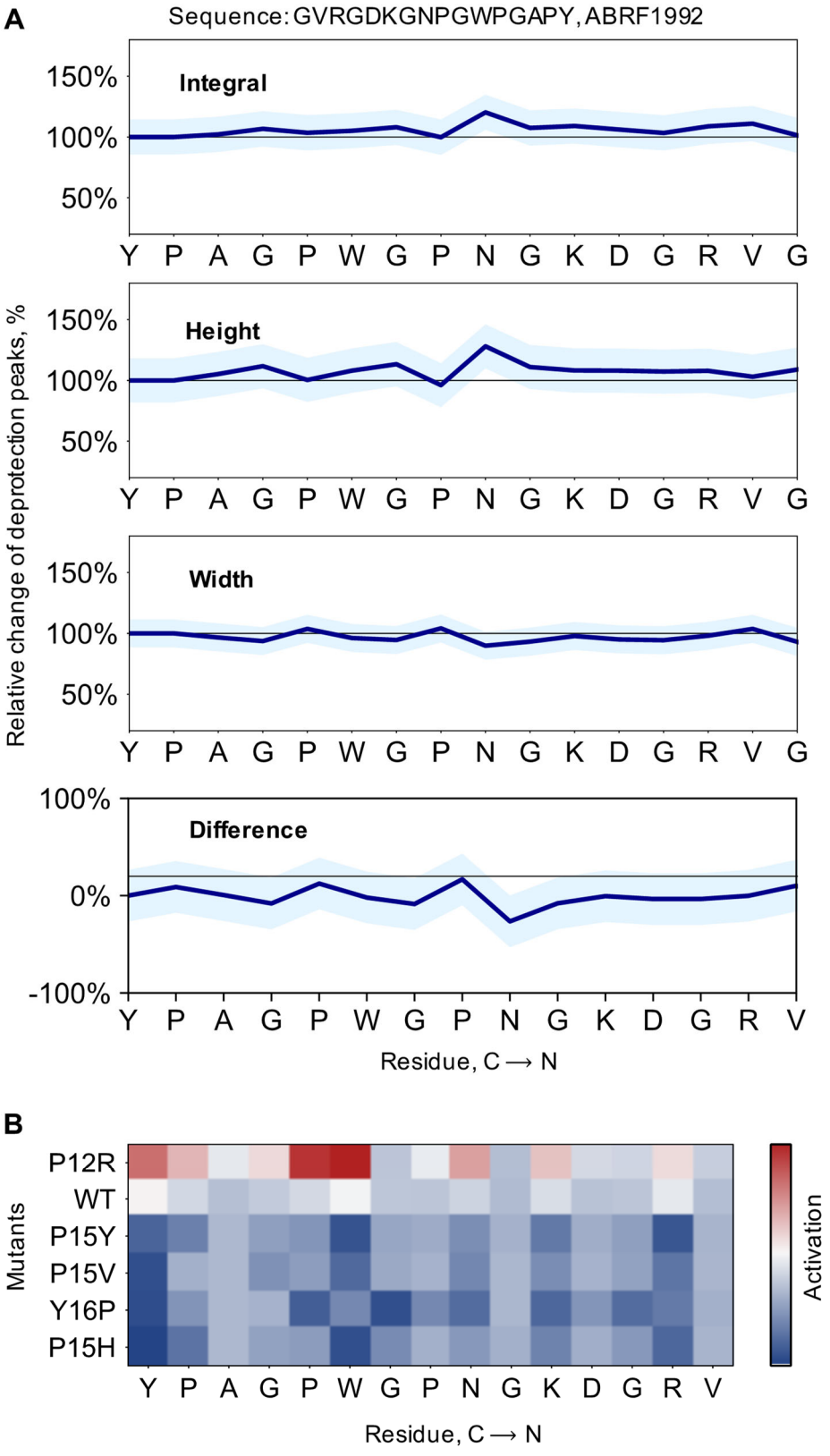
3.4.3 1-42 β-Amyloid



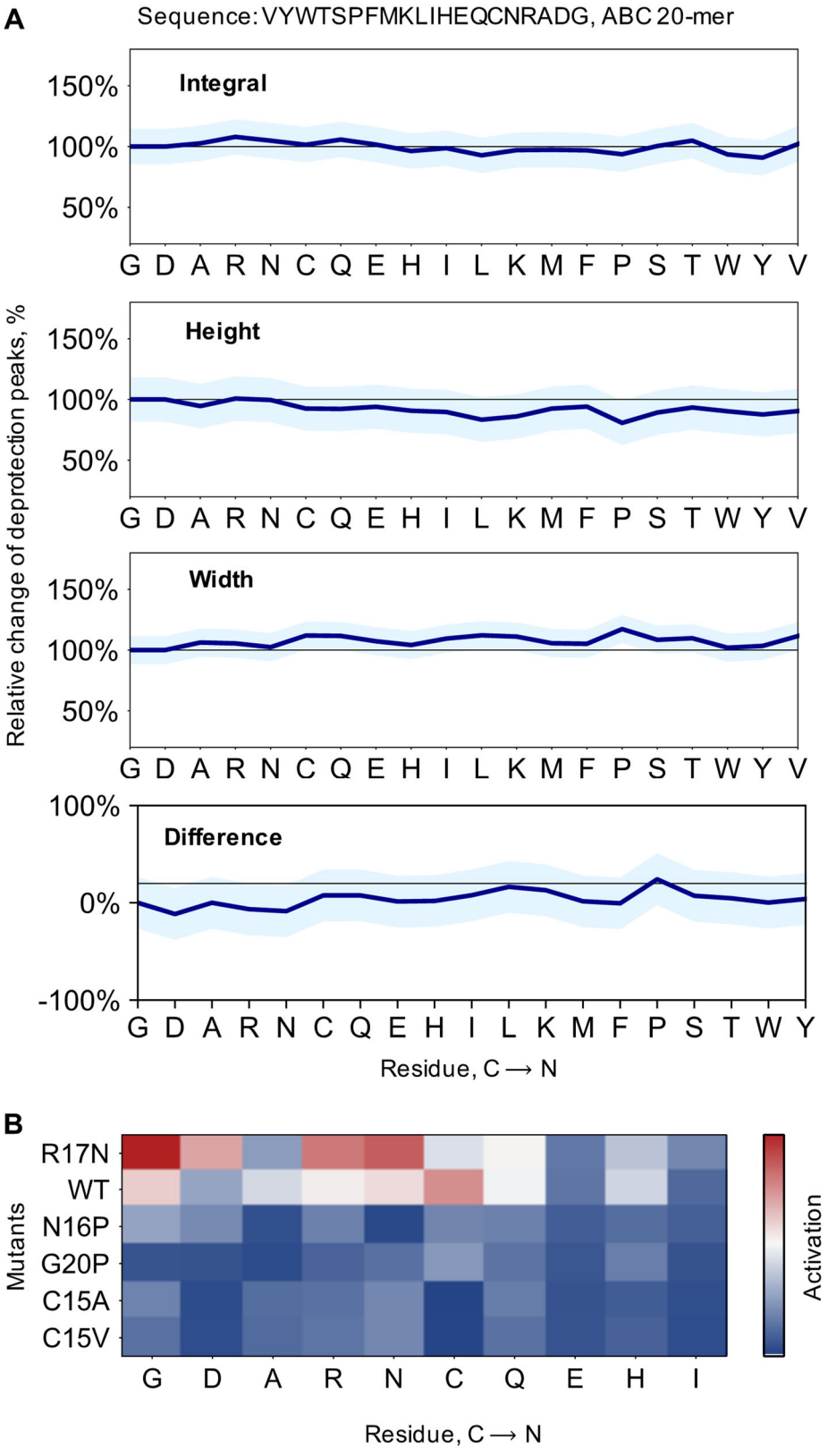
3.4.4 Thymosin



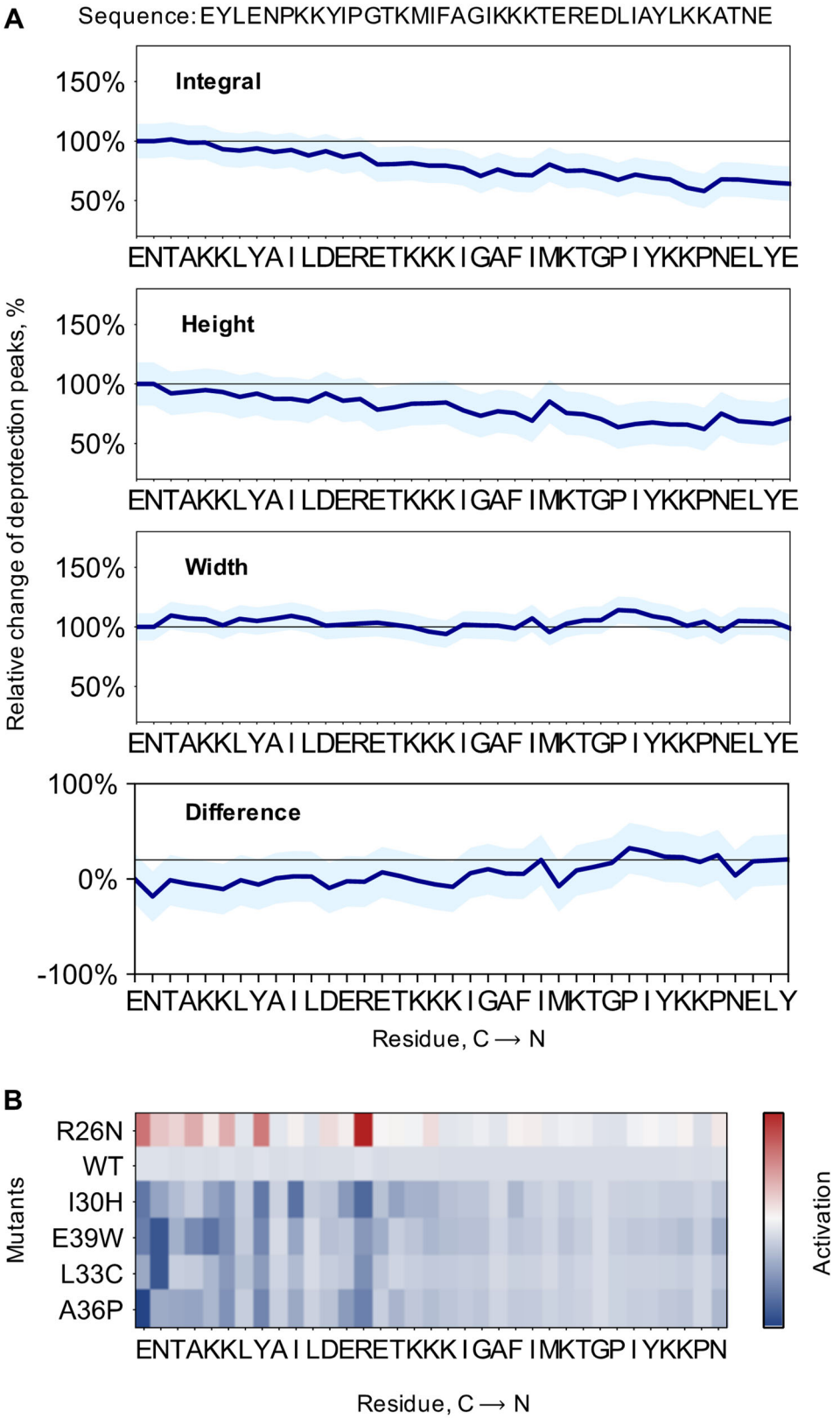
3.4.5 ABRF 1992



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

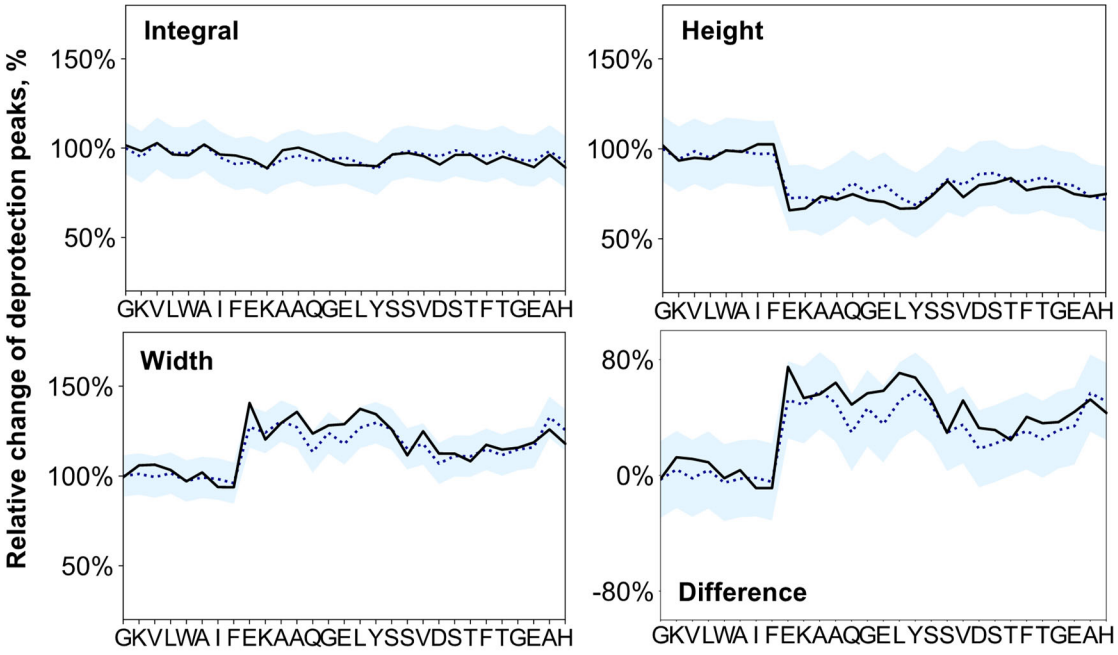
Synthesis Data for GLP-1 (R30S)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGS (30 AA)

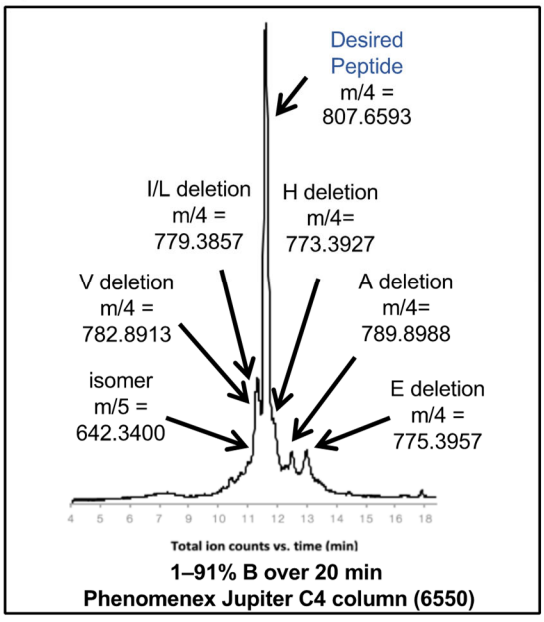
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

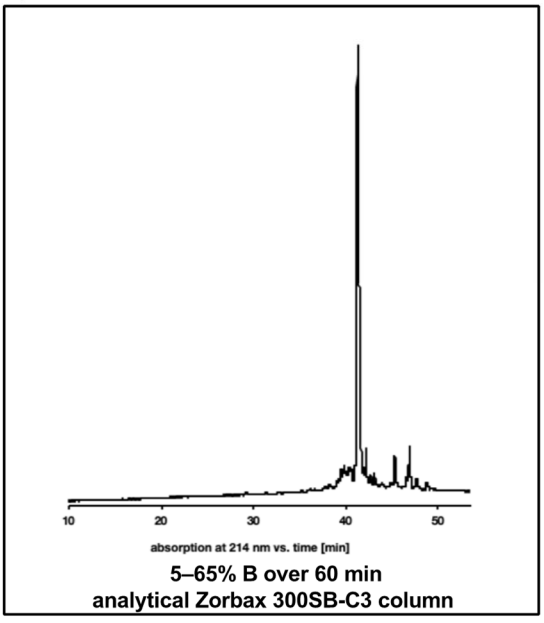
a) LCMS data, TIC



b) LCMS data, TIC



c) HPLC data



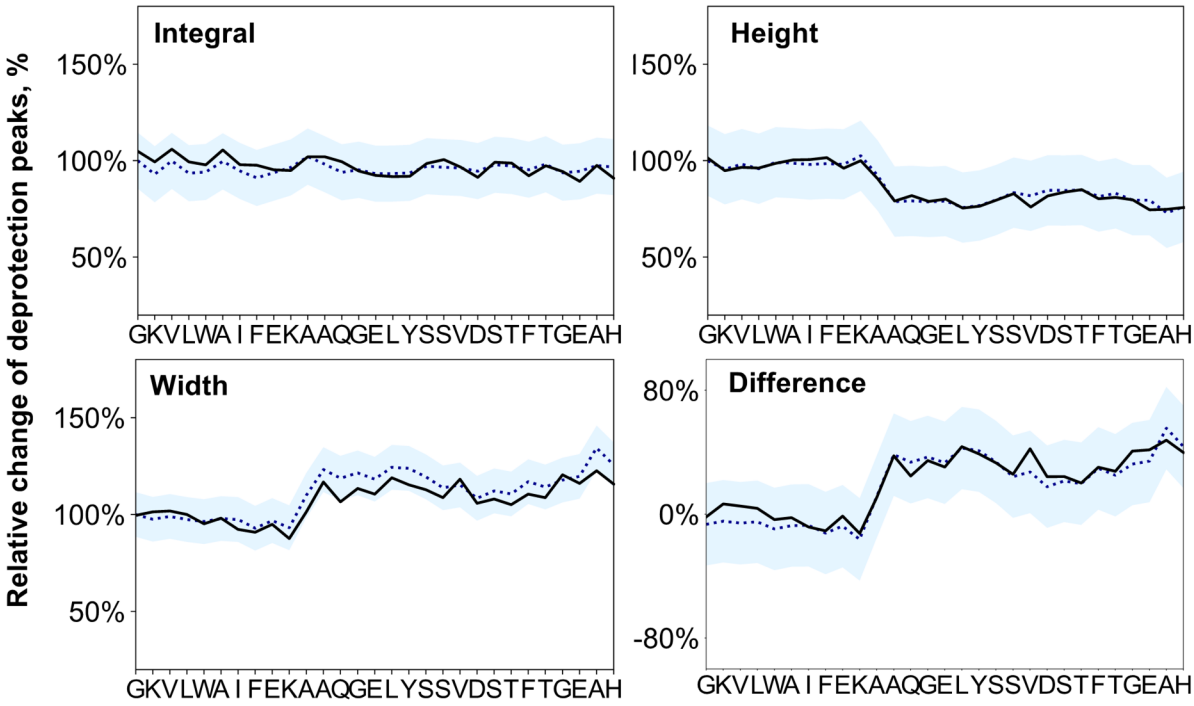
Synthesis Data for GLP-1 (WT)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR (30 AA)

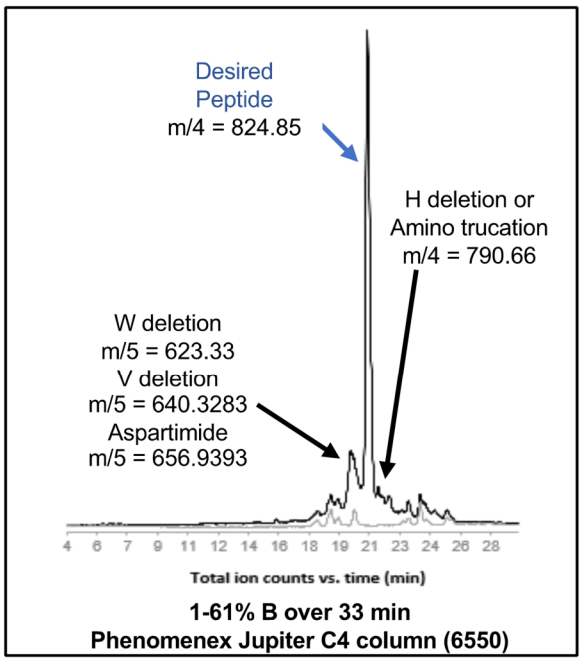
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

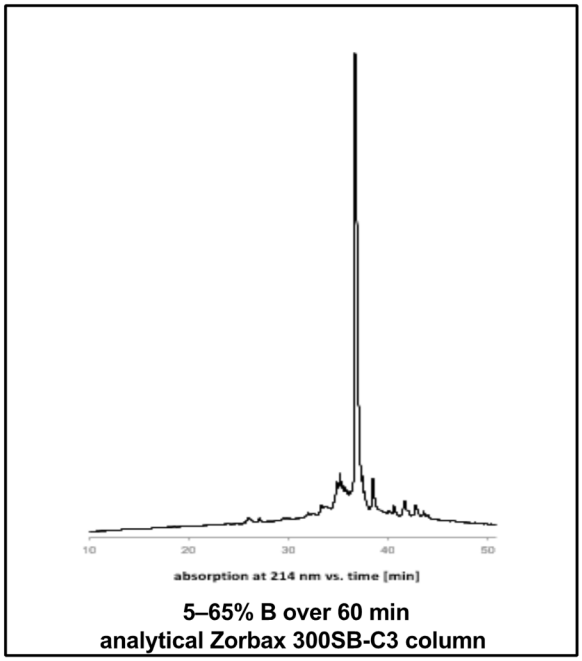


b) LCMS data, TIC



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

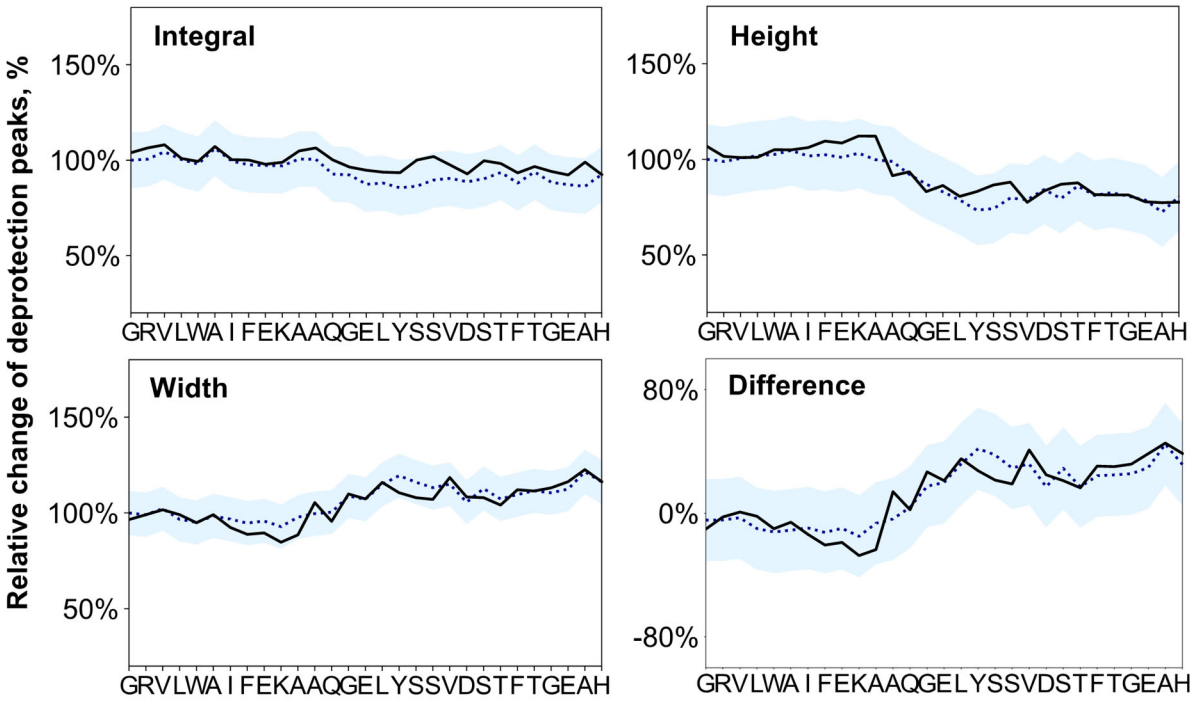
Synthesis Data for GLP-1 K28R

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAWLVRGR (30 AA)

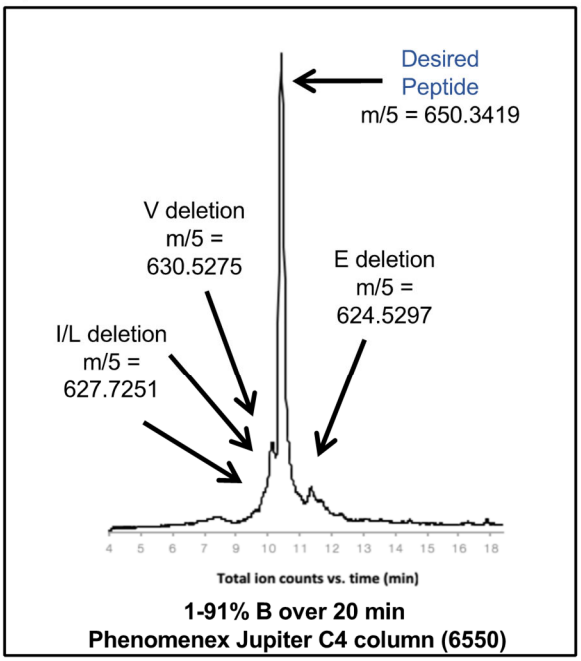
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

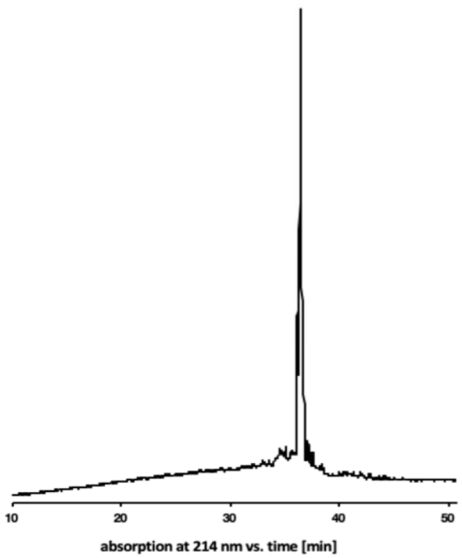


b) LCMS data, TIC



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

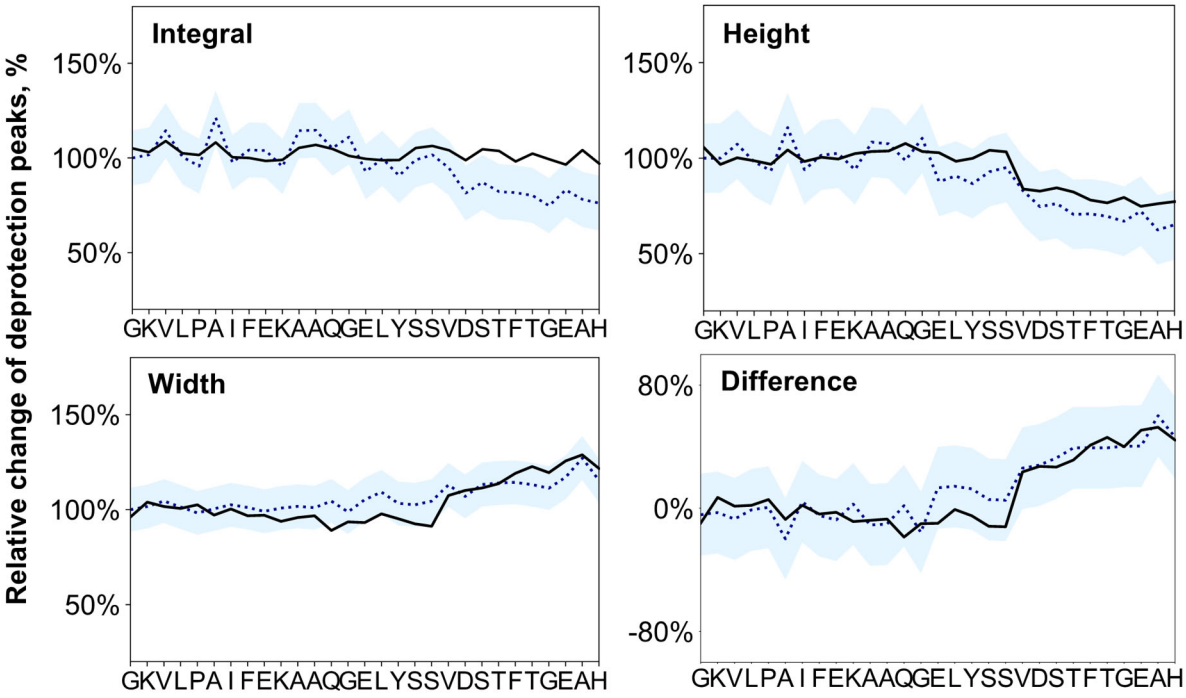
Synthesis Data for GLP-1 (W25P)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAPLVKGR (30 AA)

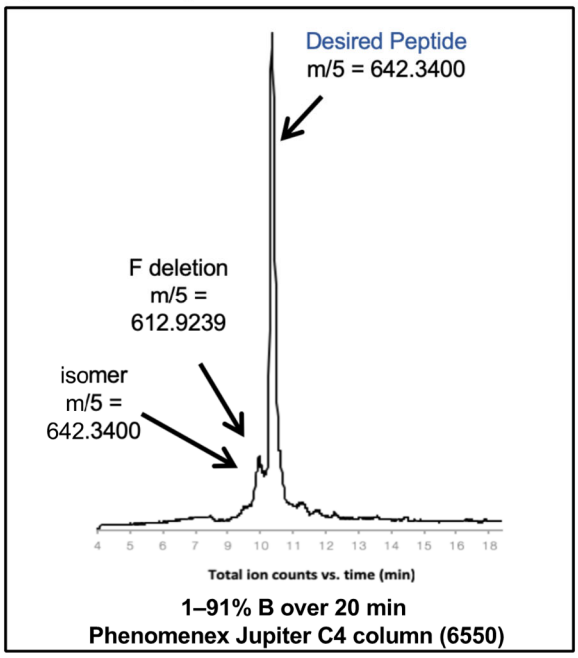
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

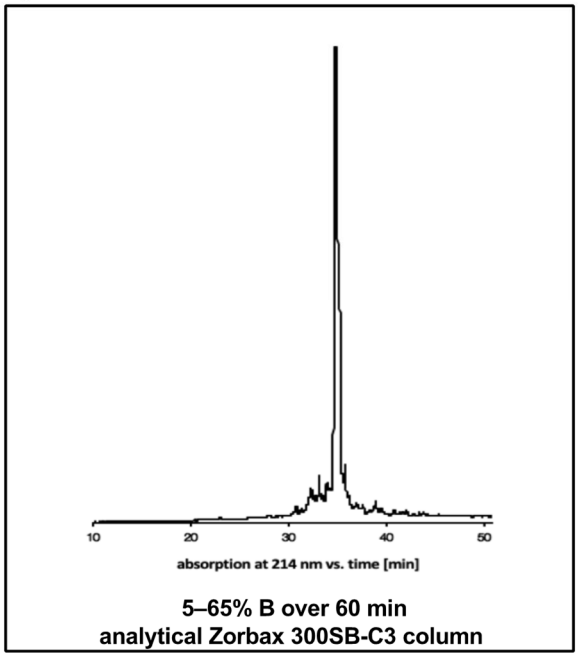


b) LCMS data, TIC



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

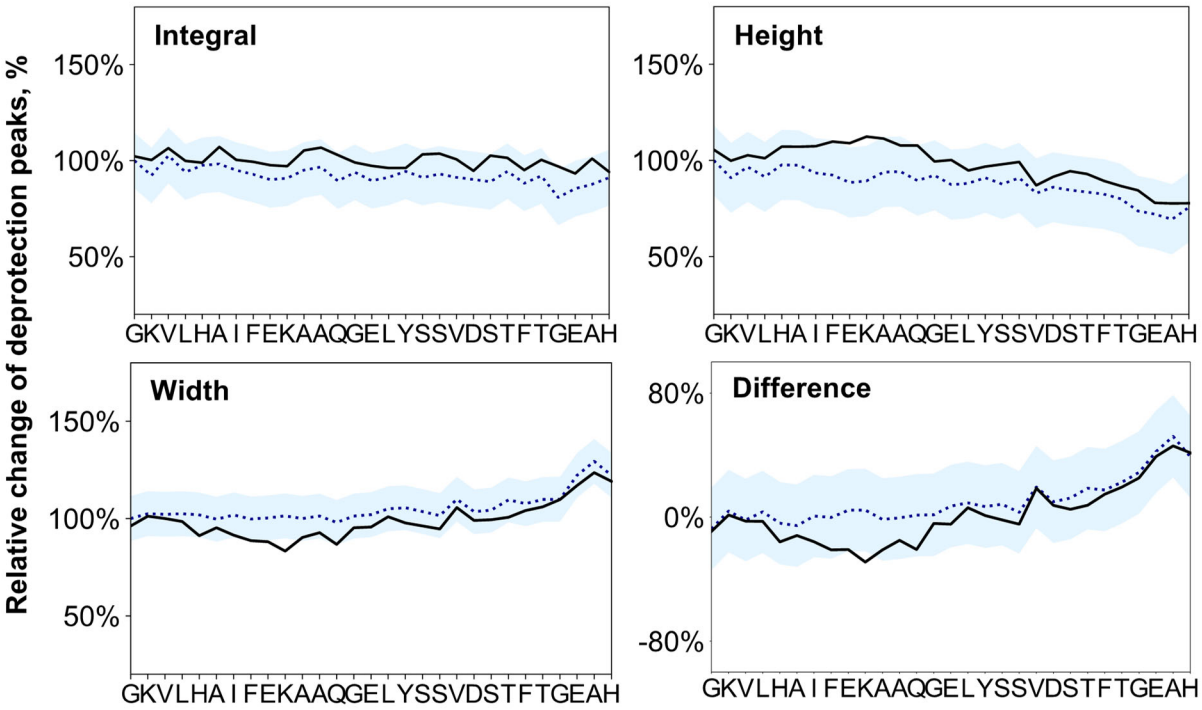
Synthesis Data for GLP-1 (W25H)

Sequence: HAEGTFTSDV SSYLEGQAAK EFIAHLVKGR (30 AA)

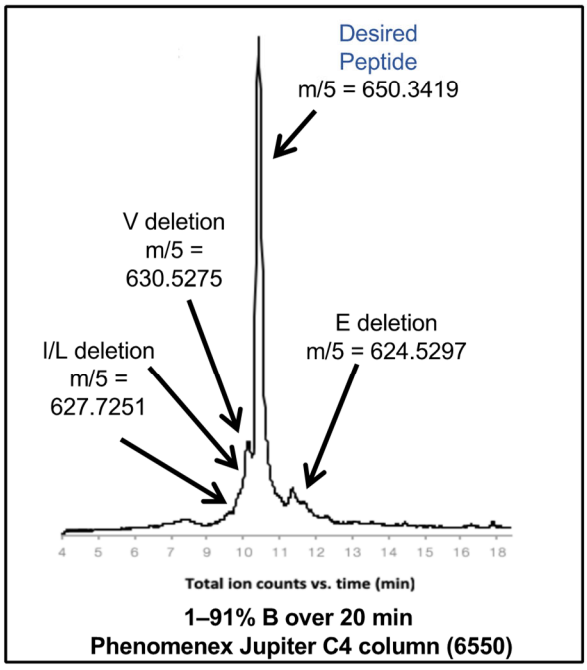
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 1.1 h

a) LCMS data, TIC

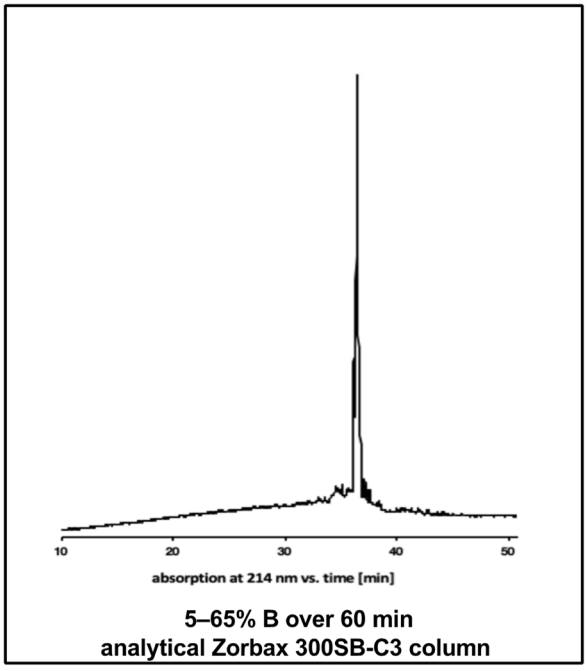


b) LCMS data, TIC



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

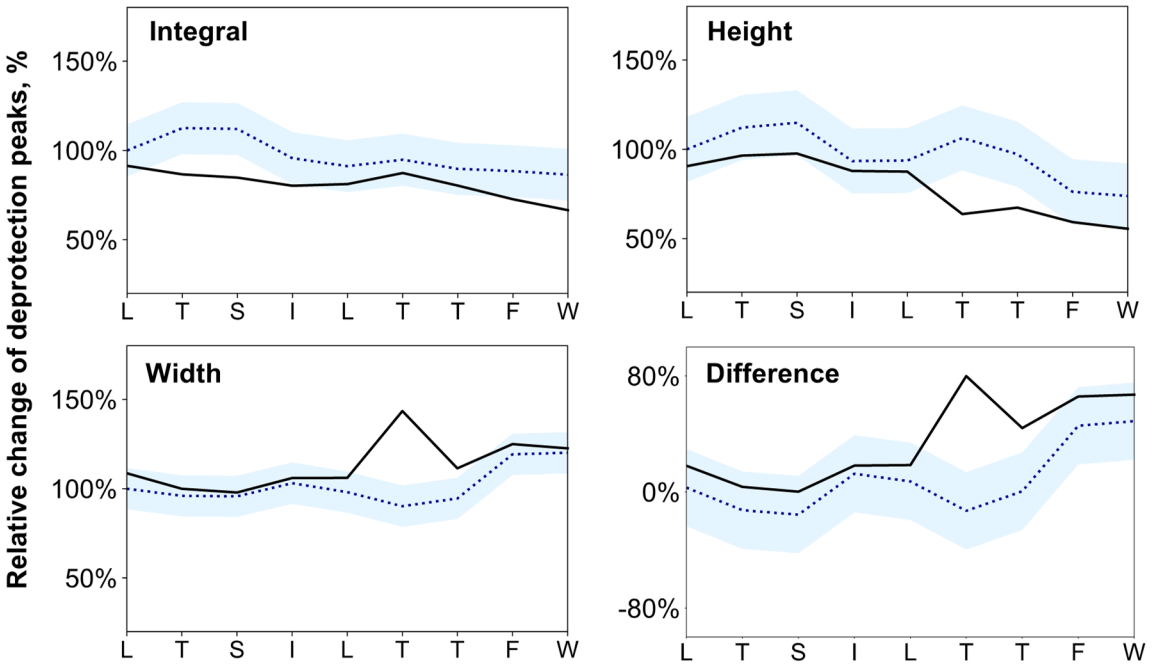
Synthesis Data for JR-10 (I9L)

Sequence: WFFTL ISTLM (10 AA)

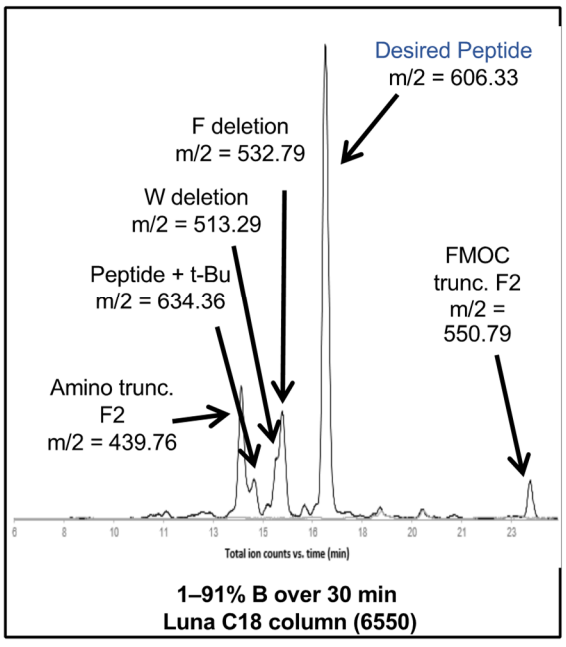
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

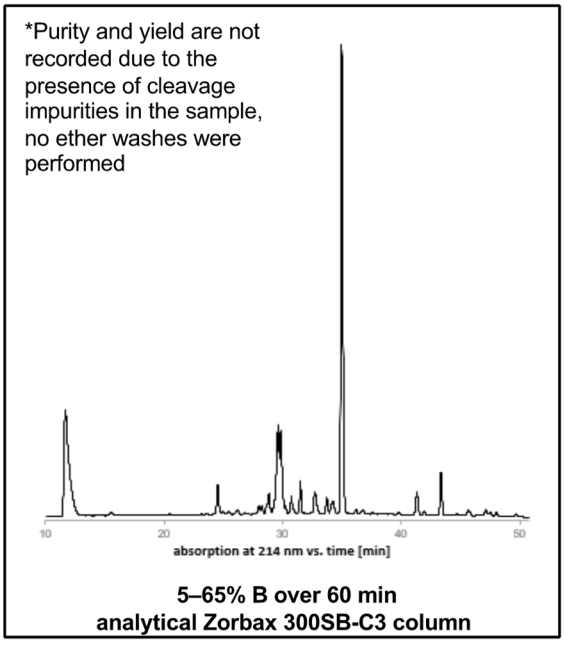


b) LCMS data, TIC



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

c) HPLC data



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

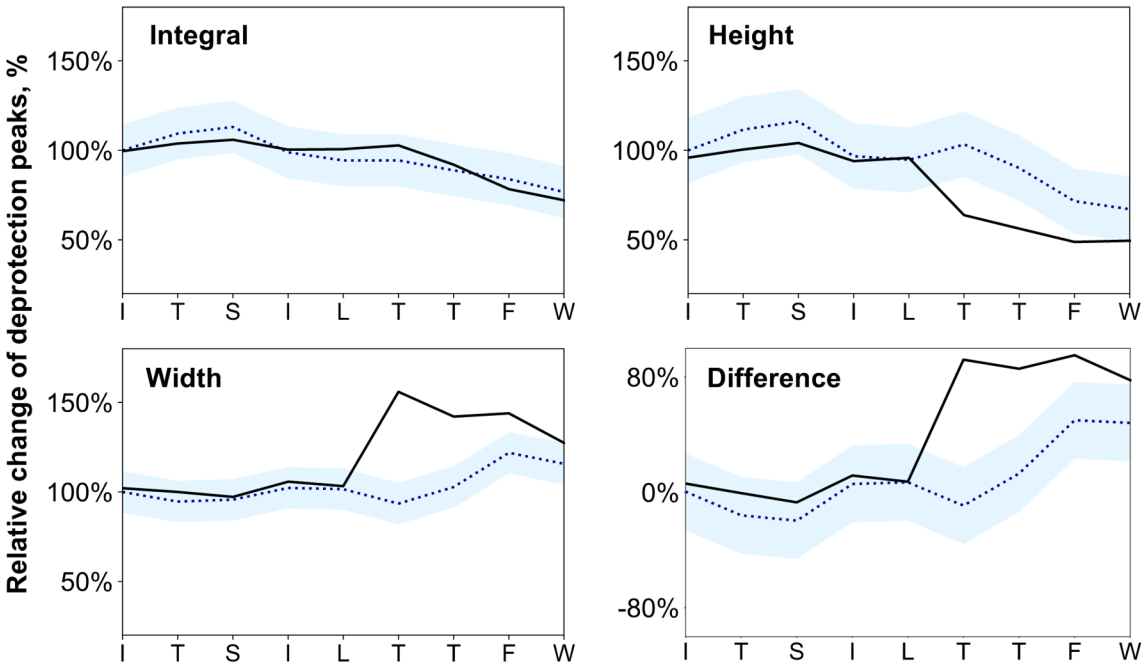
Synthesis Data for JR-10 (WT)

Sequence: WFFTL ISTIM (10 AA)

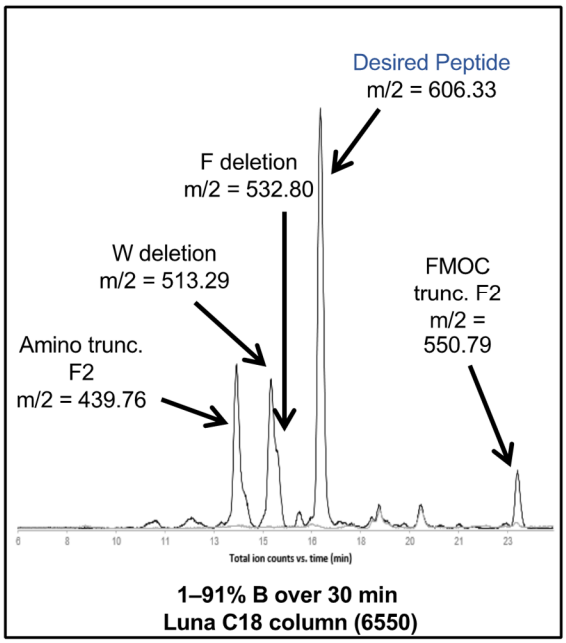
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

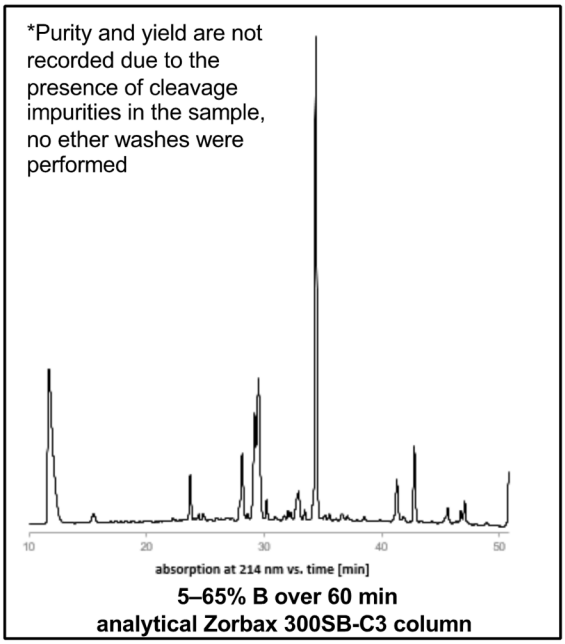


b) LCMS data, TIC



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

c) HPLC data



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

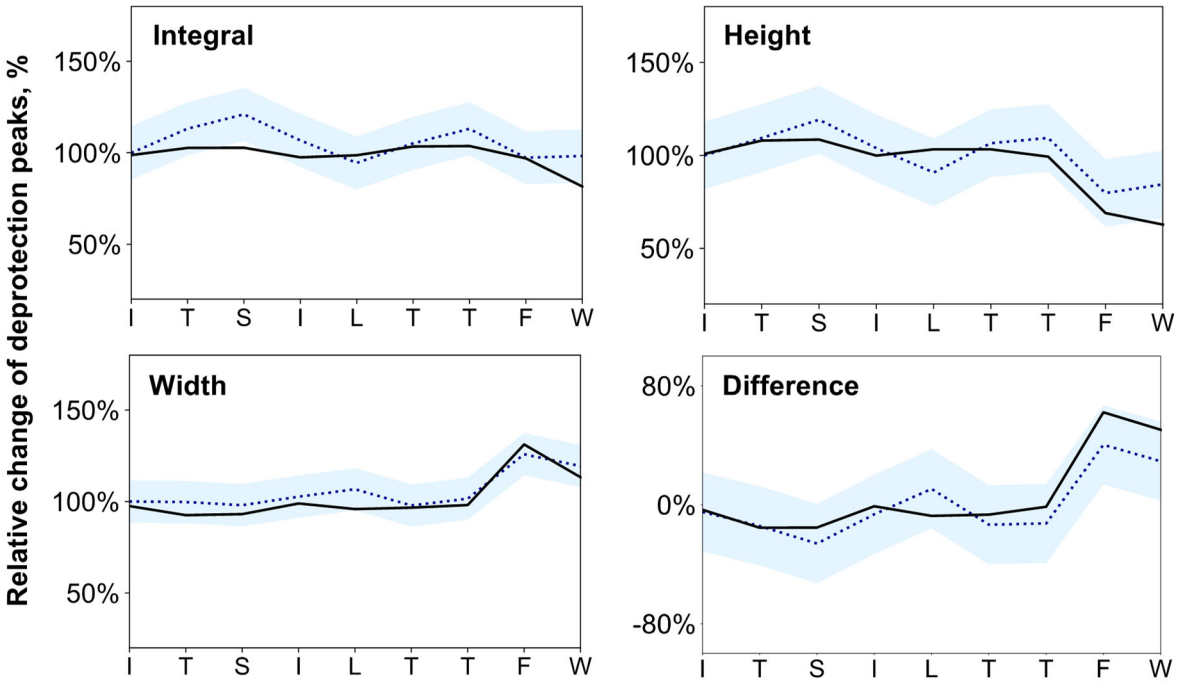
Synthesis Data for JR-10 (M10K)

Sequence: WFFTL ISTIK (10 AA)

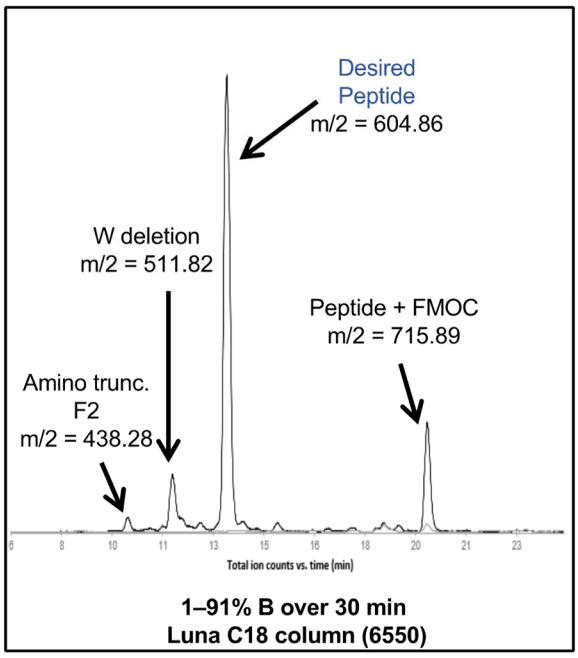
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

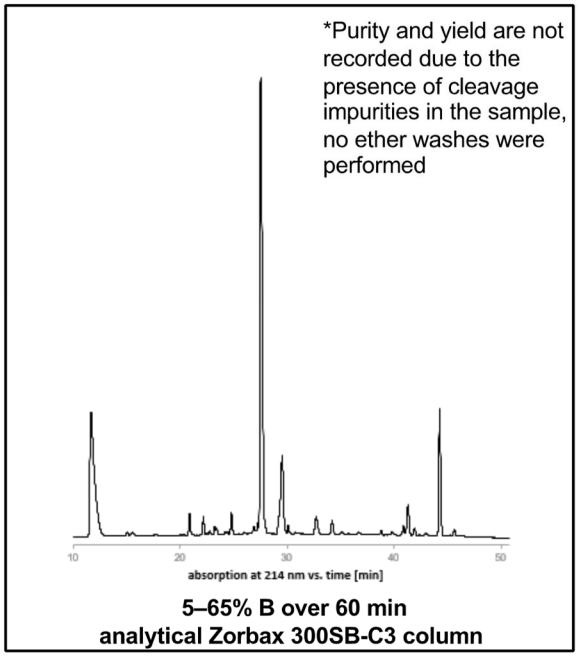


b) LCMS data, TIC



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

c) HPLC data



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

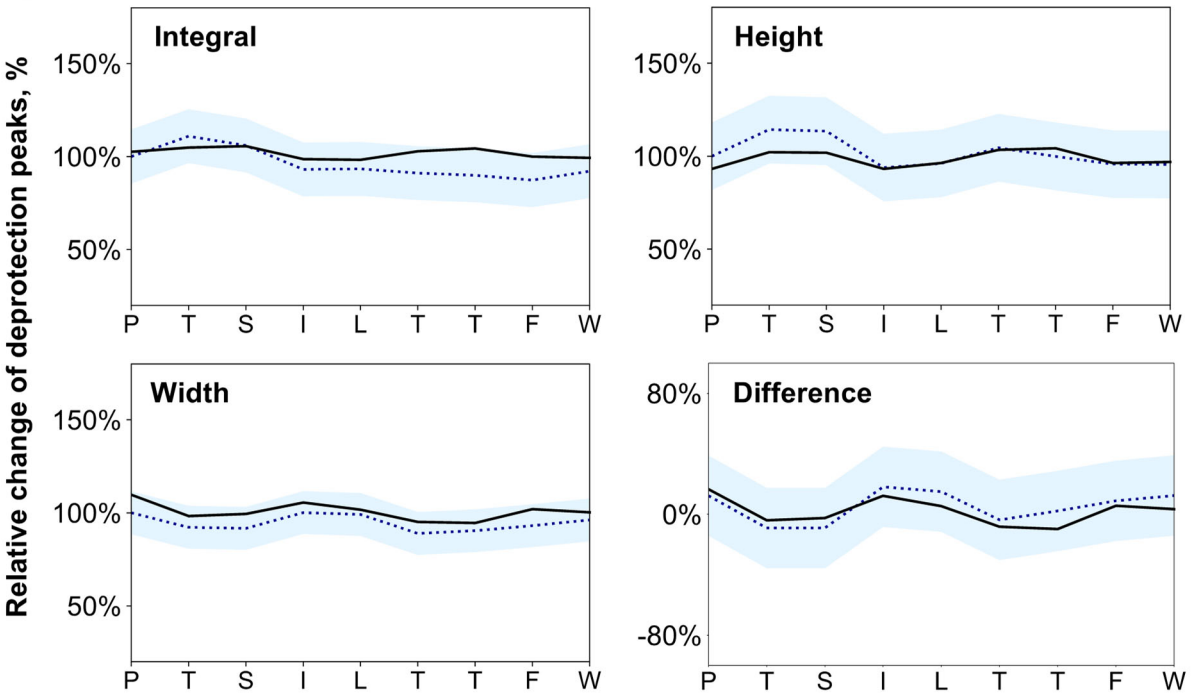
Synthesis Data for JR-10 (I9P)

Sequence: WFFTL ISTPM (10 AA)

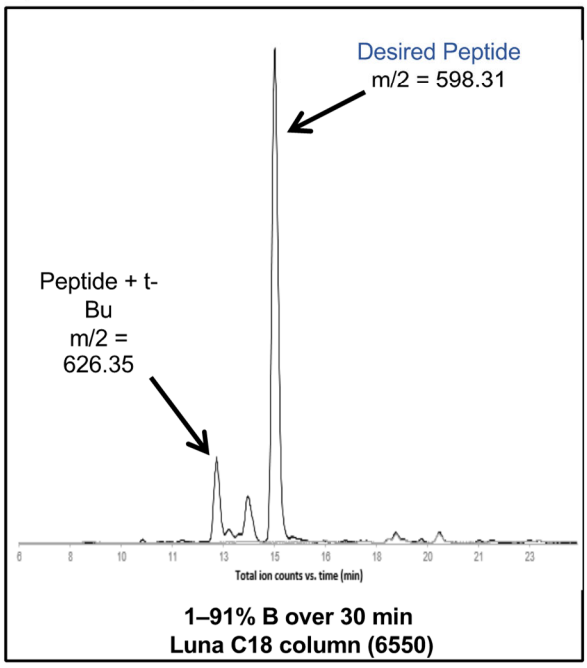
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

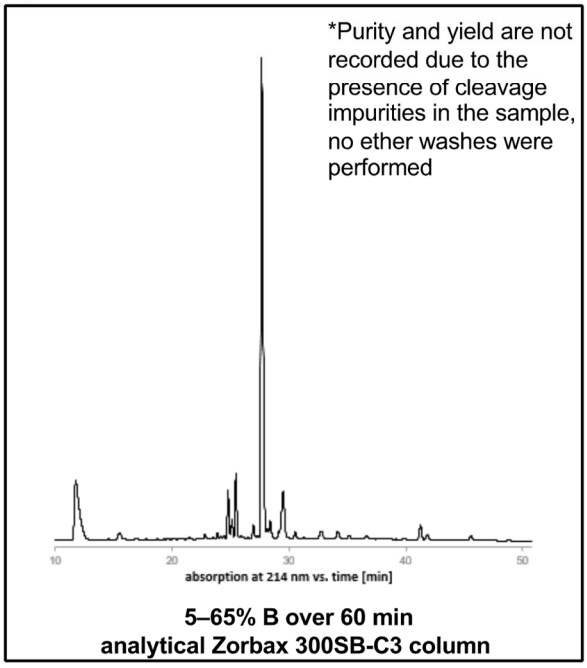


b) LCMS data, TIC



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

c) HPLC data



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

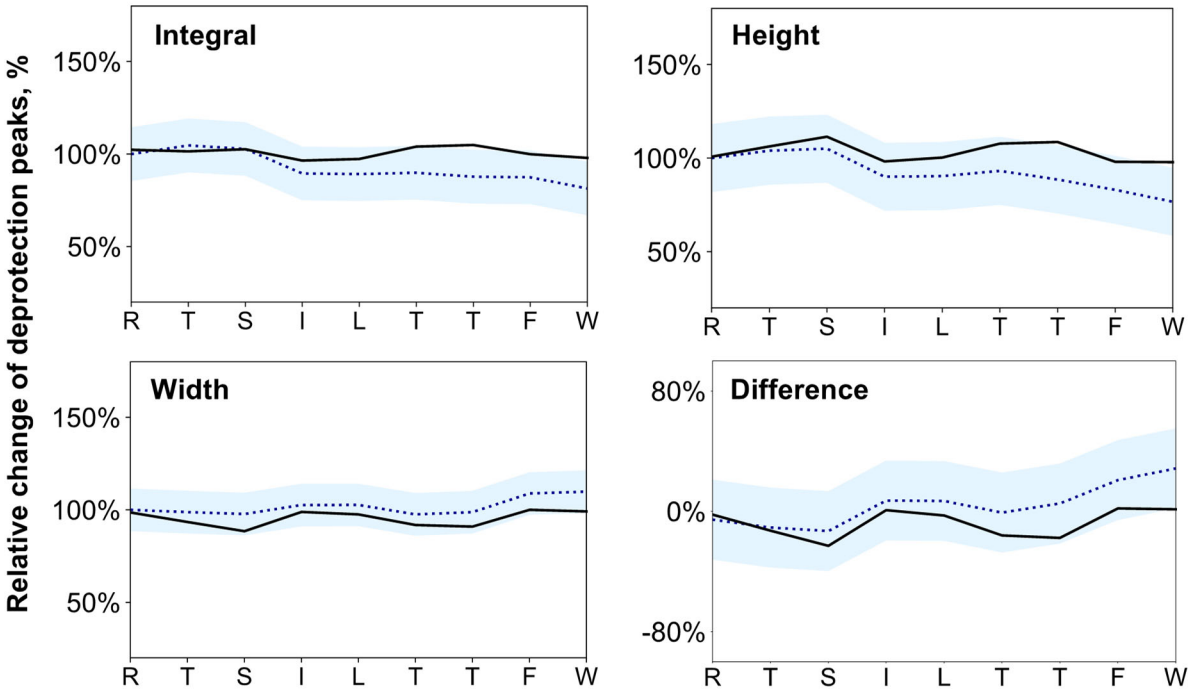
Synthesis Data for JR-10 (I9R)

Sequence: WFFTL ISTRM (10 AA)

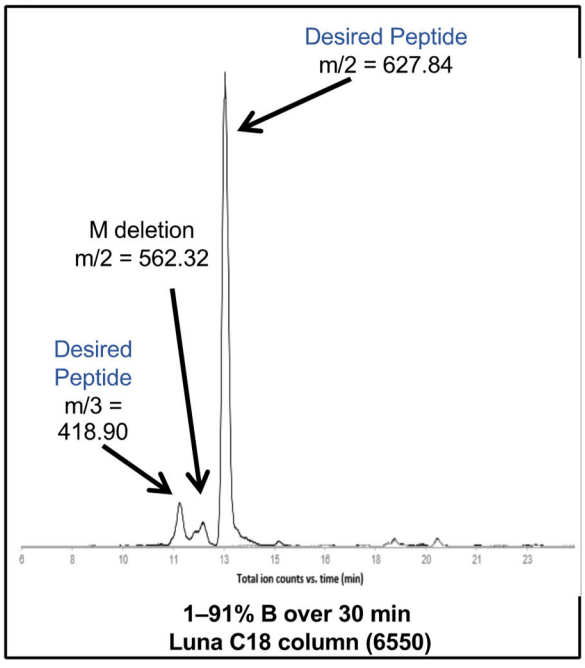
Resin: 100 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

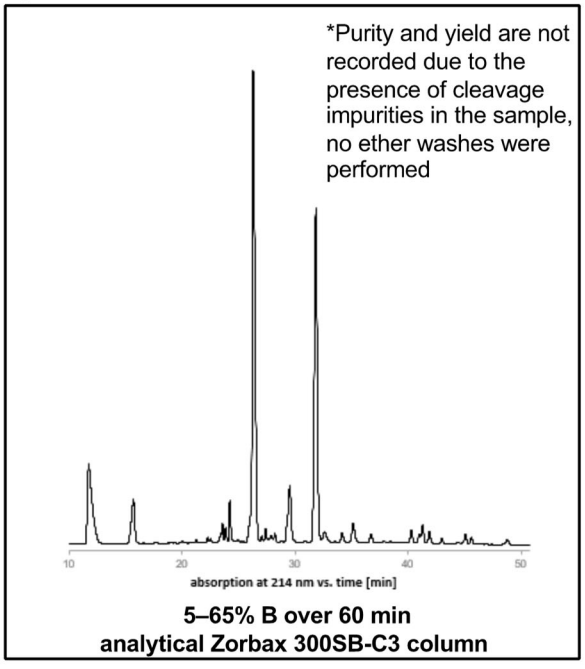


b) LCMS data, TIC



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

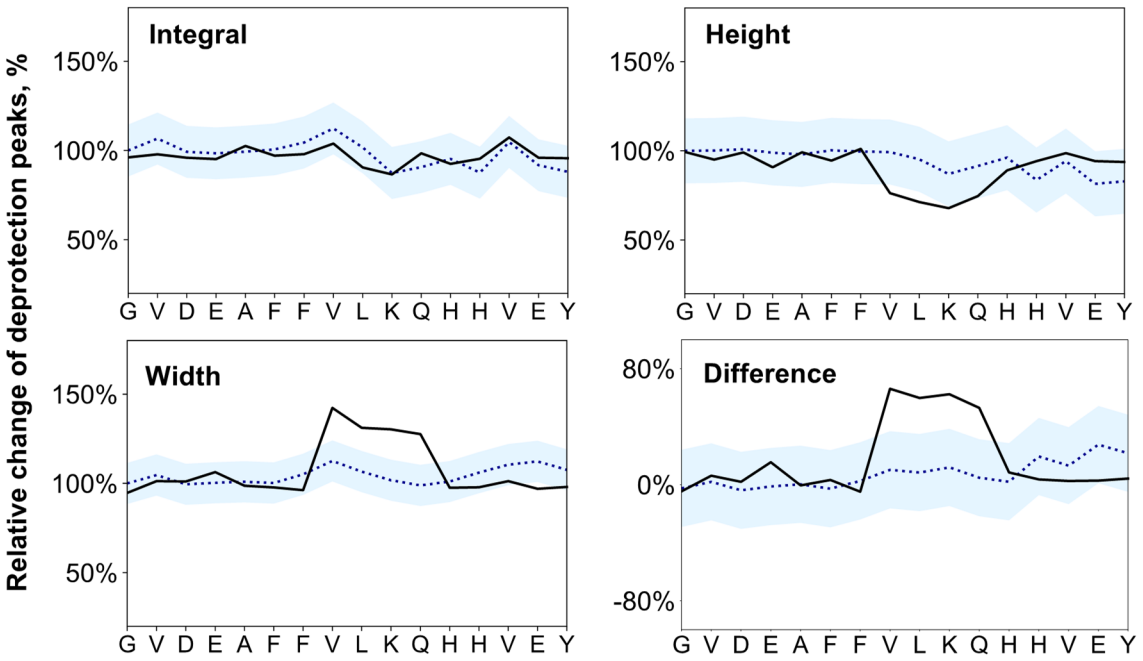
Synthesis Data for Aß[10-26]

Sequence: YEVHHQKLVFF AEDVGS (16 AA)

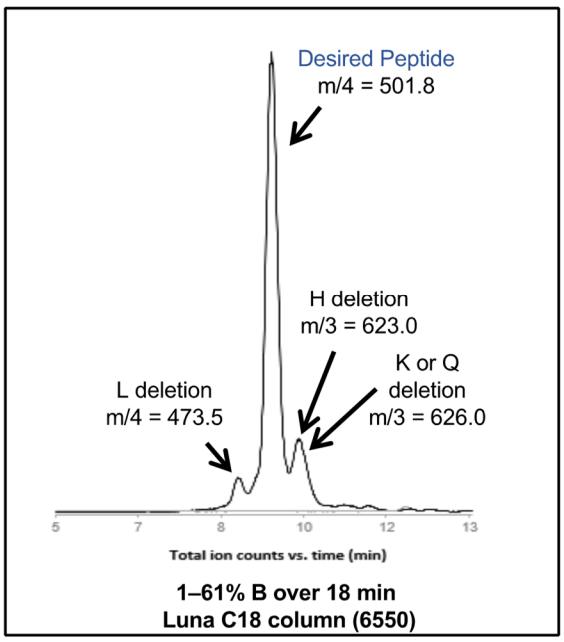
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 38 min

a) LCMS data, TIC

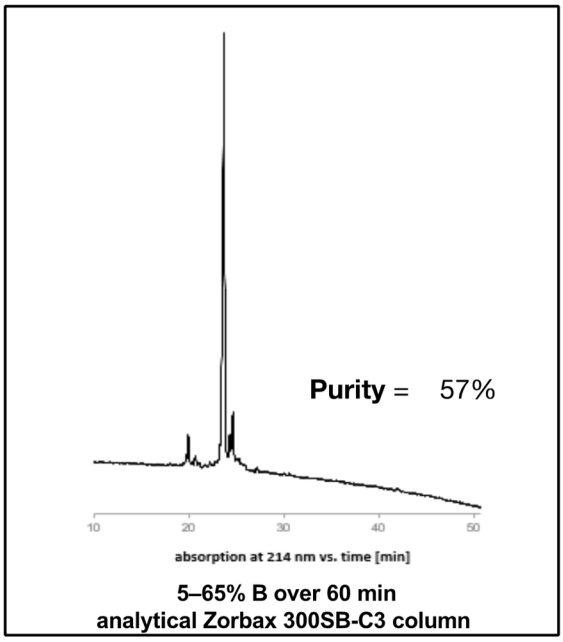


b) LCMS data, TIC



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

c) HPLC data



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

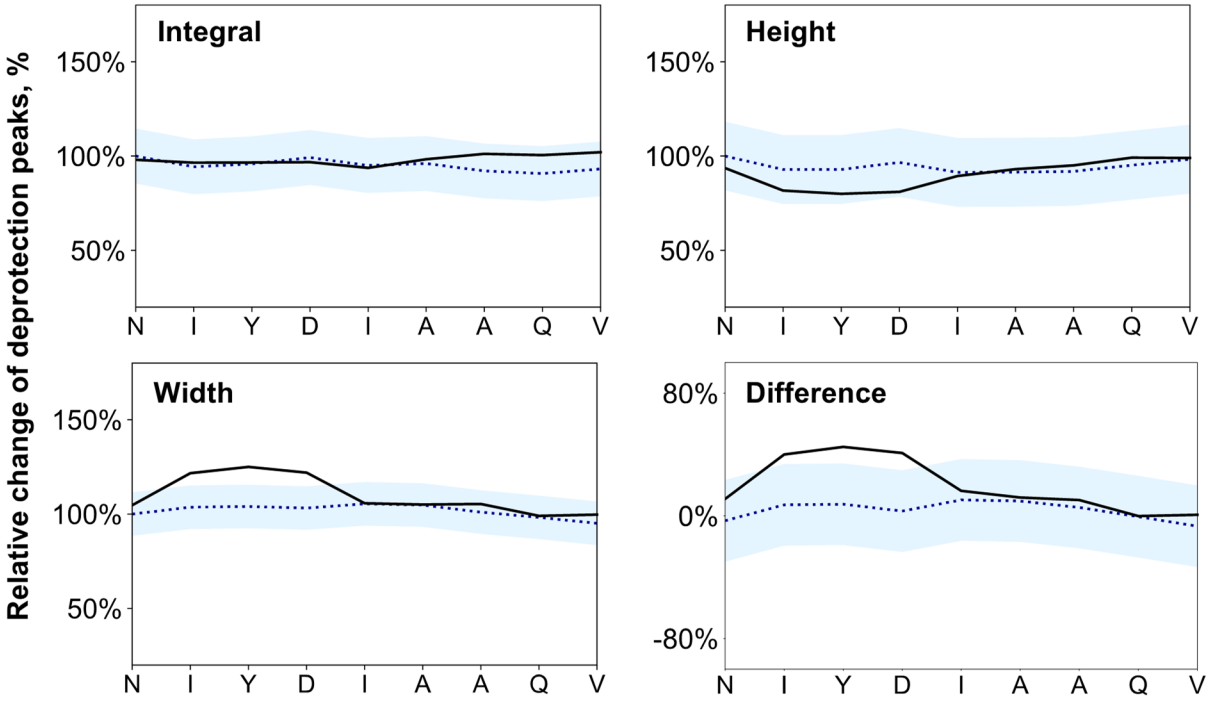
Synthesis Data for ACP[65-74]

Sequence: VQAAIDYING (10 AA)

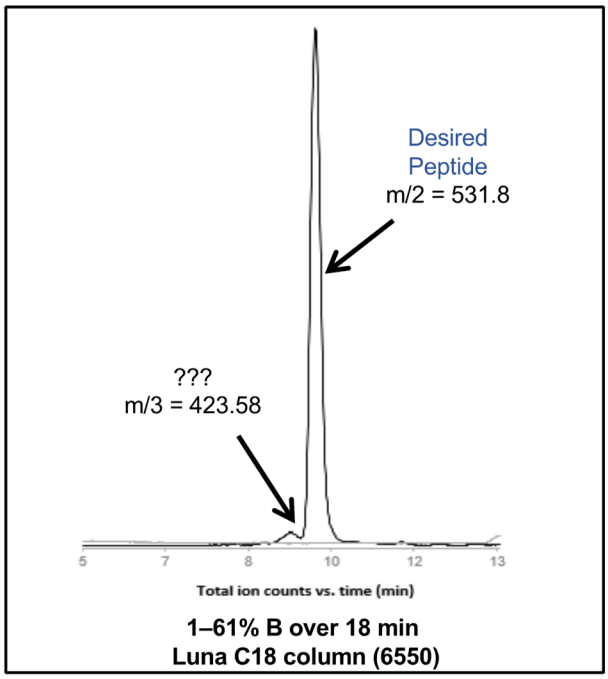
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 23 min

a) LCMS data, TIC

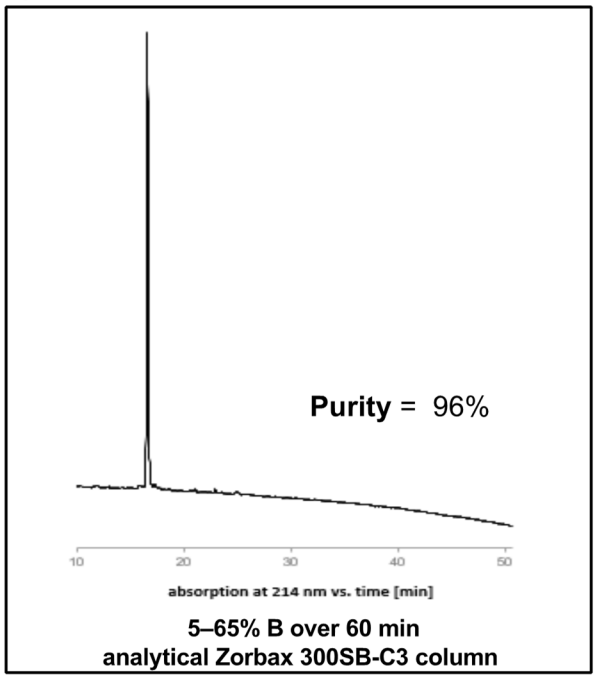


b) LCMS data, TIC



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

c) HPLC data



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

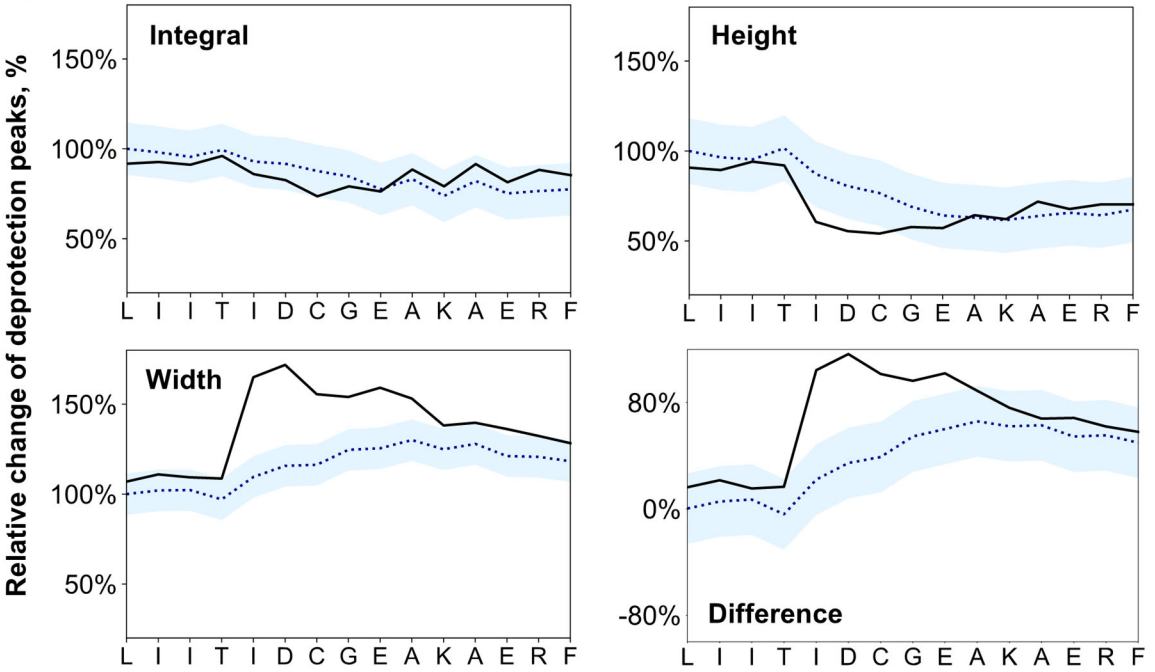
Synthesis Data for barstar[75-90]

Sequence: FREAKAEGCD ITIILS (16 AA)

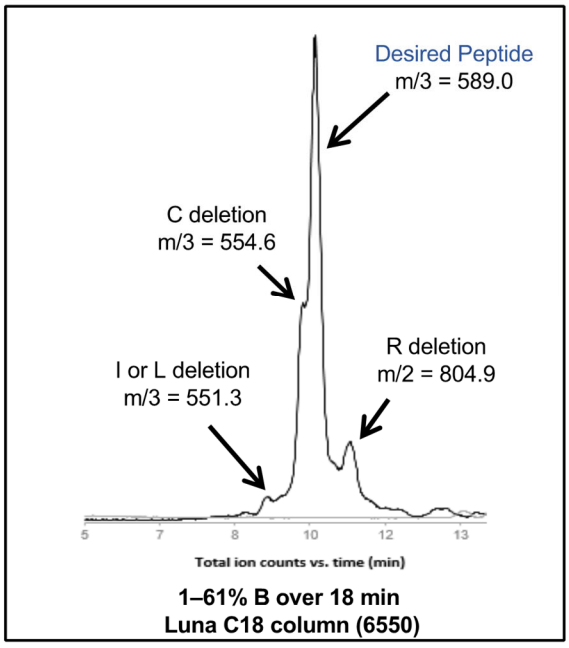
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 38 min

a) LCMS data, TIC

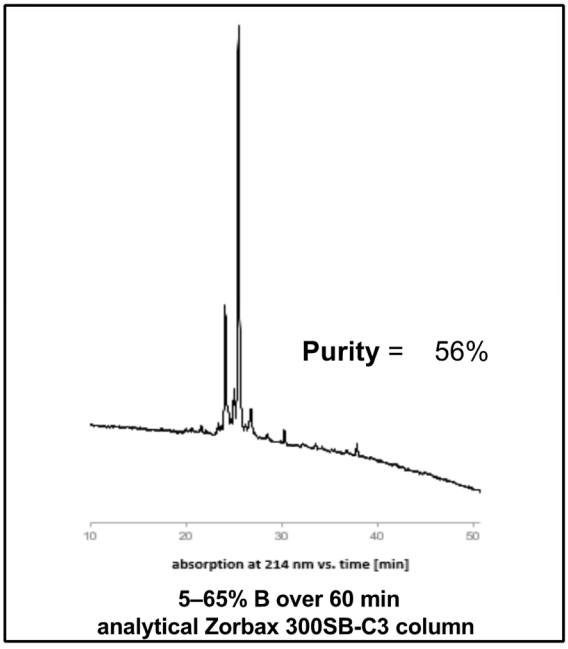


b) LCMS data, TIC



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

c) HPLC data



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

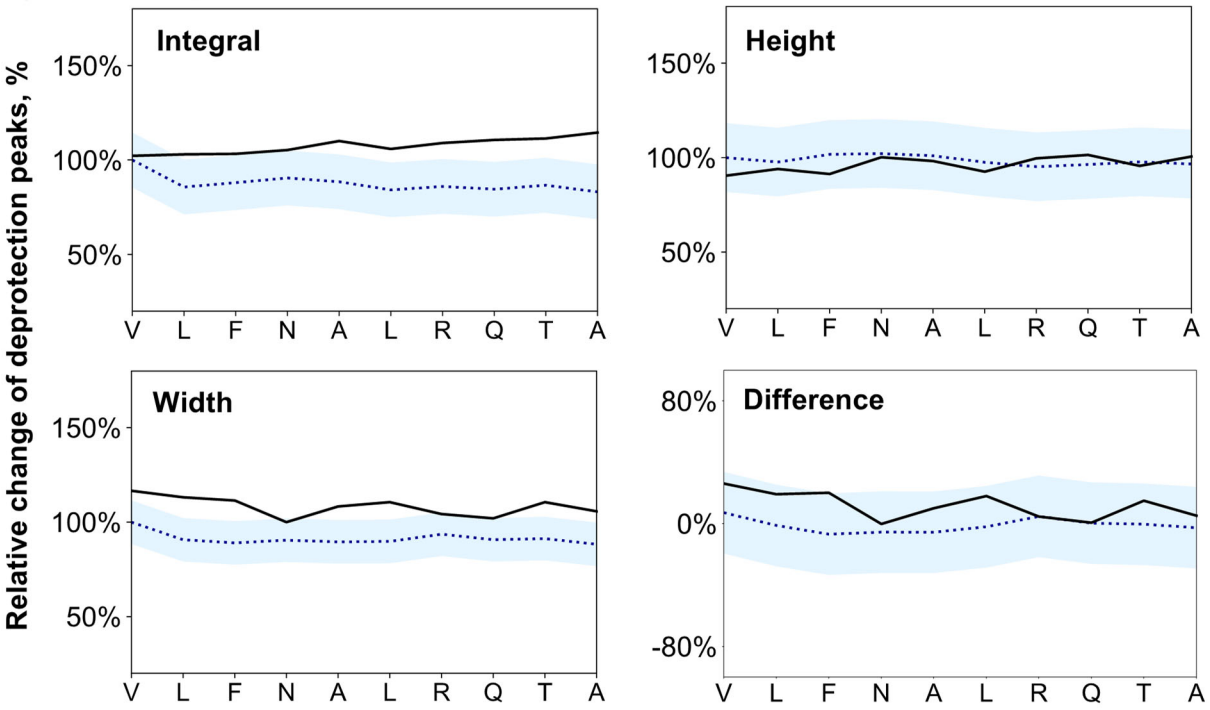
Synthesis Data for IAPP[1-18]

Sequence: ATQRLANFLV H (11 AA)

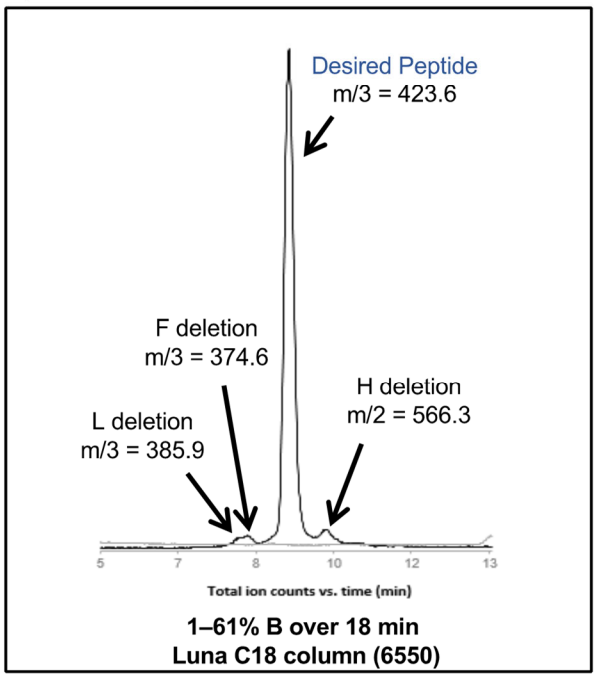
Resin: 80 mg of RINK amide ChemMatrix® (0.49 mmol/g), yielding the C-terminal amide after cleavage

Synthesis time: 25 min

a) LCMS data, TIC

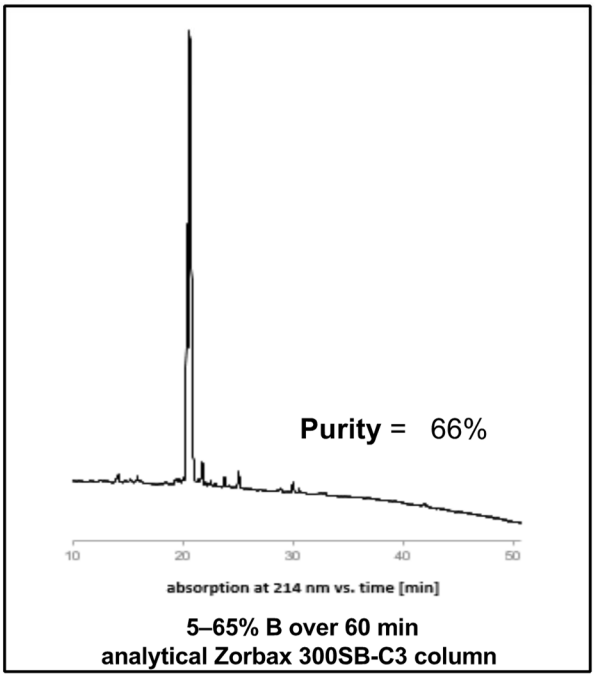


b) LCMS data, TIC



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?

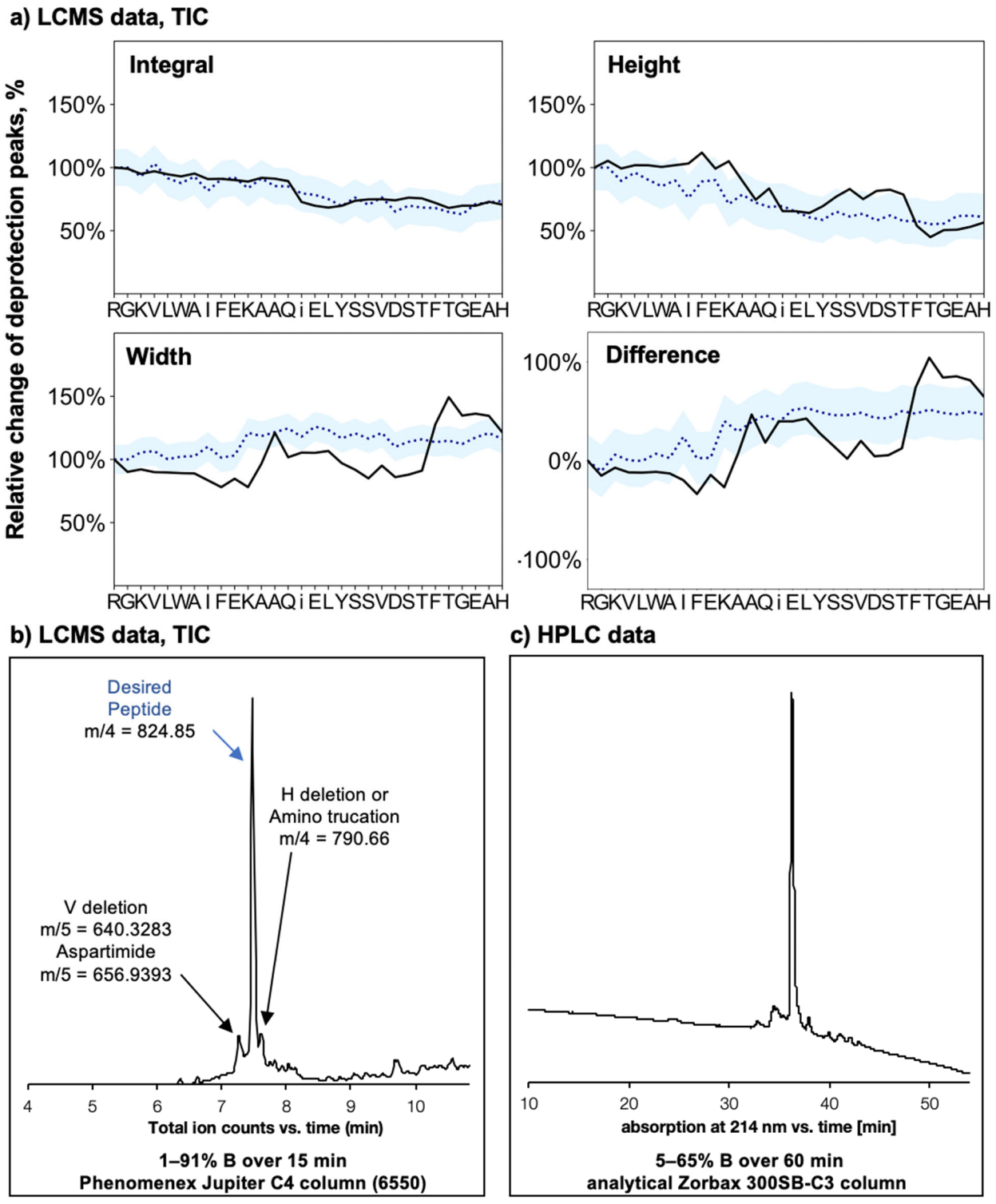
SI Table 3. Point mutations in GLP-1 with building blocks with are new to the model.

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

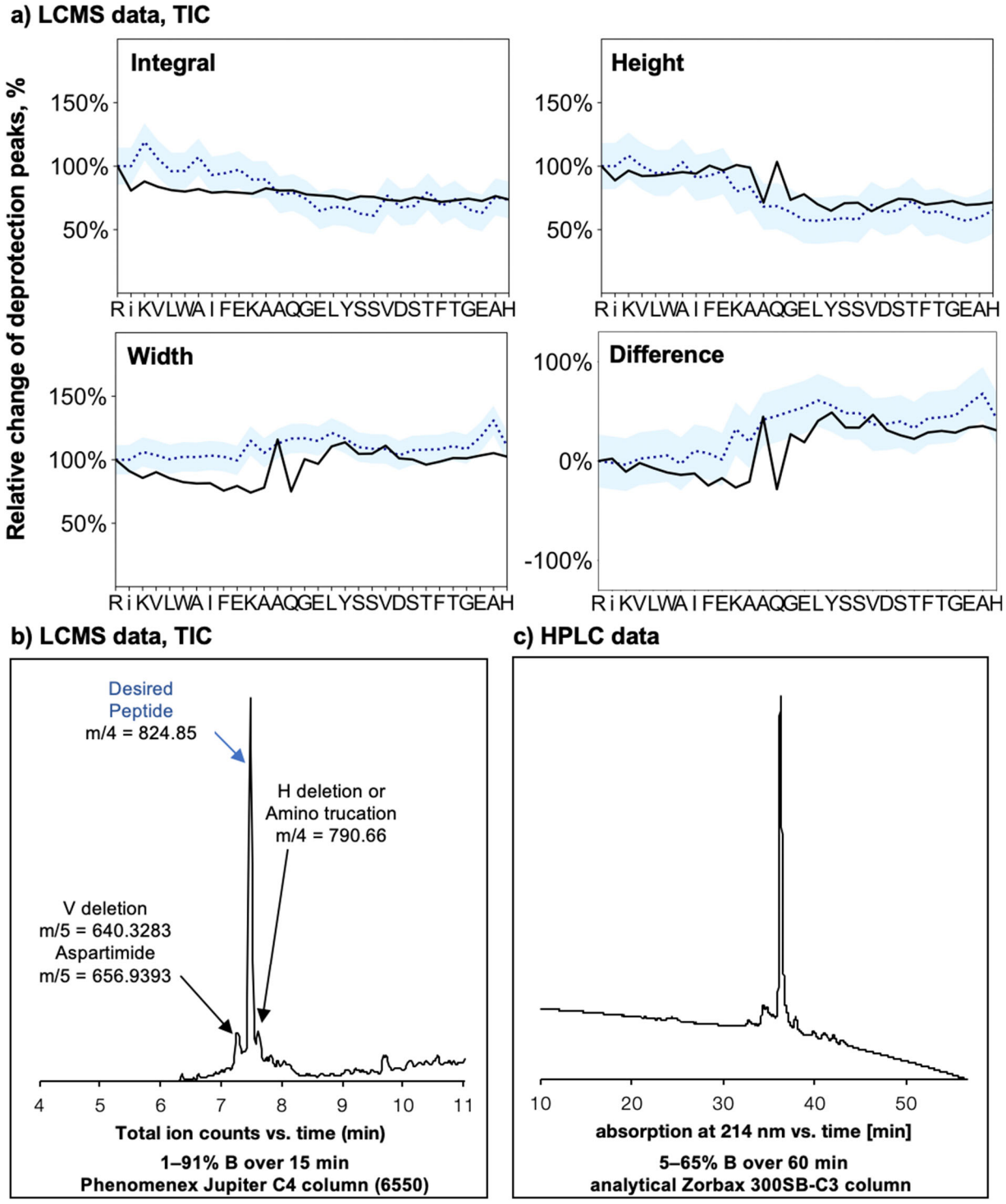


Synthesis Data for GLP-1 (Fmoc-(DMB)Gly-OH 2)

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

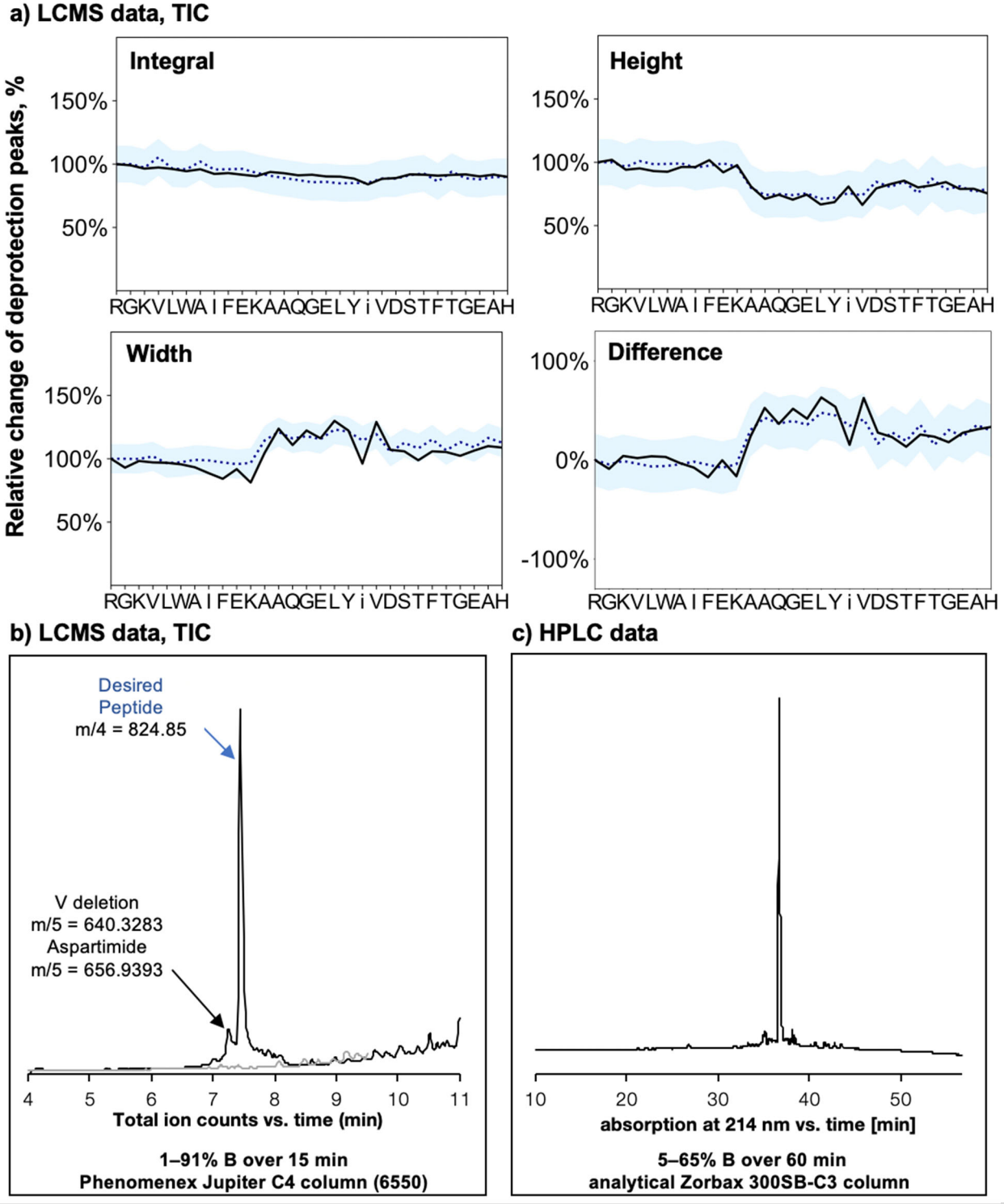


Synthesis Data for GLP-1 (Fmoc-Ser(*t*-Bu)-Ser($\Psi^{Me,Me}$pro)-OH)

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

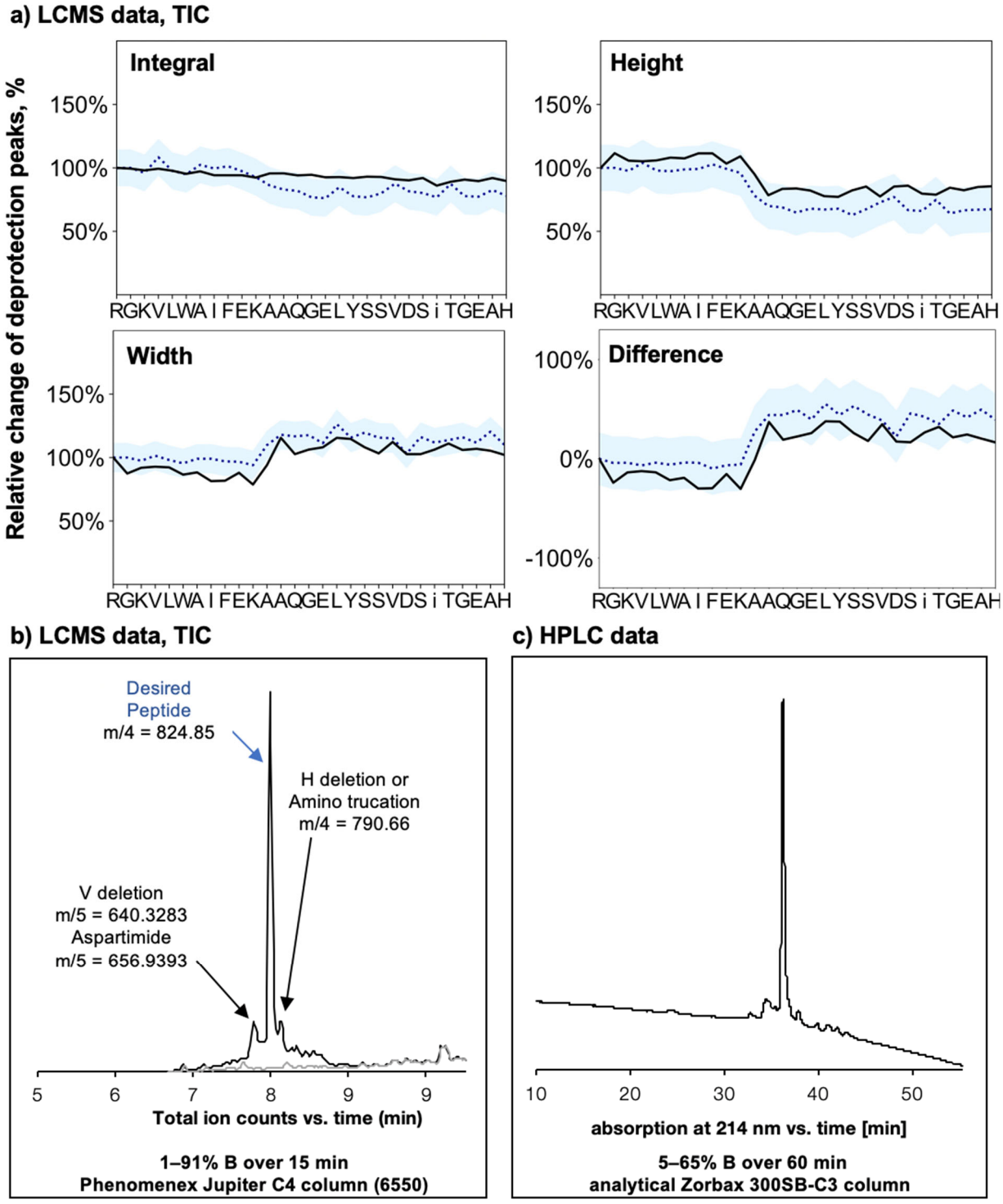


Synthesis Data for GLP-1 (Fmoc-Phe-Thr($\Psi^{Me,Me}$pro)-OH)

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

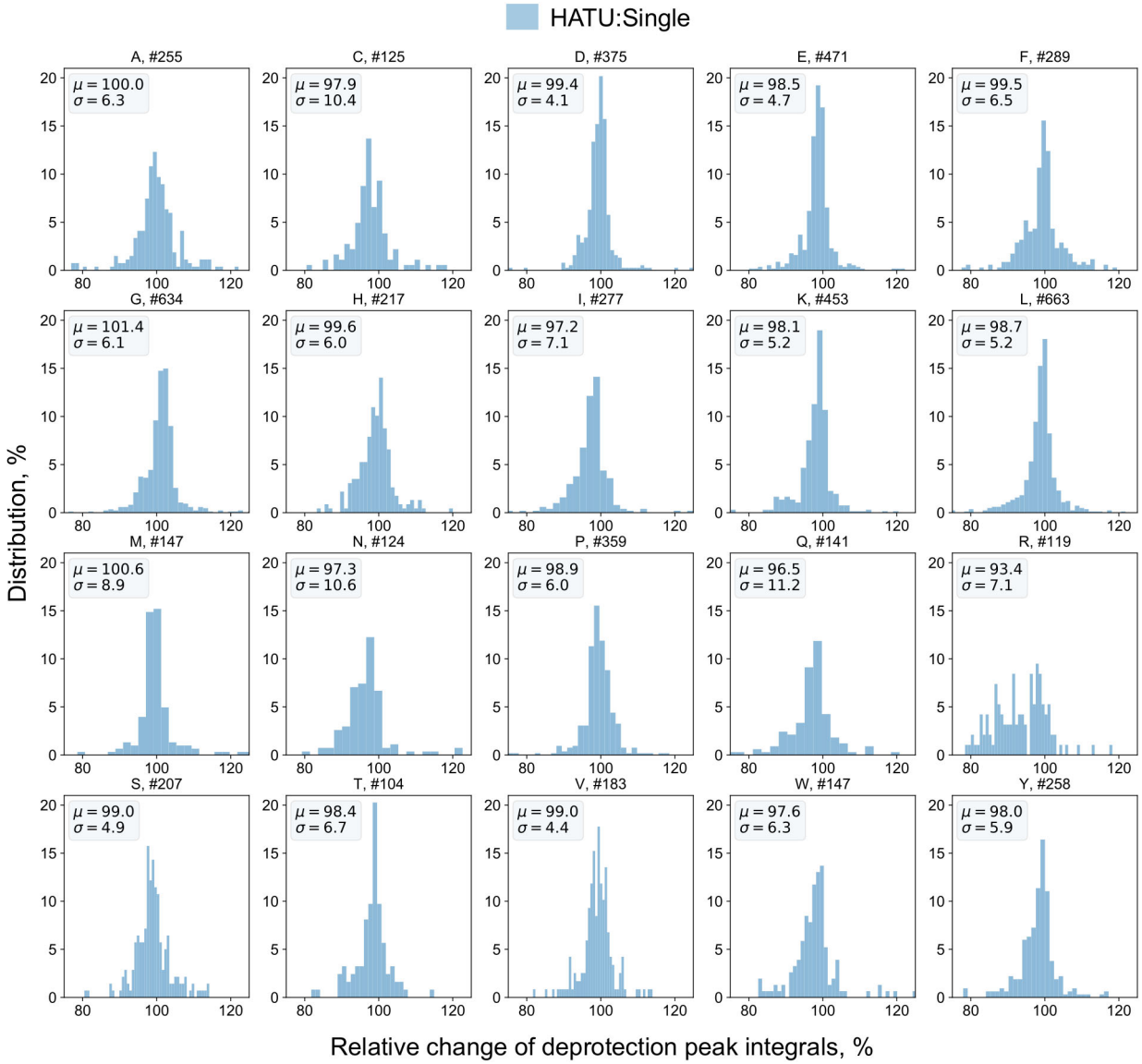
Synthesis time: 1.1 h



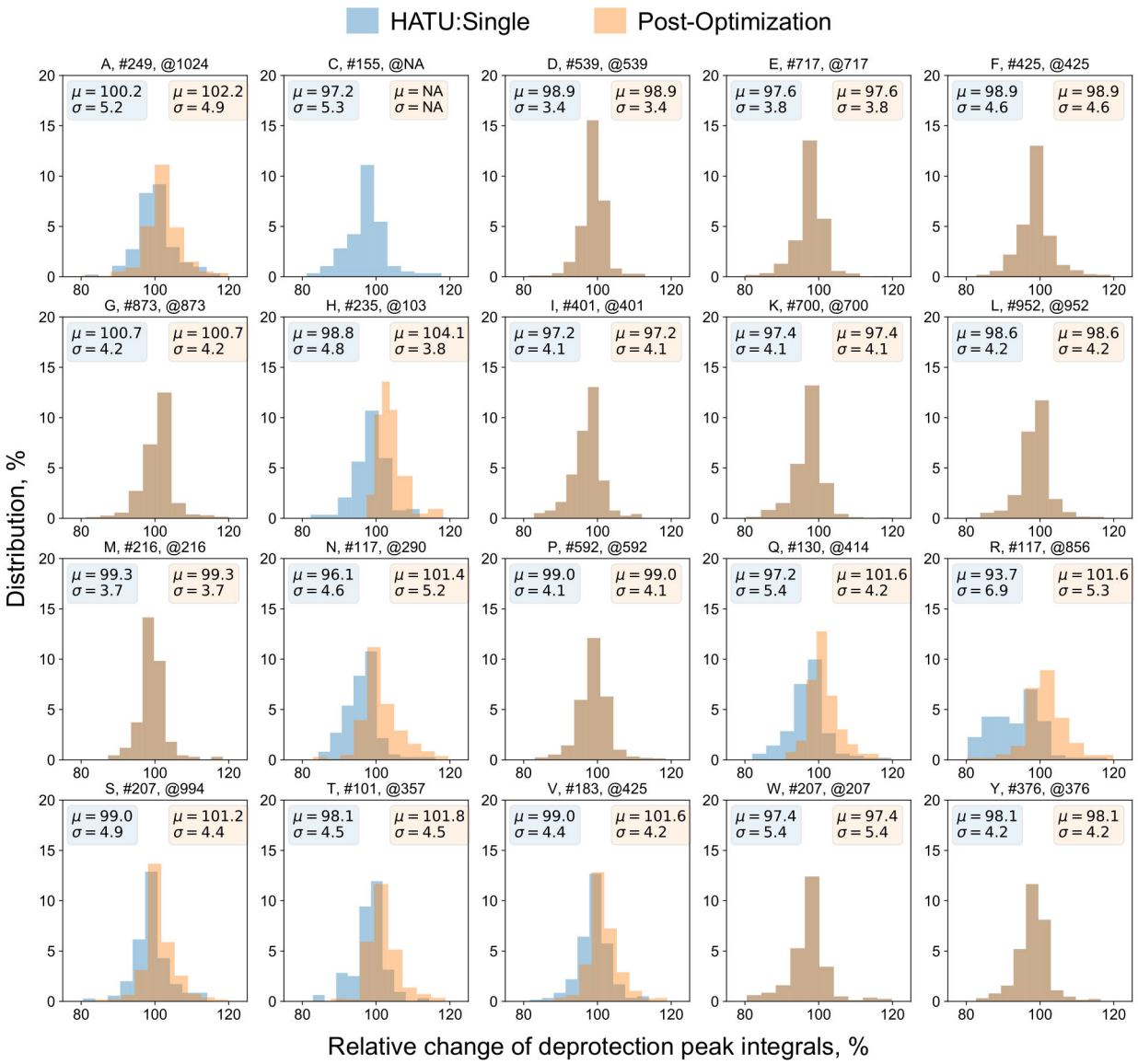
5 Statistical analysis of AFPS data set

5.1 Distribution of integrals for different synthesis parameters

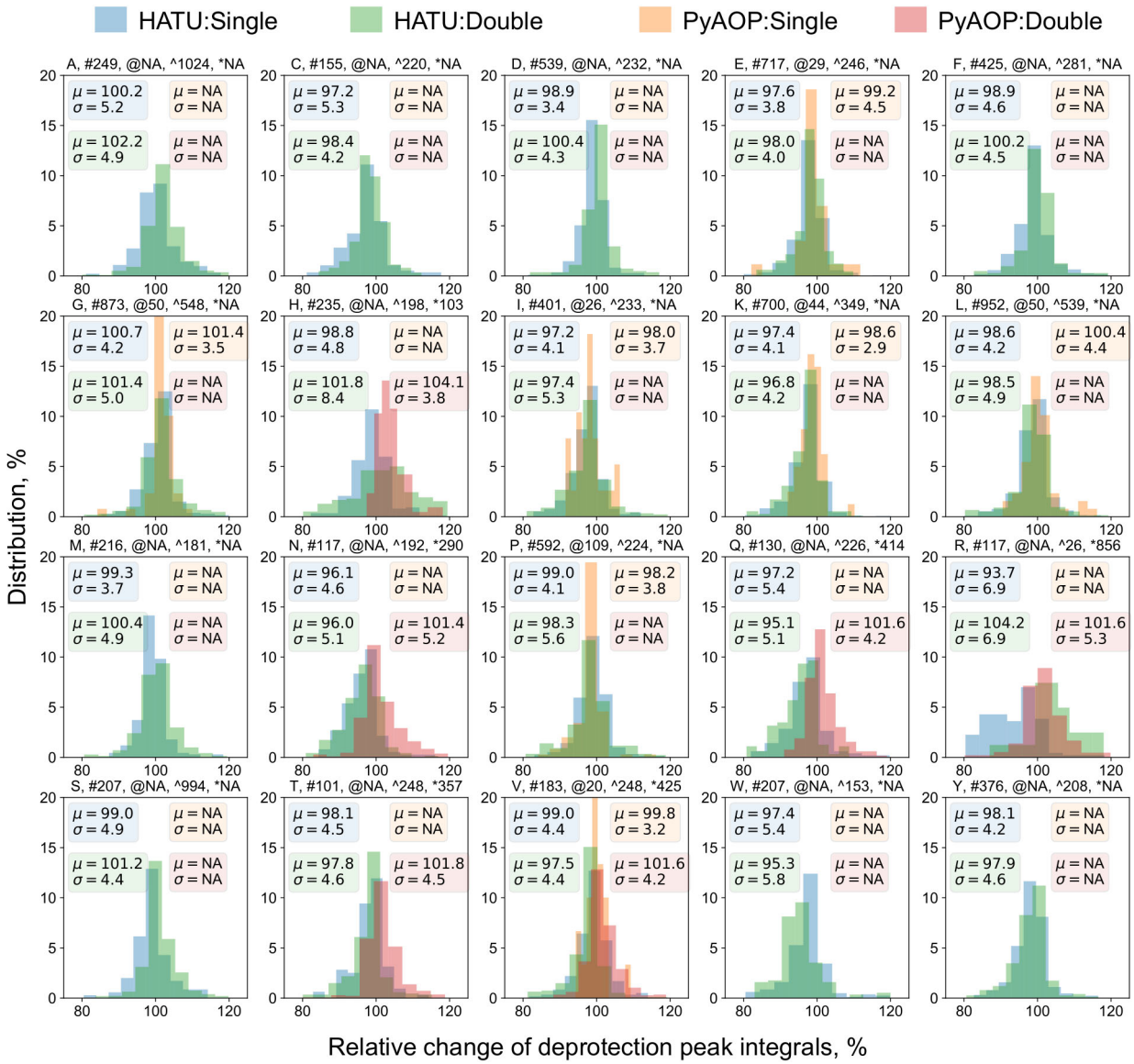
Deprotection peak integrals were analyzed for different combinations of synthesis parameters for coupling of each incoming amino acid (**SI Figure 3-5, Appendix 2**). On an average, PyAOP and double coupling strokes are seen to be more effective than HATU and single coupling strokes respectively. Apart 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 #.



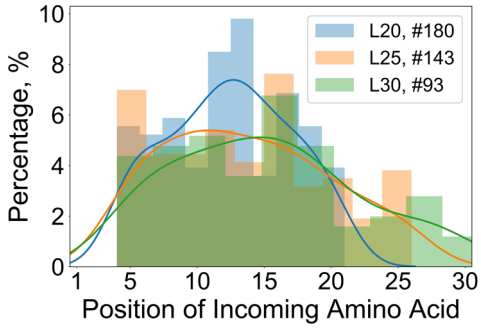
SI Figure 4. Distribution of deprotection peak integrals by amino acid for HATU coupling agent and single coupling stroke (blue) and after optimization parameters (red).(*3*) In the optimized synthesis protocol, A and S were coupled with HATU (double coupling) and H, N, Q, R, T and V were coupled with 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.

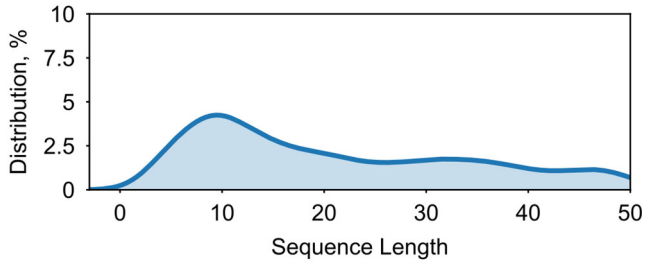


SI Figure 6. The onset of aggregation is analyzed for sequences of length (L) greater than 20, 25 and 30. # followed by the numerical quantity indicates the number of sequences in the dataset with the unique pre-chain where aggregation starts.

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).(*10*) 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**).



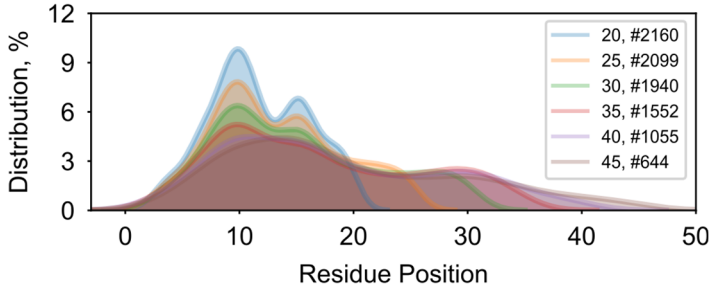
SI Figure 7. Distribution of sequences for different sequence lengths.

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

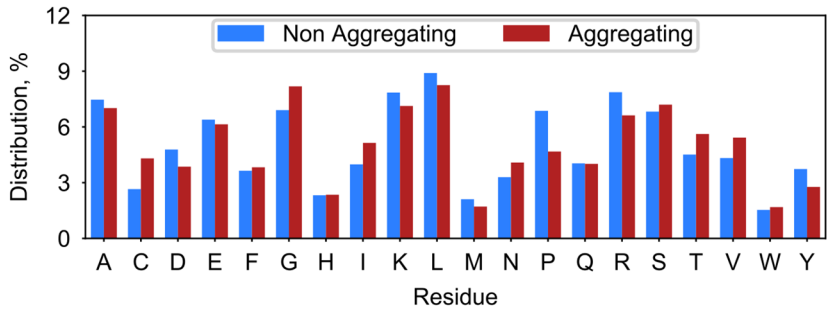
Onset of aggregation for the PDB sequences was calculated for sequences with different minimum lengths, similar to SI Section 5.1 (**SI Figure 8**).



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.

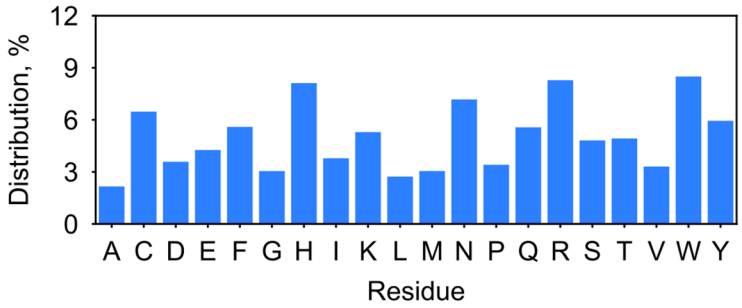


SI Figure 9. Distribution of amino acids in non-aggregating sequences and pre-chain at the aggregating step of aggregating sequences is similar.

6.5 Activation analysis

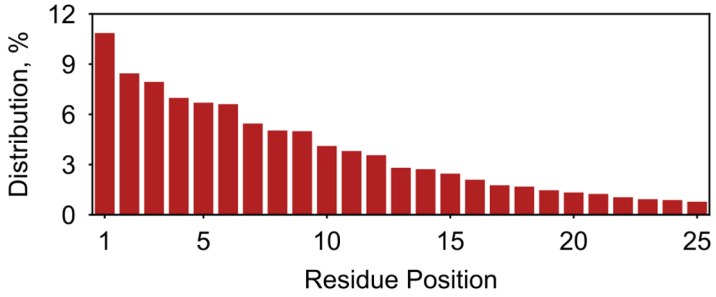
Activation maps for pre-chains of all aggregating sequences were calculated. For analysis, sum of all activations by position, residue and bit-indices for each residue was done for sequences with greater than equal to 25 residues.

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

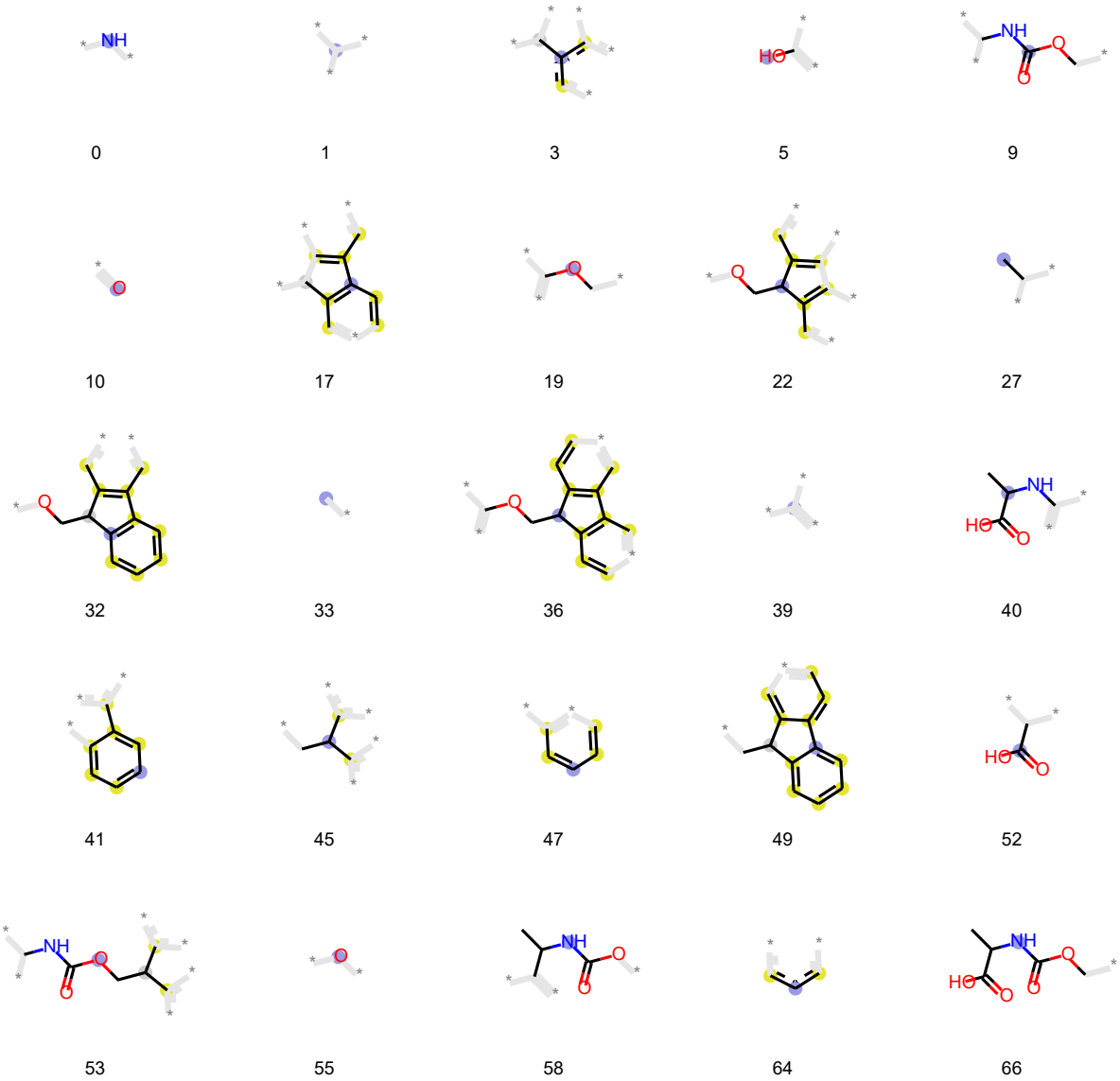
1. M. D. Simon, thesis, Massachusetts Institute of Technology (2017).
2. A. J. Mijalis, D. A. Thomas III, M. D. Simon, A. Adamo, R. Beaumont, K. F. Jensen, B. L. Pentelute, A fully automated flow-based approach for accelerated peptide synthesis. *Nat. Chem. Biol.* **13**, 464 (2017).
3. N. Hartrampf, A. Saebi, M. Poskus, Z. P. Gates, A. J. Callahan, A. E. Cowfer, S. Hanna, S. Antilla, C. K. Schissel, A. J. Quartararo, X. Ye, A. J. Mijalis, M. D. Simon, A. Loas, C. Jessen, T. E. Nielsen, B. L. Pentelute, Synthesis of Proteins by Automated Flow Chemistry. *ChemRxiv. Prepr.* (2020), doi:10.26434/chemrxiv.11833503.v1.
4. H. L. Morgan, The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107–113 (1965).
5. G. Landrum, RDKit: Open-source cheminformatics. *RDKit: Open-source cheminformatics* (2006).
6. S. Clark, P. Hayes, SigOpt WebPage. *SigOpt Web page* (2019).
7. C. K. Schissel, S. Mohapatra, J. M. Wolfe, C. M. Fadzen, Interpretable Deep Learning for De Novo Design of Cell-Penetrating Abiotic Polymers. *bioRxiv* (2020) (available at https://doi.org/10.1101/2020.04.10.036566).
8. J. M. Collins, K. A. Porter, S. K. Singh, G. S. Vanier, High-efficiency solid phase peptide synthesis (he -Spps). *Org. Lett.* **16**, 940–943 (2014).
9. E. Atherton, V. Woolley, R. C. Sheppard, Internal association in solid phase peptide synthesis. Synthesis of cytochrome C residues 66-104 on polyamide supports. *J. Chem. Soc. Chem. Commun.*, 970–971 (1980).
10. H. M. Berman, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

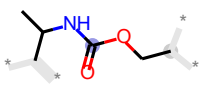
8 Appendix 1

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



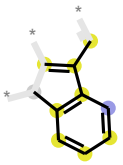


67

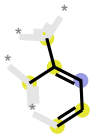
80

81

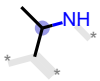
86



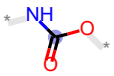
87



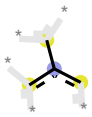
88



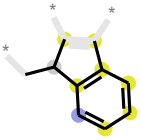
93



94

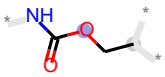


96



97

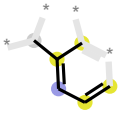
100



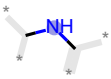
109

110

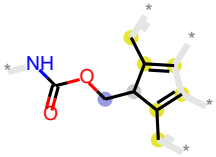
112



115



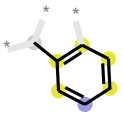
117



120

123

125



127

Cysteine

0

1

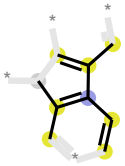
2

3

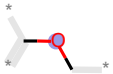
5

9

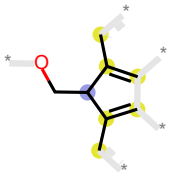
10



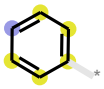
17



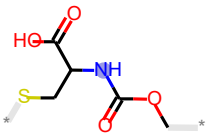
19



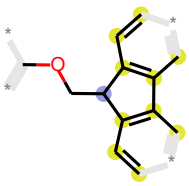
22



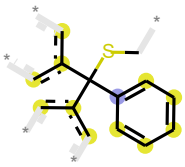
27



32

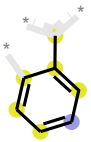


36

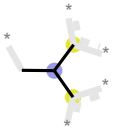


37

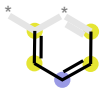
39



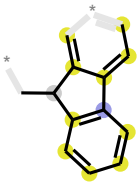
41



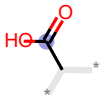
45



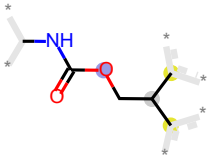
47



49

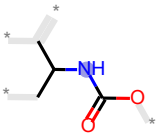


52

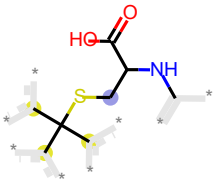


53

55

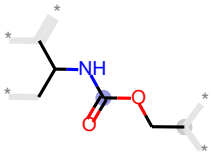


57

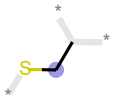


61

64



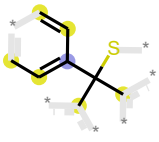
66



69

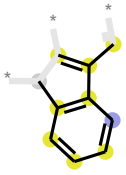
80

81

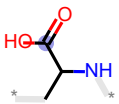


83

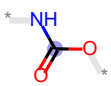
86



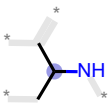
87



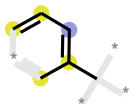
88



94

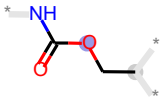


96

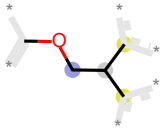


97

100



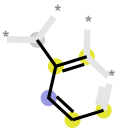
109



110

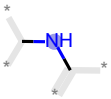
112

114

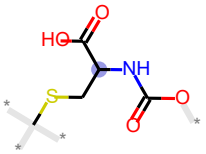


115

116



117

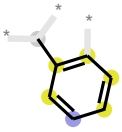


120

123

124

125



127

Aspartic acid

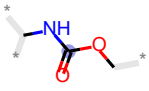
0

1

3

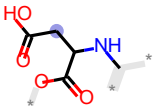
5

8

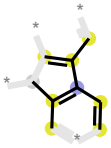


9

10

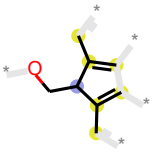


16

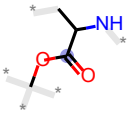


17

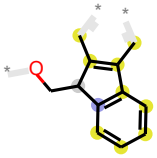
19



22

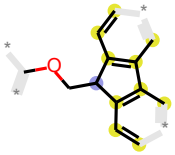


28

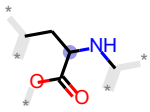


32

33



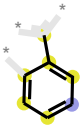
36



37

38

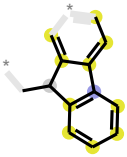
39



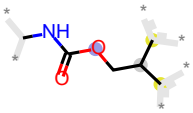
41

45

47

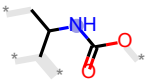


49



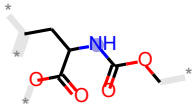
53

55



57

64



66

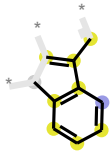
67

73

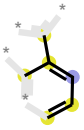
80

81

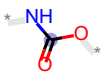
86



87

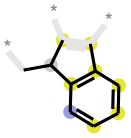


88



94

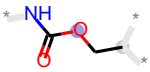
96



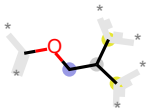
97

100

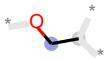
103



109

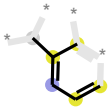


110

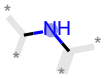


112

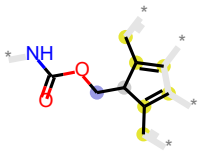
114



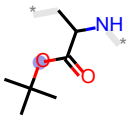
115



117



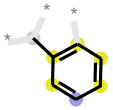
120



122

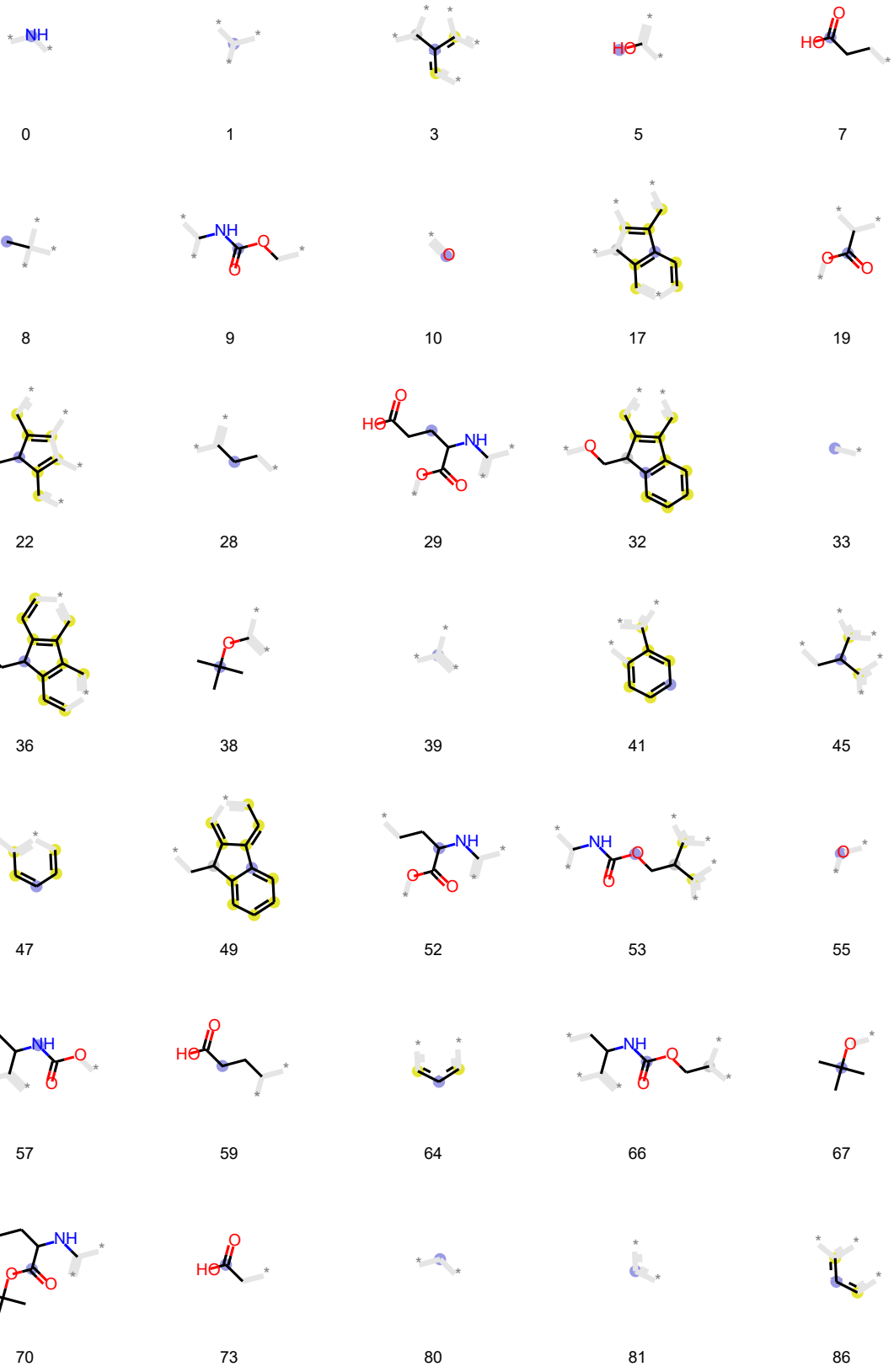
123

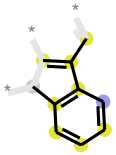
125



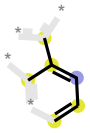
127

Glutamic acid

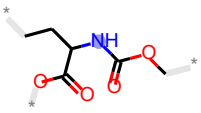




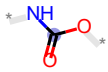
87



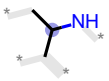
88



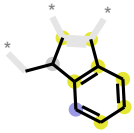
90



94



96

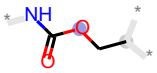


97

99

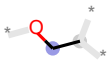
100

103



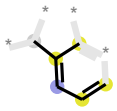
109

110

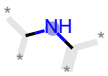


112

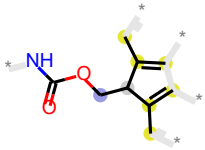
114



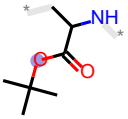
115



117



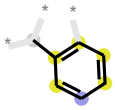
120



122

123

125



127

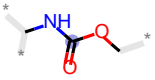
Phenylalanine

0

1

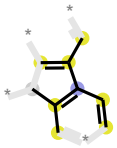
3

5



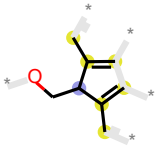
9

10



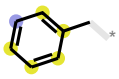
17

19

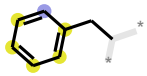


22

27

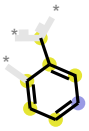


32

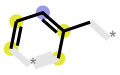


36

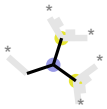
39



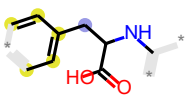
41



42

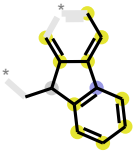


45



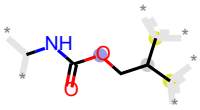
46

47



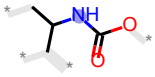
49

52



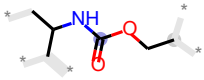
53

55



57

64

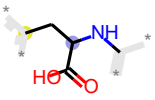


66

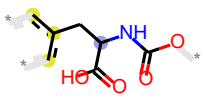
79

80

81

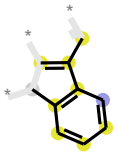


83

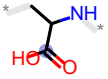


84

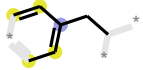
86



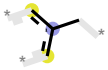
87



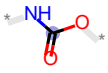
88



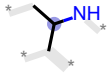
89



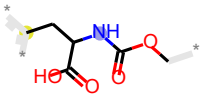
90



94

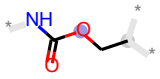


96

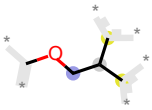


97

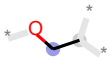
100



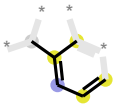
109



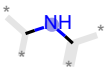
110



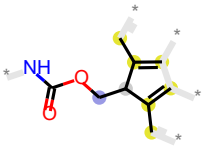
112



115



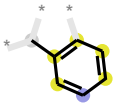
117



120

123

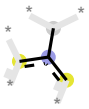
125



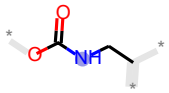
127

Glycine

0

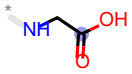


3



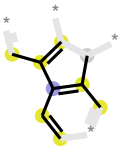
4

5

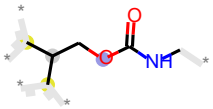


6

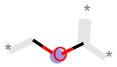
10



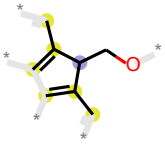
17



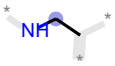
18



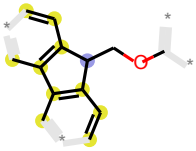
19



22

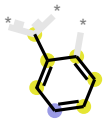


32

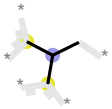


36

39

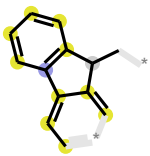


41



45

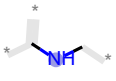
47



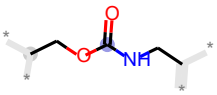
49

55

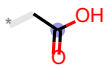
64



69



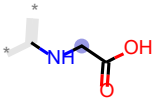
71



73

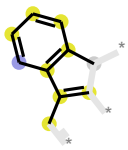
80

81

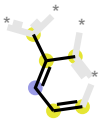


83

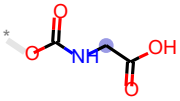
86



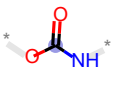
87



88

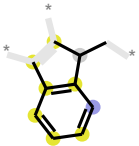


91



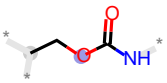
94

96

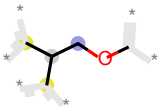


97

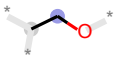
100



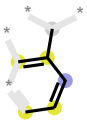
109



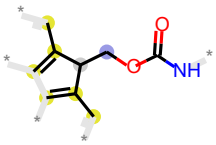
110



112



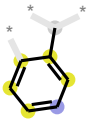
115



120

123

125

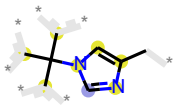


127

Histidine

0

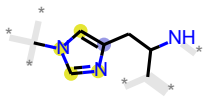
1



2

3

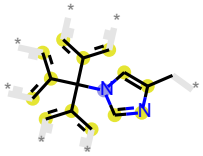
5



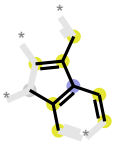
6

9

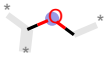
10



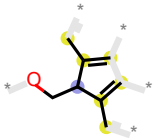
16



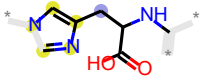
17



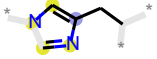
19



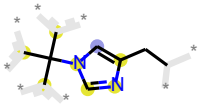
22



27



29

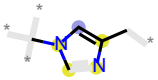


31

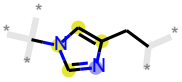
32

36

39

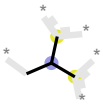


40

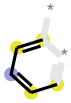


41

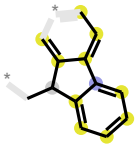
44



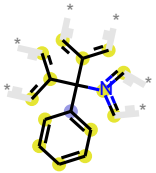
45



47

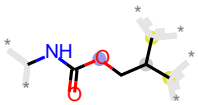


49



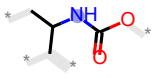
50

52

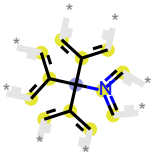


53

55

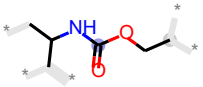


57

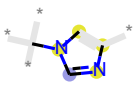


61

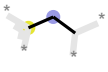
64



66



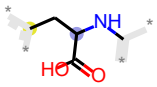
69



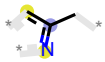
79

80

81

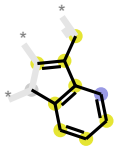


83

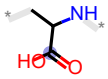


84

86

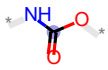


87

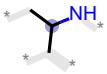


88

93



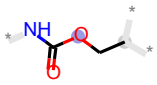
94



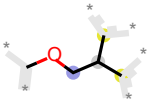
96

97

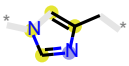
100



109



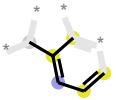
110



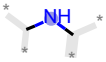
111

112

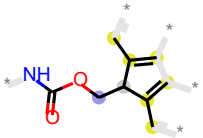
114



115



117

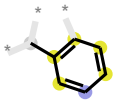


120

122

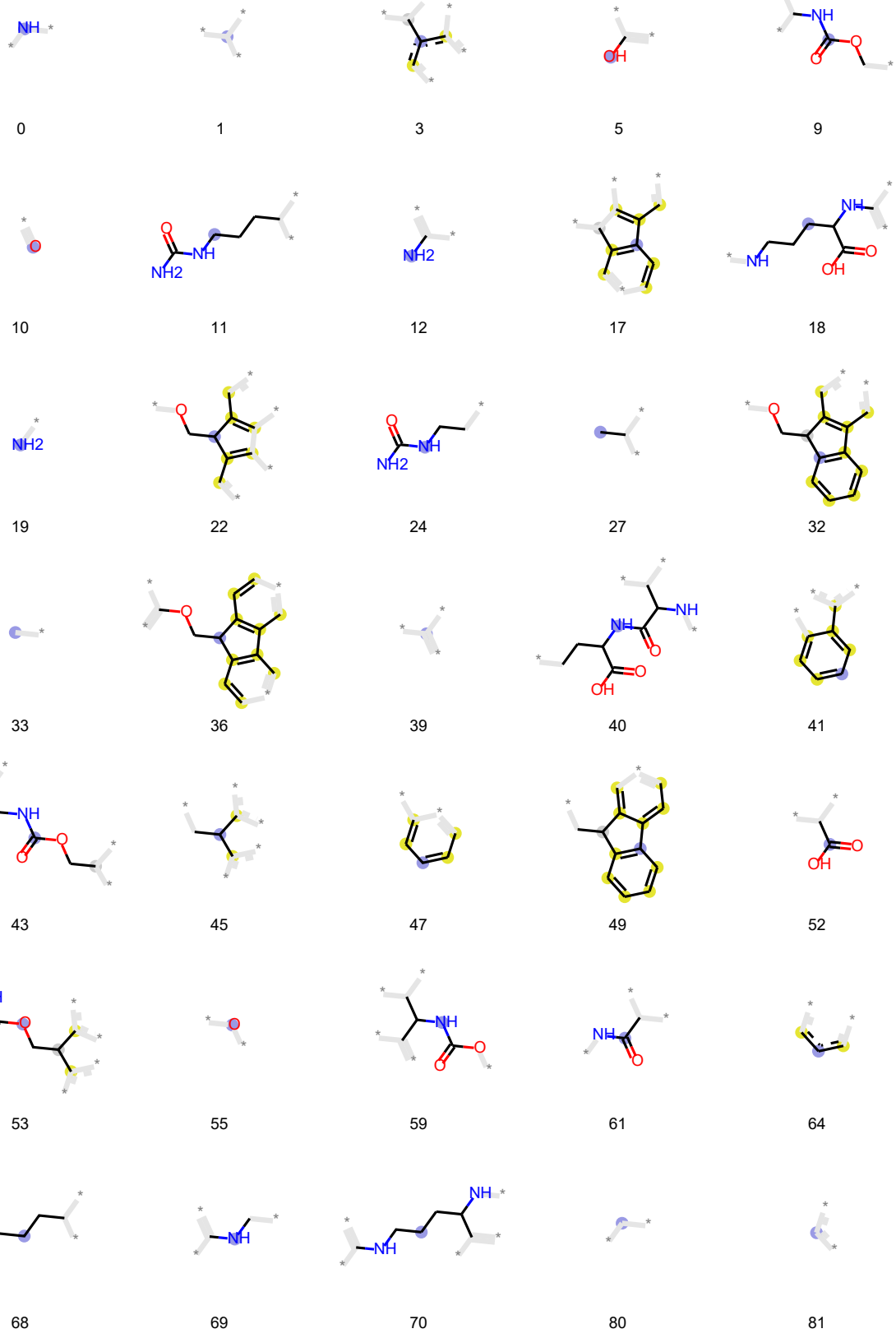
123

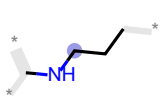
125



127

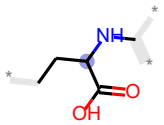
Isoleucine


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



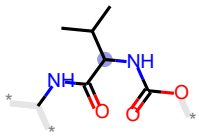
84

86

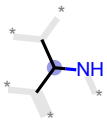


87

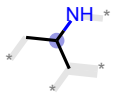
88



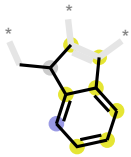
91



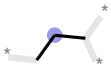
94



96



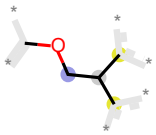
97



99

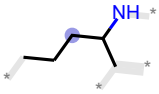
100

109

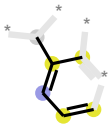


110

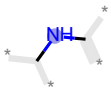
112



113



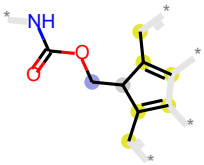
115



117

118

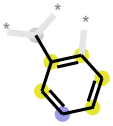
119



120

123

125

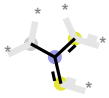


127

Lysine

0

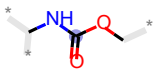
1



3

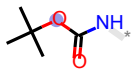
5

8

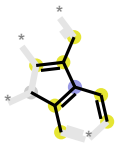


9

10

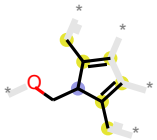


16

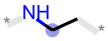


17

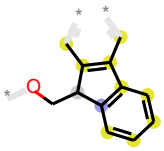
19



22

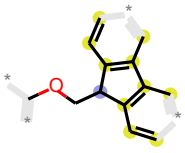


27

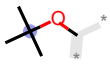


32

33

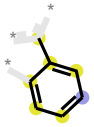


36

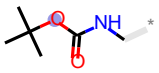


38

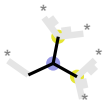
39



41

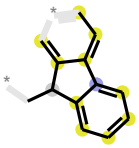


42



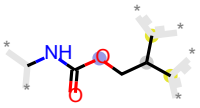
45

47



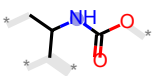
49

52



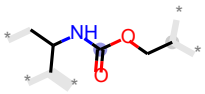
53

55



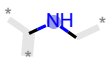
57

64



66

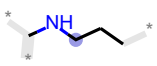
67



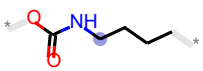
69

80

81

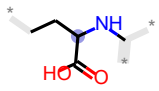


84

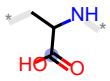


85

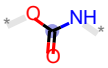
86



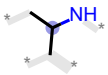
87



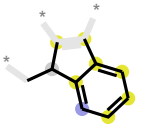
88



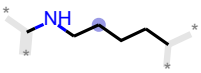
94



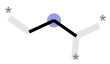
96



97



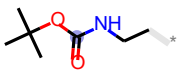
98



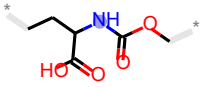
99

100

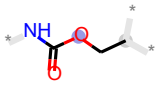
103



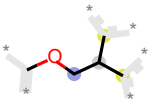
105



108

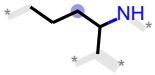


109



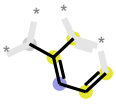
110

112

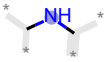


113

114

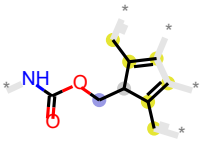


115



117

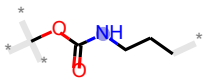
119



120

123

125

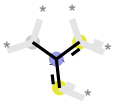


127

Leucine

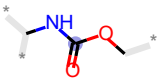
0

1



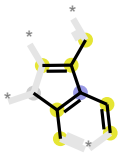
3

5



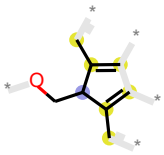
9

10



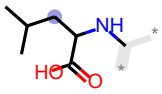
17

19

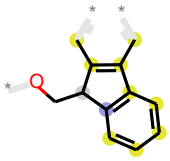


22

27

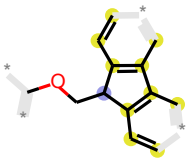


30



32

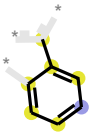
33



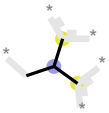
36

39

40

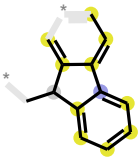


41



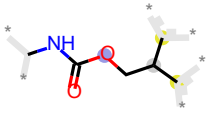
45

47



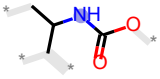
49

52



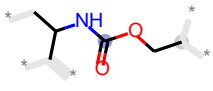
53

55

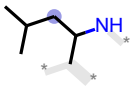


57

64



66

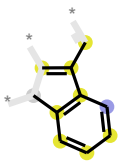


74

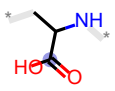
80

81

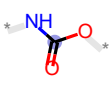
86



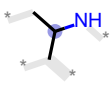
87



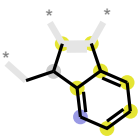
88



94

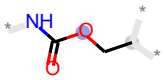


96

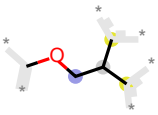


97

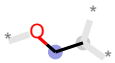
100



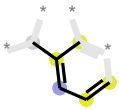
109



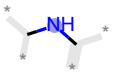
110



112

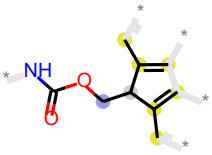


115



117

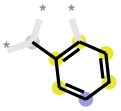
119



120

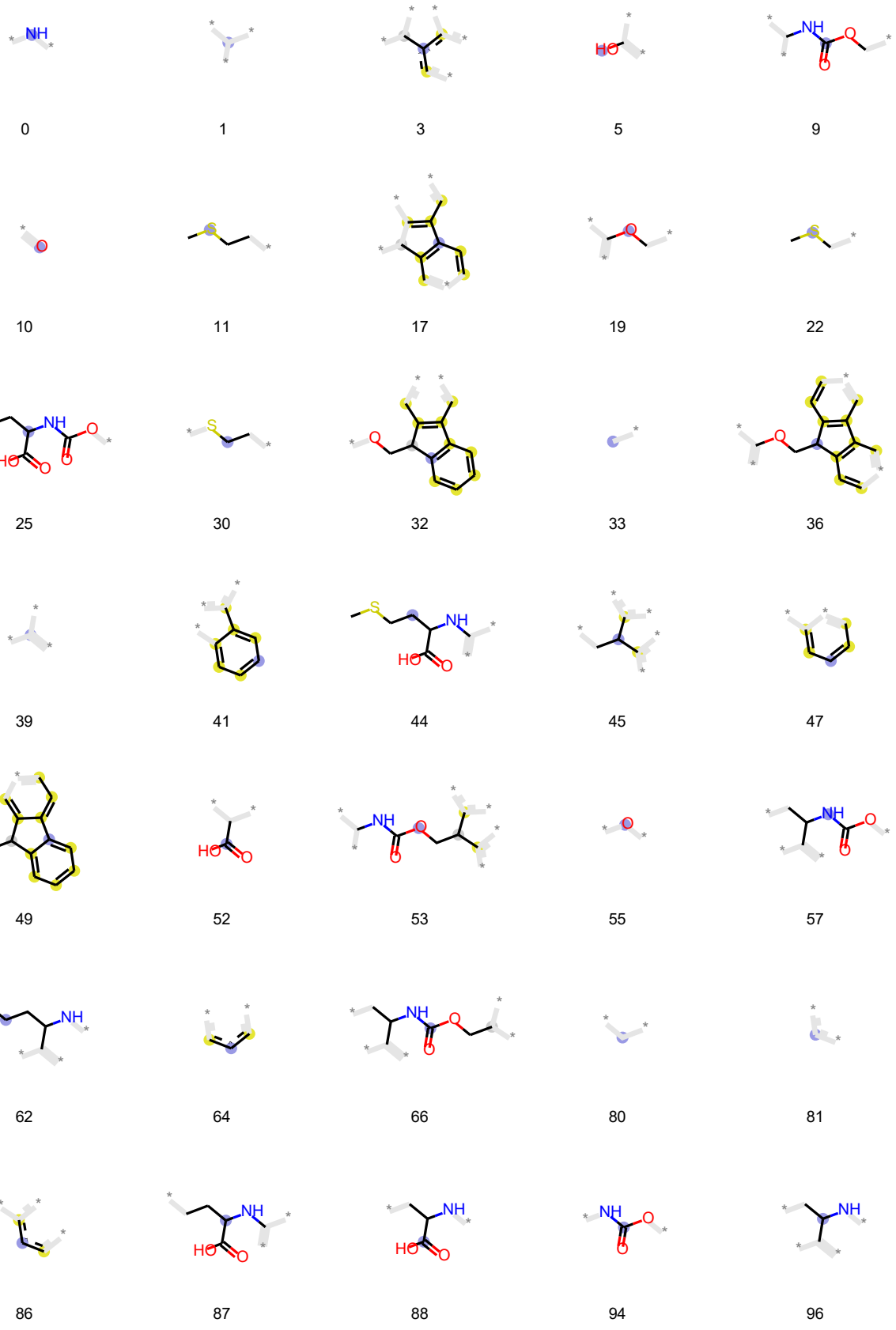
123

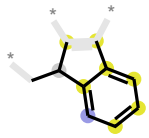
125



127

Methionine

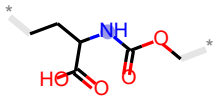

0
1
3
5
9
10
11
17
19
22
25
30
32
33
36
39
41
44
45
47
49
52
53
55
57
62
64
66
80
81
86
87
88
94
96



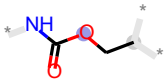
97

99

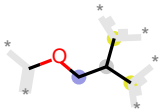
100



108

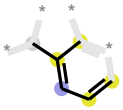


109



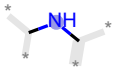
110

112

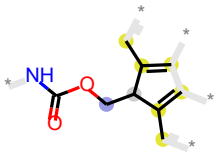


115

116



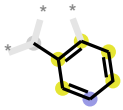
117



120

123

125



127

Asparagine

0

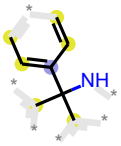
1

3

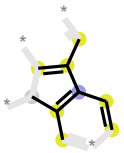
5

9

10

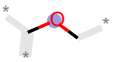


11

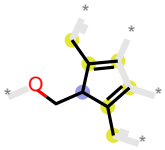


17

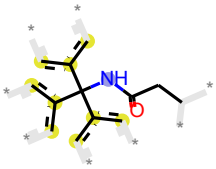
18



19

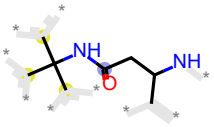


22

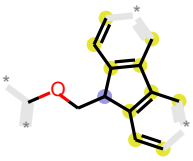


24

27



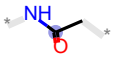
32



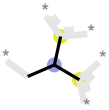
36

39

40

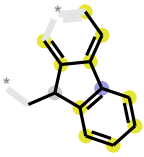


41



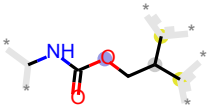
45

47



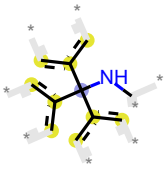
49

52

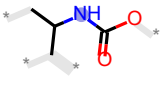


53

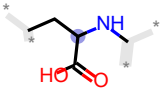
55



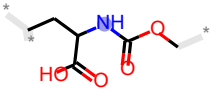
56



57

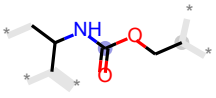


58

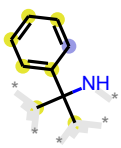


59

64



66

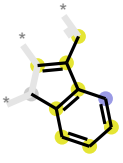


69

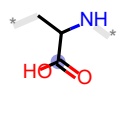
80

81

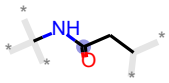
86



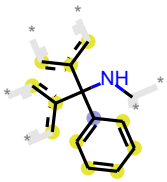
87



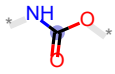
88



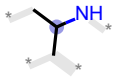
89



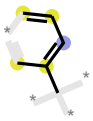
91



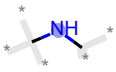
94



96



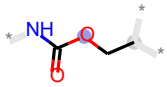
97



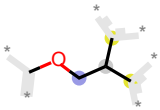
98

100

105



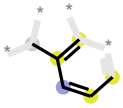
109



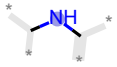
110

112

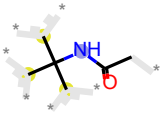
114



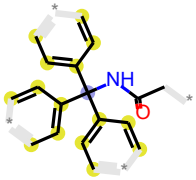
115



117



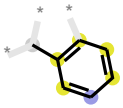
120



122

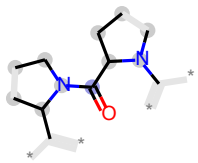
123

125

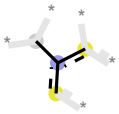


127

Proline



1

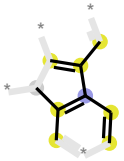


3

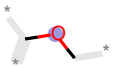
4

5

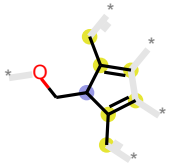
10



17



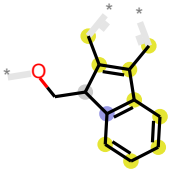
19



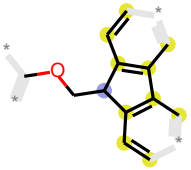
22

23

30

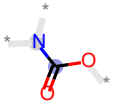


32

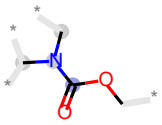


36

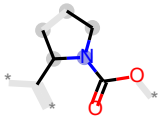
39



41



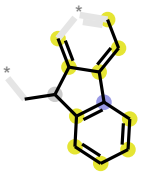
42



43

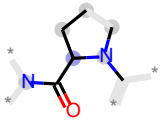
45

47



49

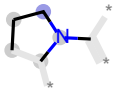
55



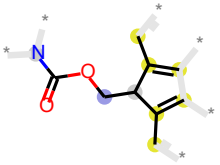
58

62

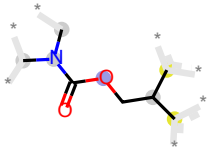
64



65

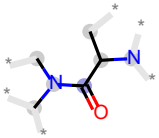


66



70

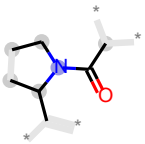
72



78

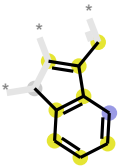
80

81

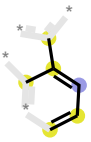


85

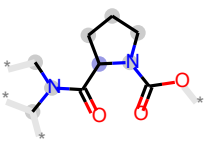
86



87

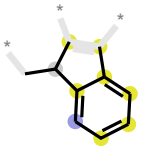


88

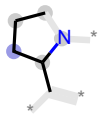


93

96

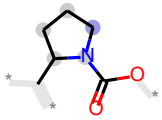


97

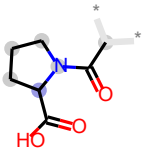


98

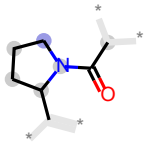
100



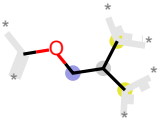
102



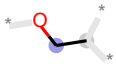
105



108

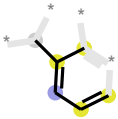


110



112

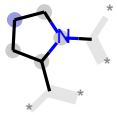
114



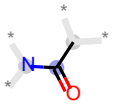
115

123

125



126

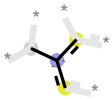


127

Glutamine

0

1

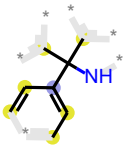


3

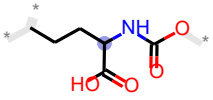
5

9

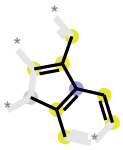
10



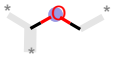
11



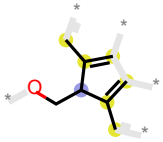
16



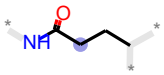
17



19

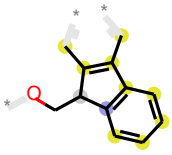


22

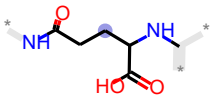


27

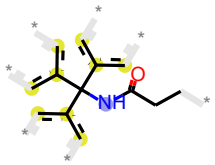
28



32

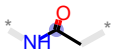


33

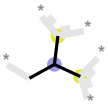


36

39



41

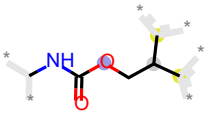


45

47

49

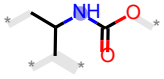
52



53

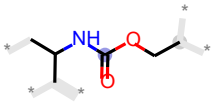
55

56

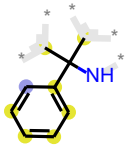


57

64



66

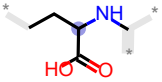


69

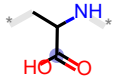
80

81

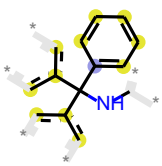
86



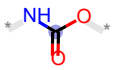
87



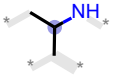
88



91

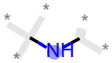


94

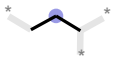


96

97

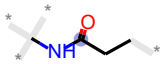


98



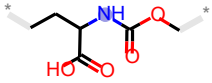
99

100

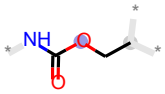


101

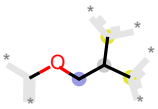
105



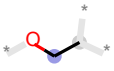
108



109

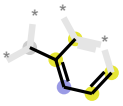


110

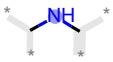


112

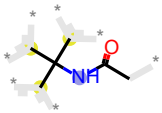
114



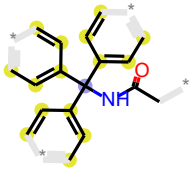
115



117



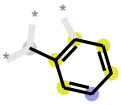
120



122

123

125



127

Arginine

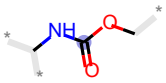
0

1

2

3

5



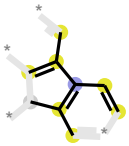
9

10

12

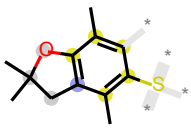
14

16

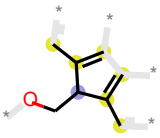


17

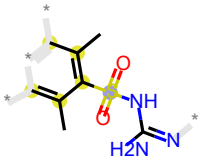
19



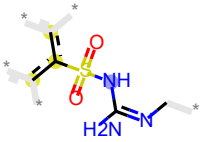
21



22

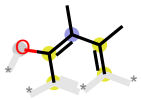


23



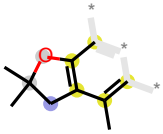
27

30



32

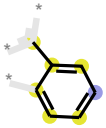
33



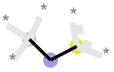
34

36

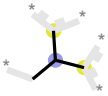
39



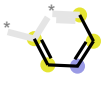
41



44



45

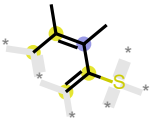


47

49

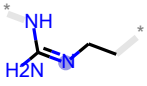
51

52

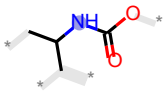


53

55

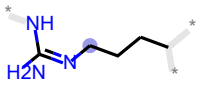


56

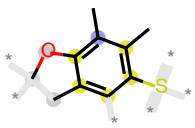


57

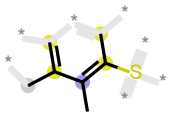
58



61

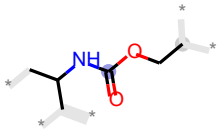


62



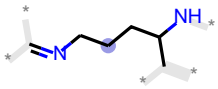
63

64

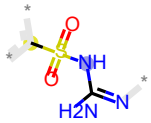


66

68



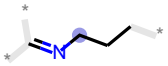
70



74

80

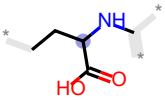
81



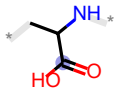
84

85

86



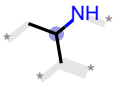
87



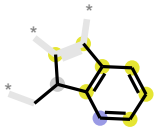
88

93

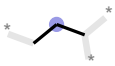
94



96

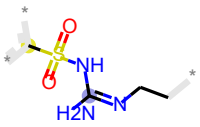


97



99

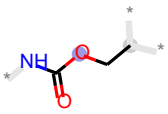
100



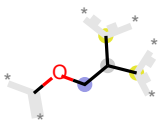
103

105

108

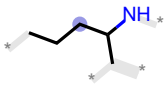


109

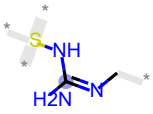


110

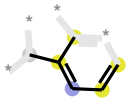
112



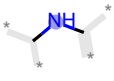
113



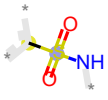
114



115

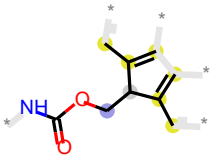


117

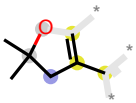


118

119



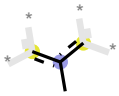
120



121

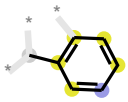
122

123



124

125

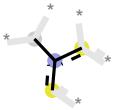


127

Serine

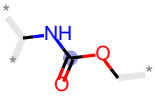
0

1



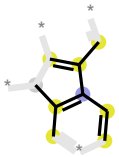
3

8

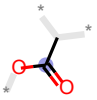


9

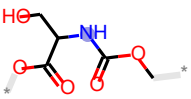
10



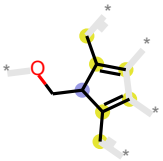
17



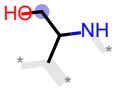
19



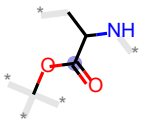
20



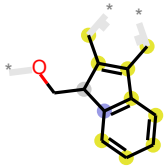
22



23

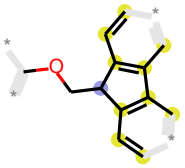


28



32

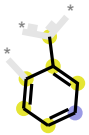
33



36

38

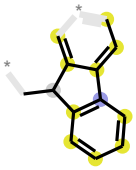
39



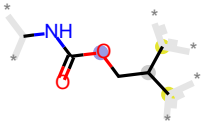
41

45

47

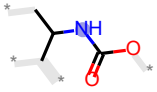


49



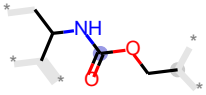
53

55



57

64



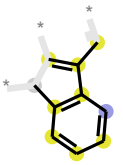
66

67

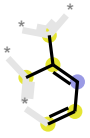
80

81

86



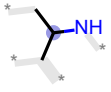
87



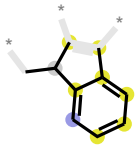
88

91

94



96

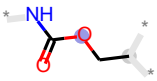


97

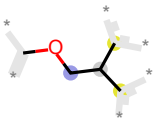
98

100

103



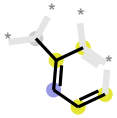
109



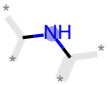
110

112

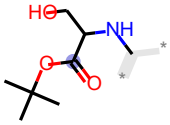
114



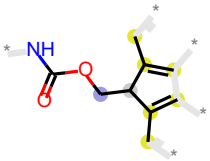
115



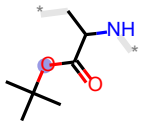
117



118



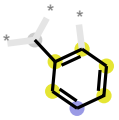
120



122

123

125



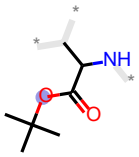
127

Threonine

0

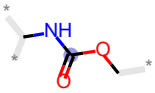
1

3



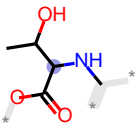
6

8

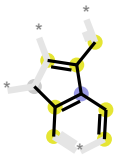


9

10

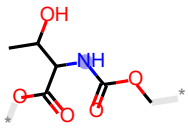


13

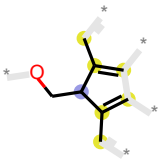


17

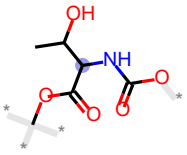
19



21



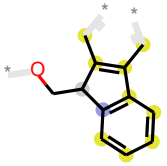
22



24

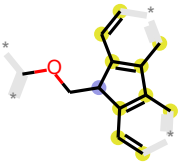
27

29



32

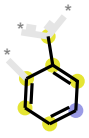
33



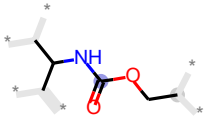
36

38

39



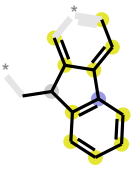
41



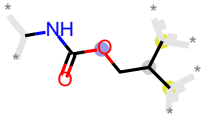
43

45

47

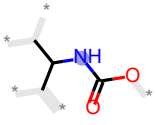


49

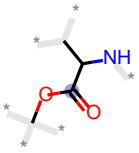


53

55

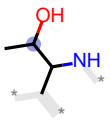


59



62

64



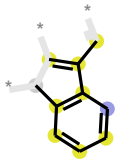
65

67

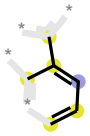
80

81

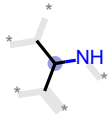
86



87

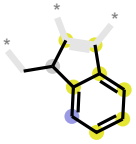


88



94

96

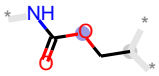


97

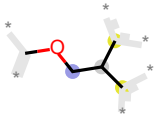
99

100

103



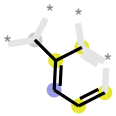
109



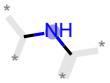
110

112

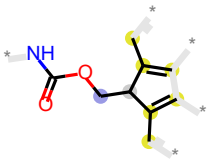
114



115



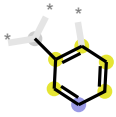
117



120

123

125

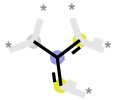


127

Valine

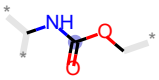
0

1



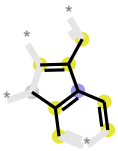
3

5



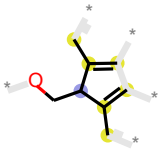
9

10



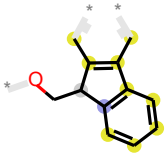
17

19



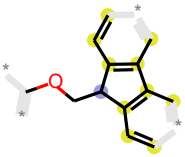
22

27



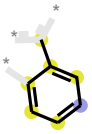
32

33

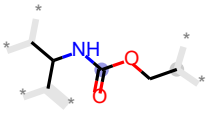


36

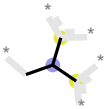
39



41

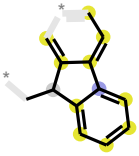


43



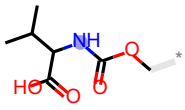
45

47



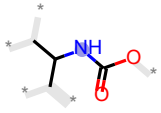
49

52

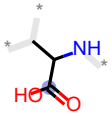


53

55

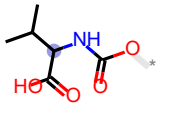


59



63

64

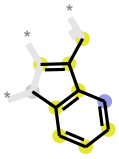


65

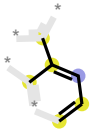
80

81

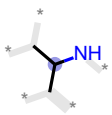
86



87

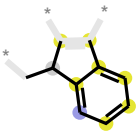


88



94

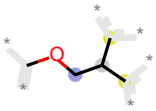
96



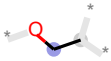
97

100

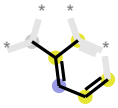
109



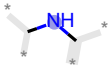
110



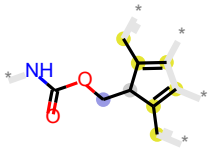
112



115



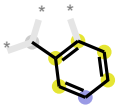
117



120

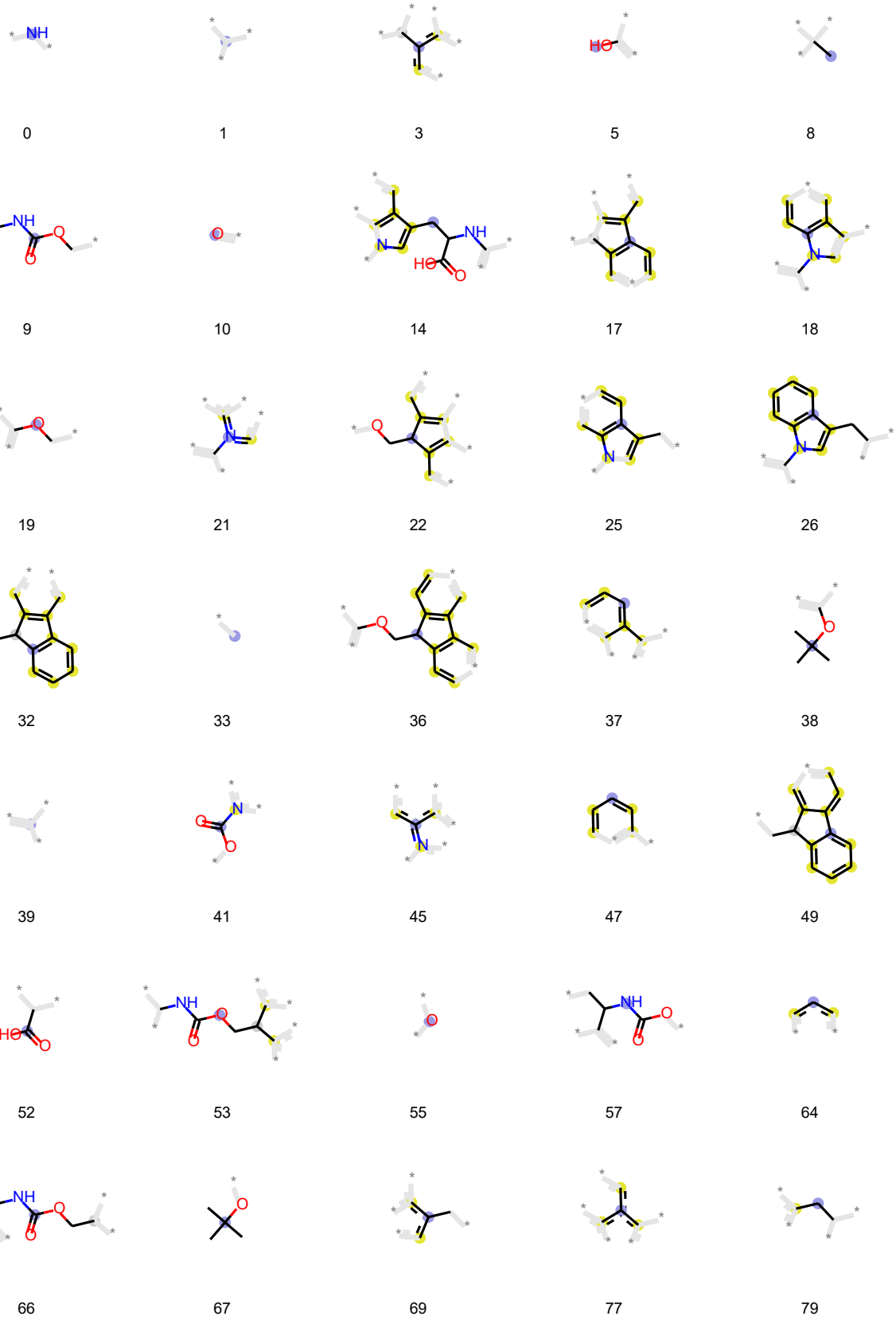
123

125



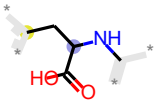
127

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

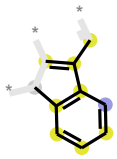
80

81

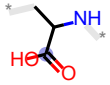


83

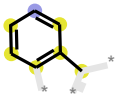
86



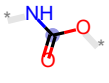
87



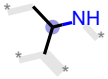
88



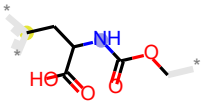
93



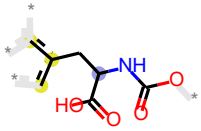
94



96

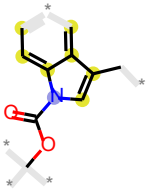


97

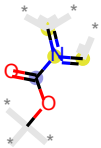


98

100

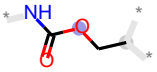


101

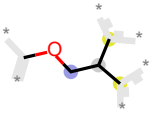


102

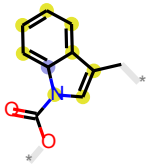
103



109



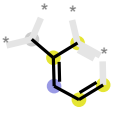
110



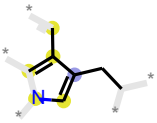
111

112

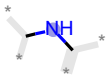
114



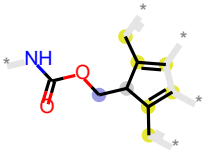
115



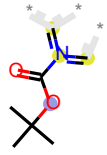
116



117



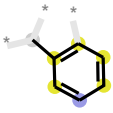
120



121

123

125



127

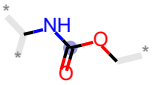
Tyrosine

0

1

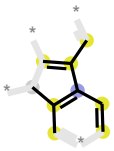
3

8



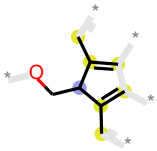
9

10

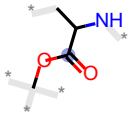


17

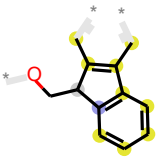
19



22

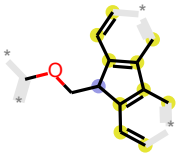


28



32

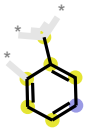
33



36

38

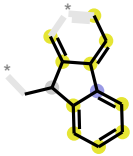
39



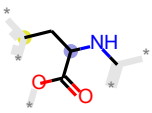
41

45

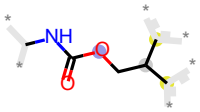
47



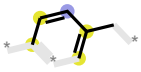
49



50

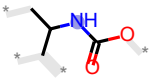


53



54

55

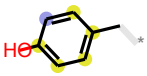


57

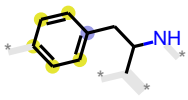
64

66

67



74

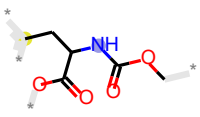


78

79

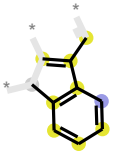
80

81

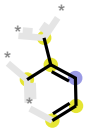


85

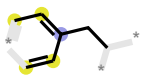
86



87

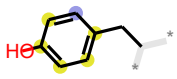


88

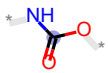


89

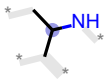
90



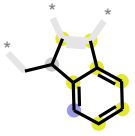
93



94



96

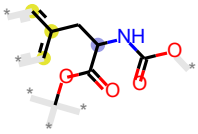


97

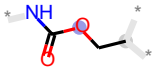
100

103

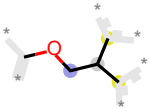
105



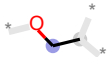
107



109

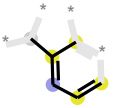


110

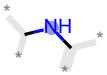


112

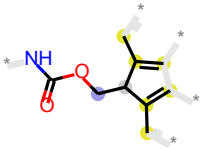
114



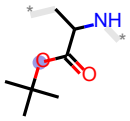
115



117



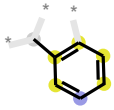
120



122

123

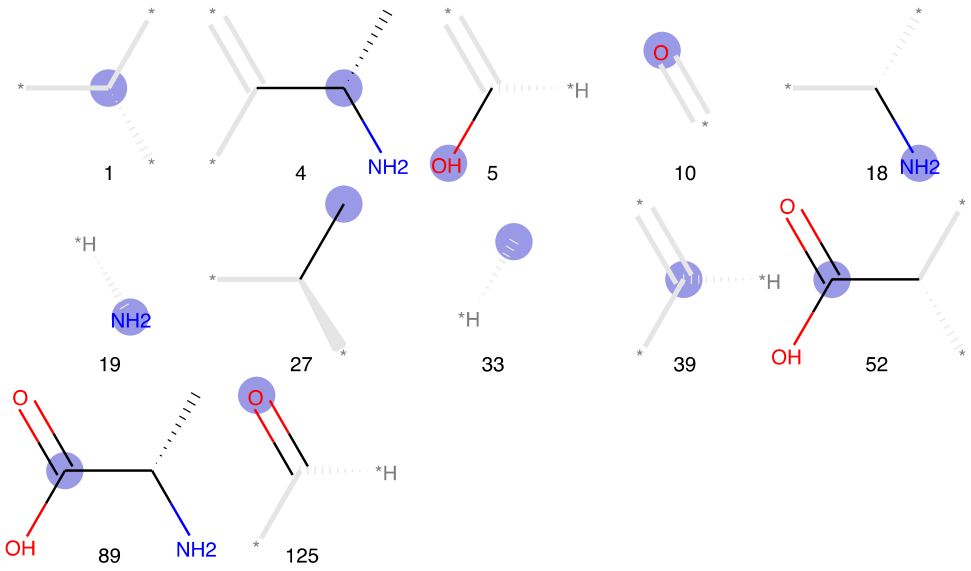
125



127

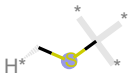
8.2 Substructures for pre-chain residues

Alanine



Cysteine

1



2

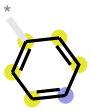
5

9

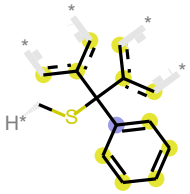
10

18

19

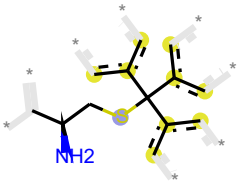


27

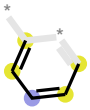


37

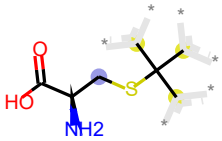
39



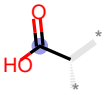
45



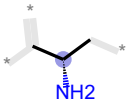
47



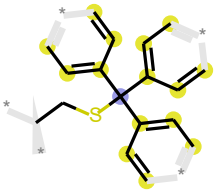
49



52

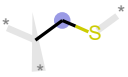


57



61

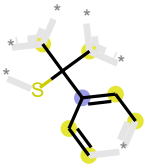
64



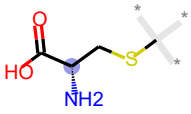
69

80

81

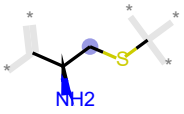


83

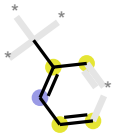


85

86

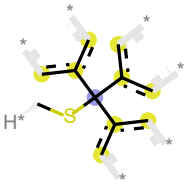


88



97

100



112

114

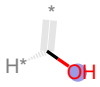
116

124

125

Aspartic acid

1



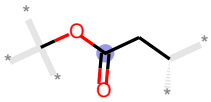
5

8

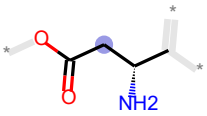
10

18

19

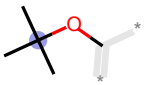


22



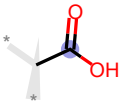
26

33



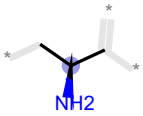
38

39

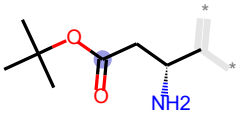


52

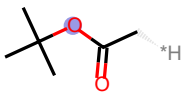
55



57



59



60

67

80

103

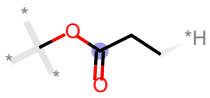
106

114

125

Glutamic acid

1



4

5

8

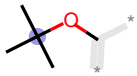
10

18

19

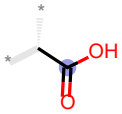
28

33



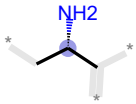
38

39

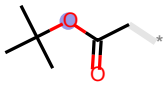


52

55

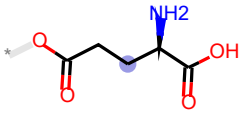


57



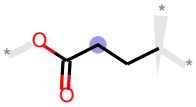
60

67

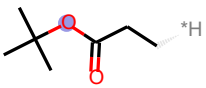


75

80



88

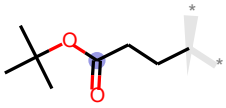


90

99

103

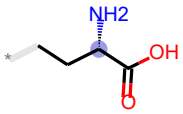
106



109

114

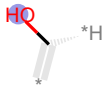
125



127

Phenylalanine

1

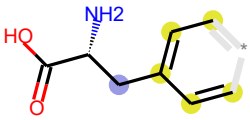


5

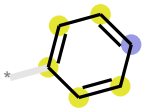
10

18

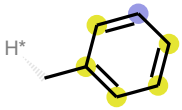
19



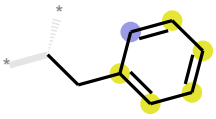
22



27

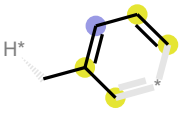


32



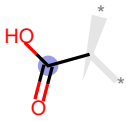
36

39

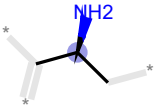


42

47

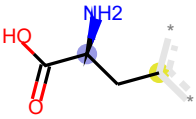


52

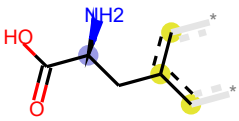


57

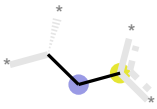
64



69



73

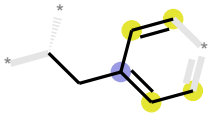


79

80

81

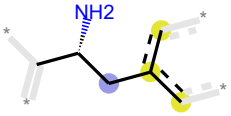
86



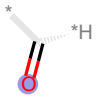
89

90

100

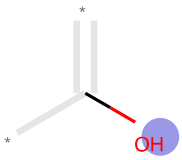


121

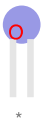


125

Glycine

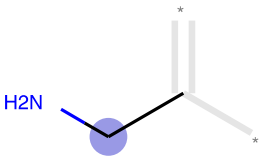


5

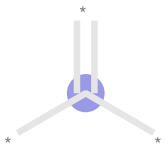


10

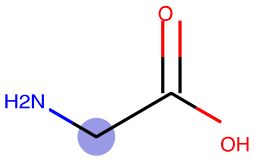
19



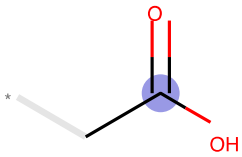
27



39

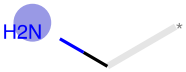


70

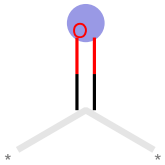


73

80



85



125

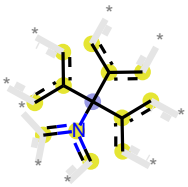
Histidine

1

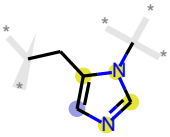
5

9

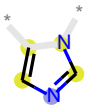
10



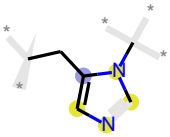
12



13



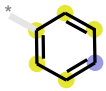
14



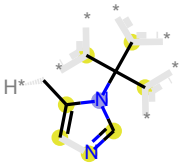
17

18

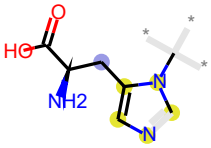
19



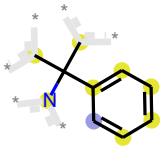
27



29



30

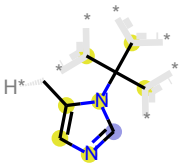


32

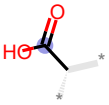
36

39

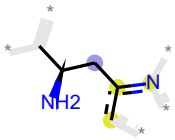
47



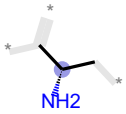
49



52



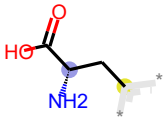
54



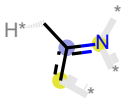
57

64

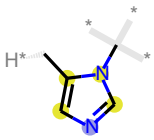
67



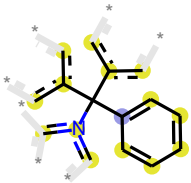
69



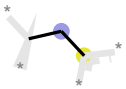
70



72



73

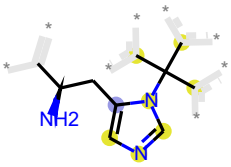


79

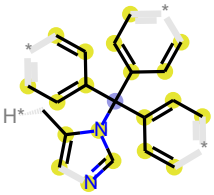
80

81

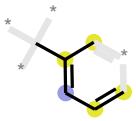
86



87

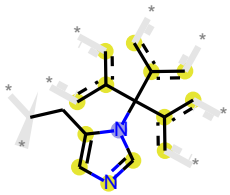


91

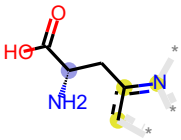


97

100

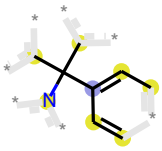


105



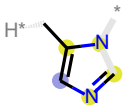
107

111



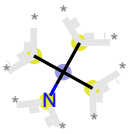
112

114



115

122

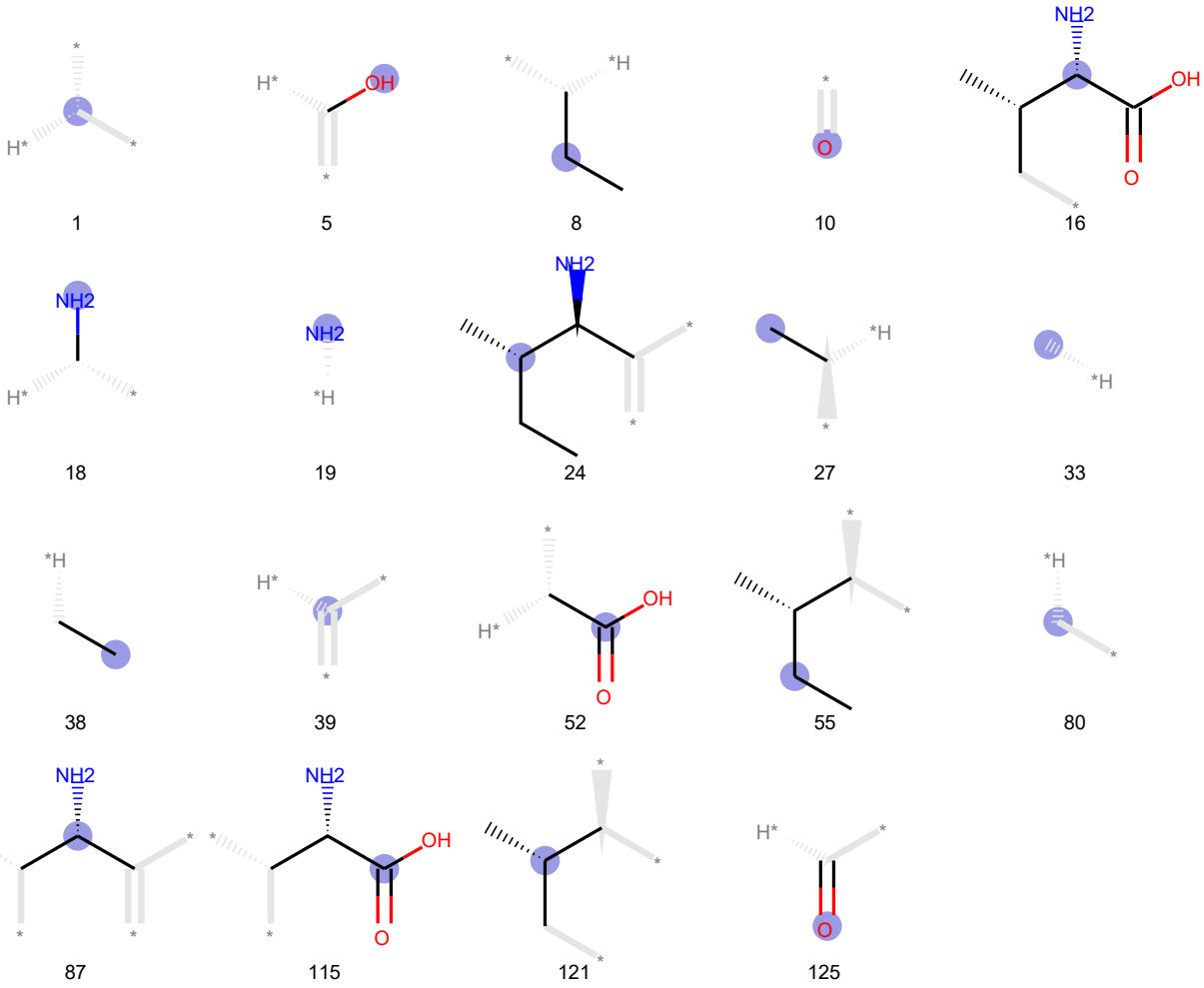


123

125

127

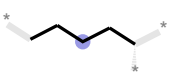
Isoleucine



Lysine

0

1



3

5

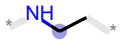
8

10

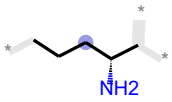
16

18

19

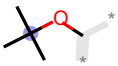


27



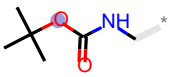
28

33

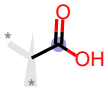


38

39



42

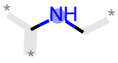


52

55

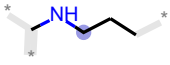
57

67

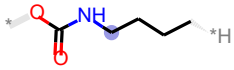


69

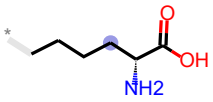
80



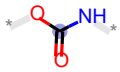
84



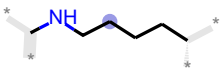
85



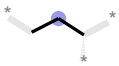
86



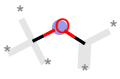
94



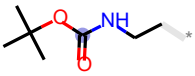
98



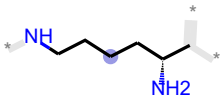
99



103



105

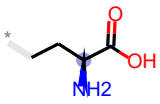


106

114

119

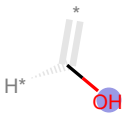
125



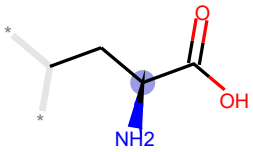
127

Leucine

1

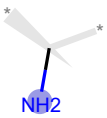


5



8

10

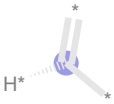


18

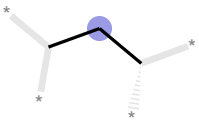
19

27

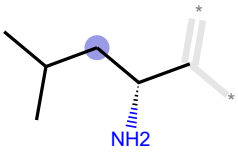
33



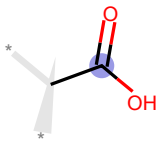
39



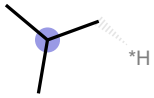
40



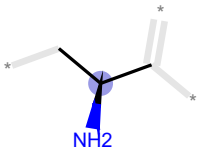
41



52

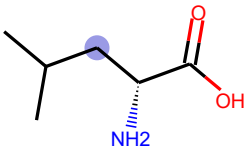


55

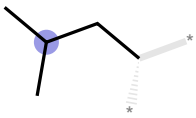


57

80



110

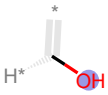


119

125

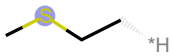
Methionine

1



5

10



11

18

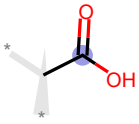
19

22

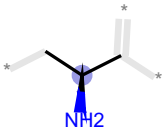
30

33

39



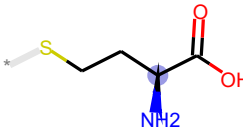
52



57

80

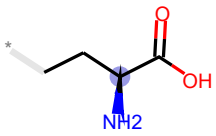
99



112

116

125

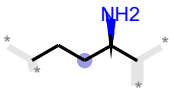


127

Asparagine

0

1

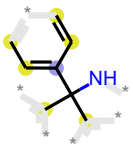


4

5

9

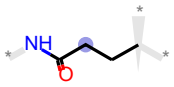
10



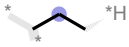
11

18

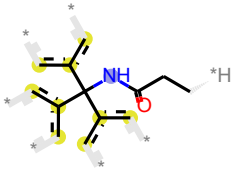
19



27

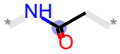


28



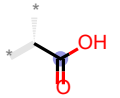
36

39

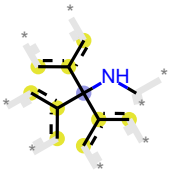


41

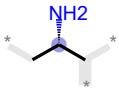
47



52

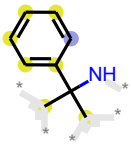


56



57

64

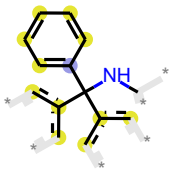


69

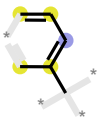
80

81

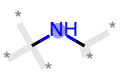
86



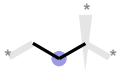
91



97

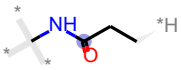


98

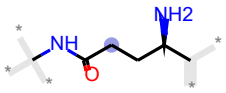


99

100

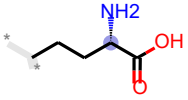


101



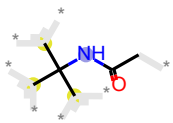
104

105

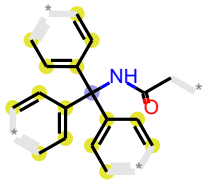


106

114

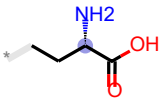


120



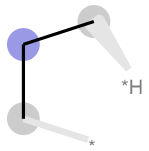
122

125

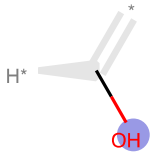


127

Proline

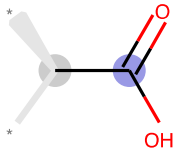


4



5

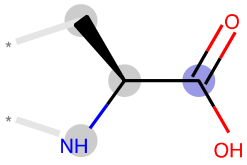
10



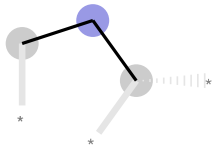
23

30

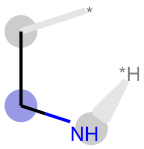
39



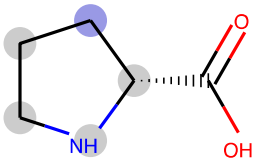
41



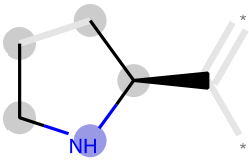
45



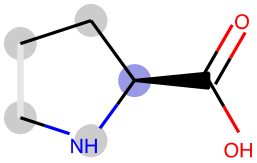
49



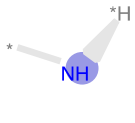
63



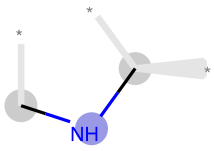
72



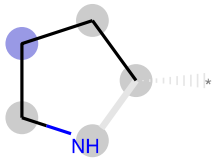
74



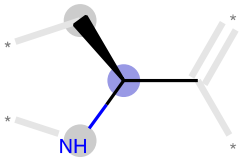
90



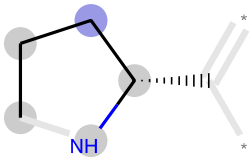
99



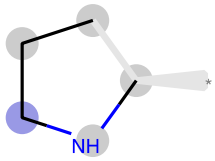
102



106

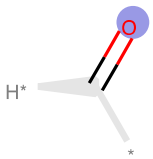


118



120

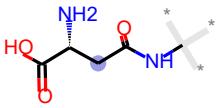
123



125

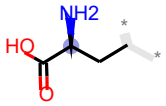
0

1



4

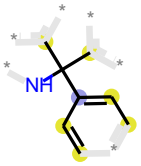
5



8

9

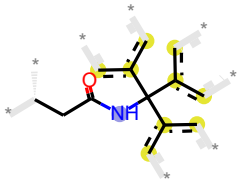
10



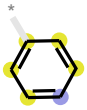
11

18

19

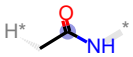


24



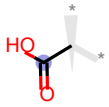
27

39

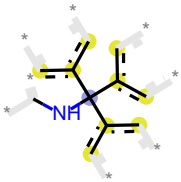


41

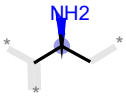
47



52

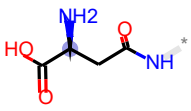


56

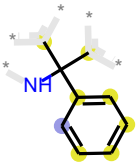


57

64



68

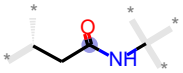


69

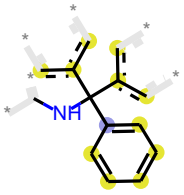
80

81

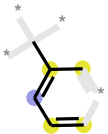
86



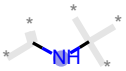
89



91



97

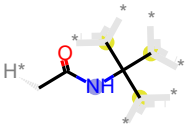


98

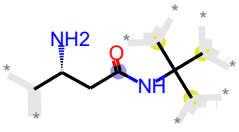
100

105

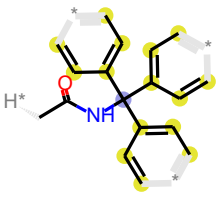
114



120



121



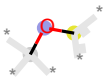
122

125

Arginine

0

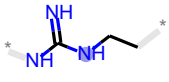
1



2

5

10



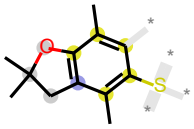
12

14

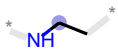
16

18

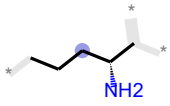
19



21

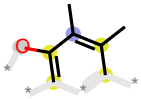


27



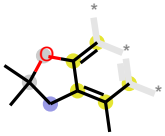
28

30



32

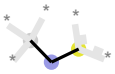
33



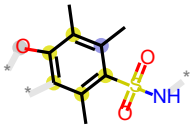
34

36

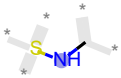
39



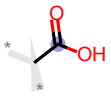
44



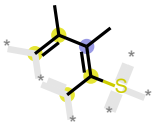
49



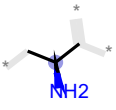
51



52

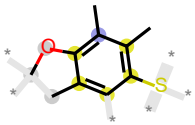


53

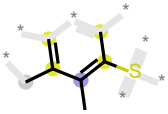


57

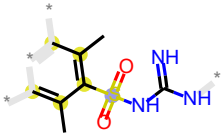
58



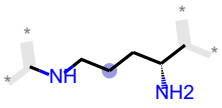
62



63

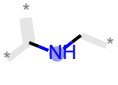


66

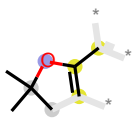


67

68

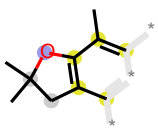


69

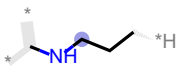


74

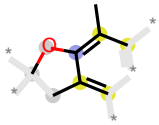
80



81



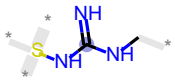
84



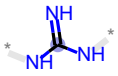
88

93

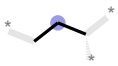
94



95

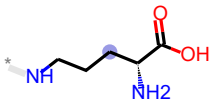


98



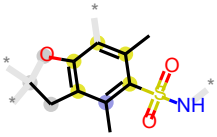
99

100



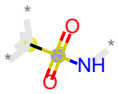
103

105



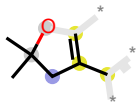
109

111



118

119

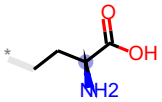


121

122

124

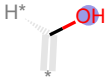
125



127

Serine

1



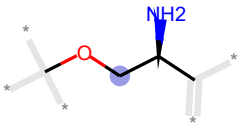
5

8

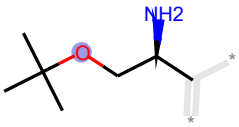
10

18

19



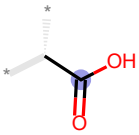
21



29

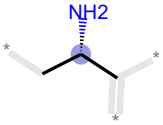
33

39

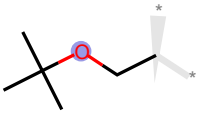


52

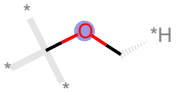
55



57

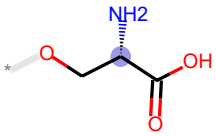


60



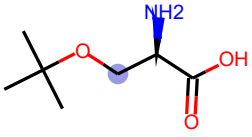
64

67



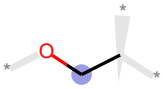
75

80



98

114

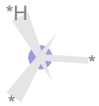


118

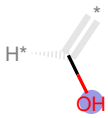
124

125

Threonine



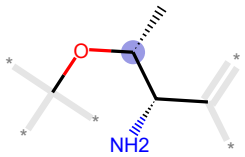
1



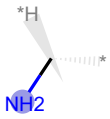
5

8

10

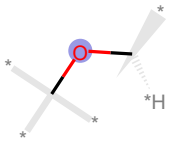


17

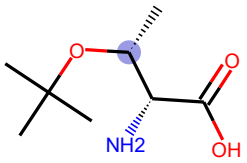


18

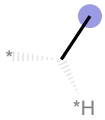
19



22



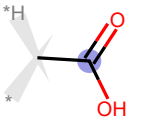
24



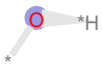
27

33

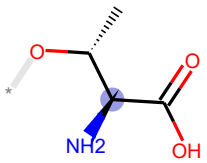
39



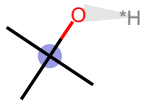
52



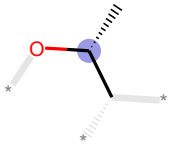
55



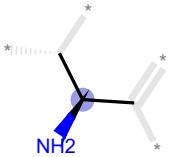
59



67

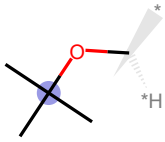


82



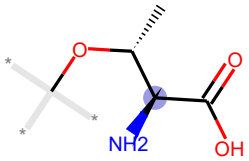
87

114



115

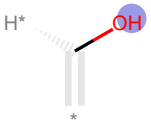
125



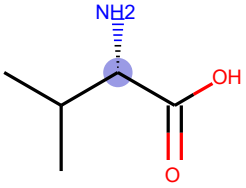
126

Valine

1



5

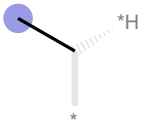


7

10

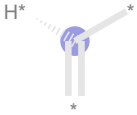
18

19

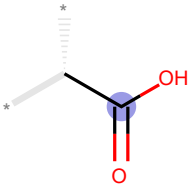


27

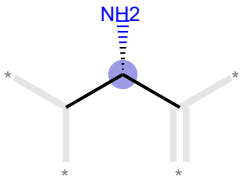
33



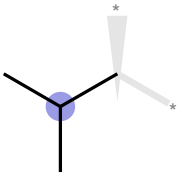
39



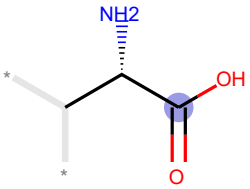
52



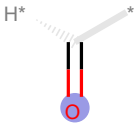
87



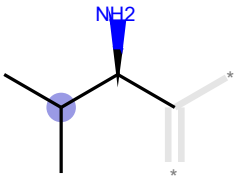
109



115



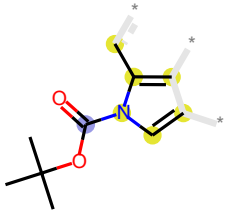
125



126

Tryptophan


1


3

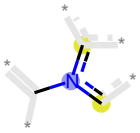

5

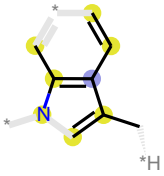

8

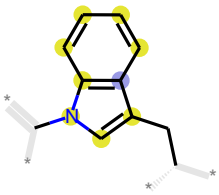

10


18

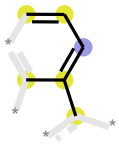

19


21


25

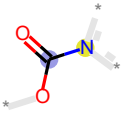

26

33


37

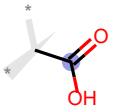

38


39

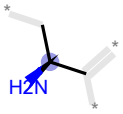

41


45


47


52


55


57


64


67


69


77


79

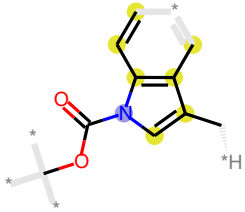

80

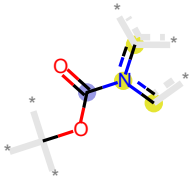

81


86

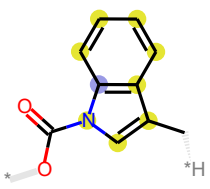

93

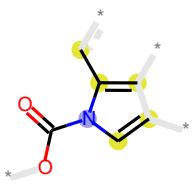

100


101

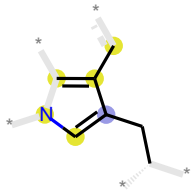

102


103

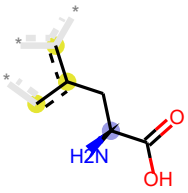

111


112

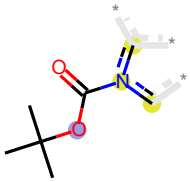
114



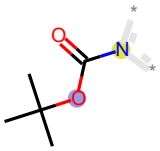
116



120



121



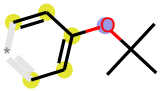
123

125

Tyrosine

1

3



4

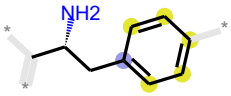
5

8

10

18

19

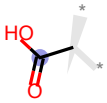


22

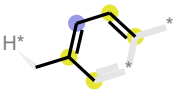
33

36

39

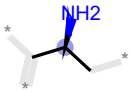


52

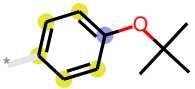


54

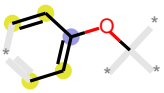
55



57



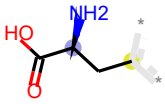
60



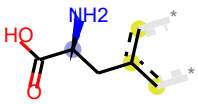
61

66

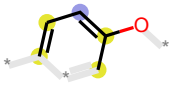
67



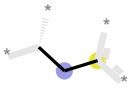
69



73



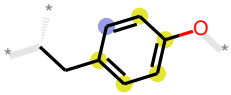
78



79

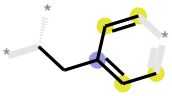
80

81

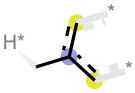


84

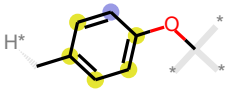
86



89



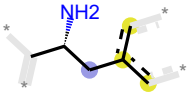
90



91

100

114

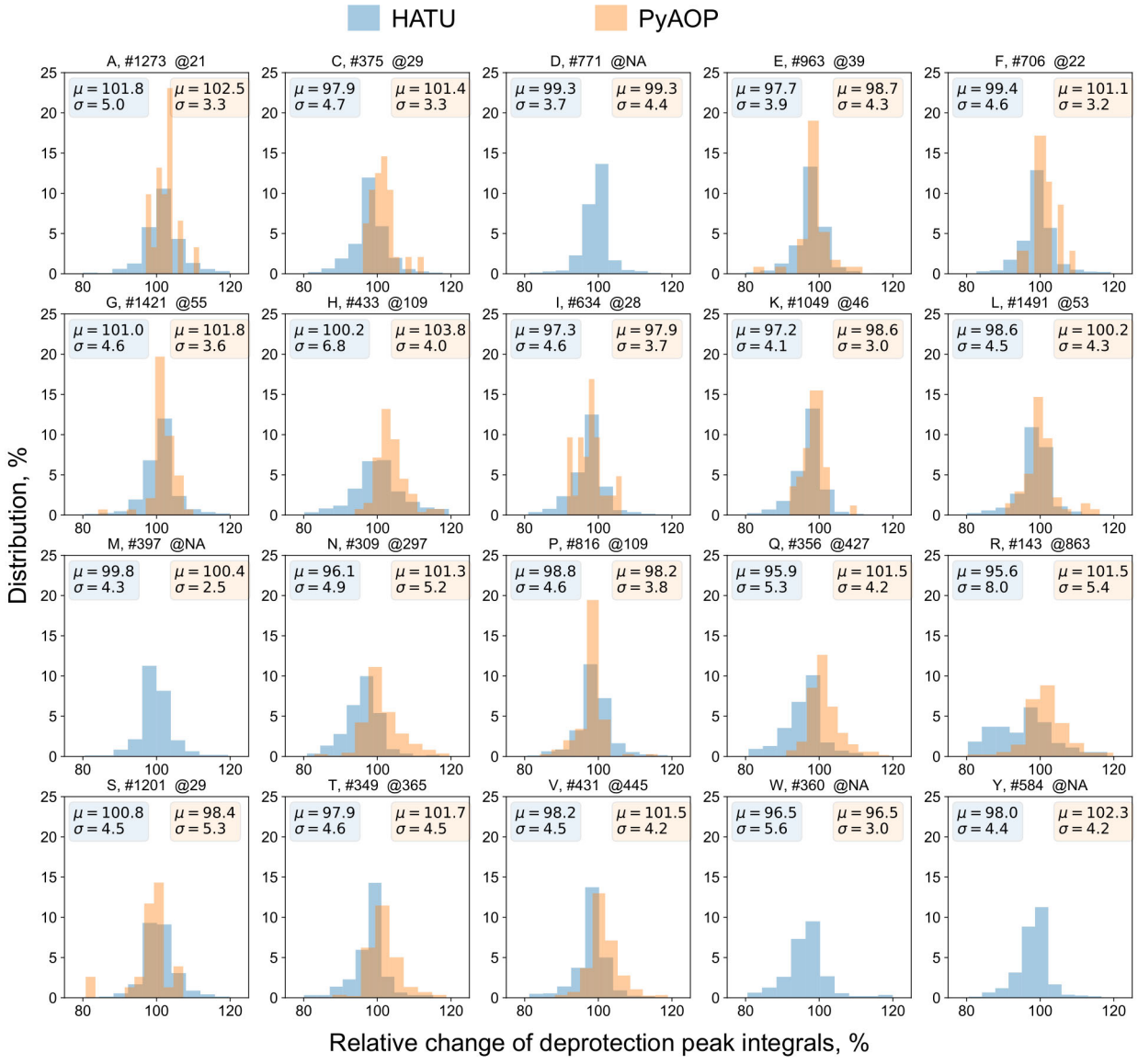


121

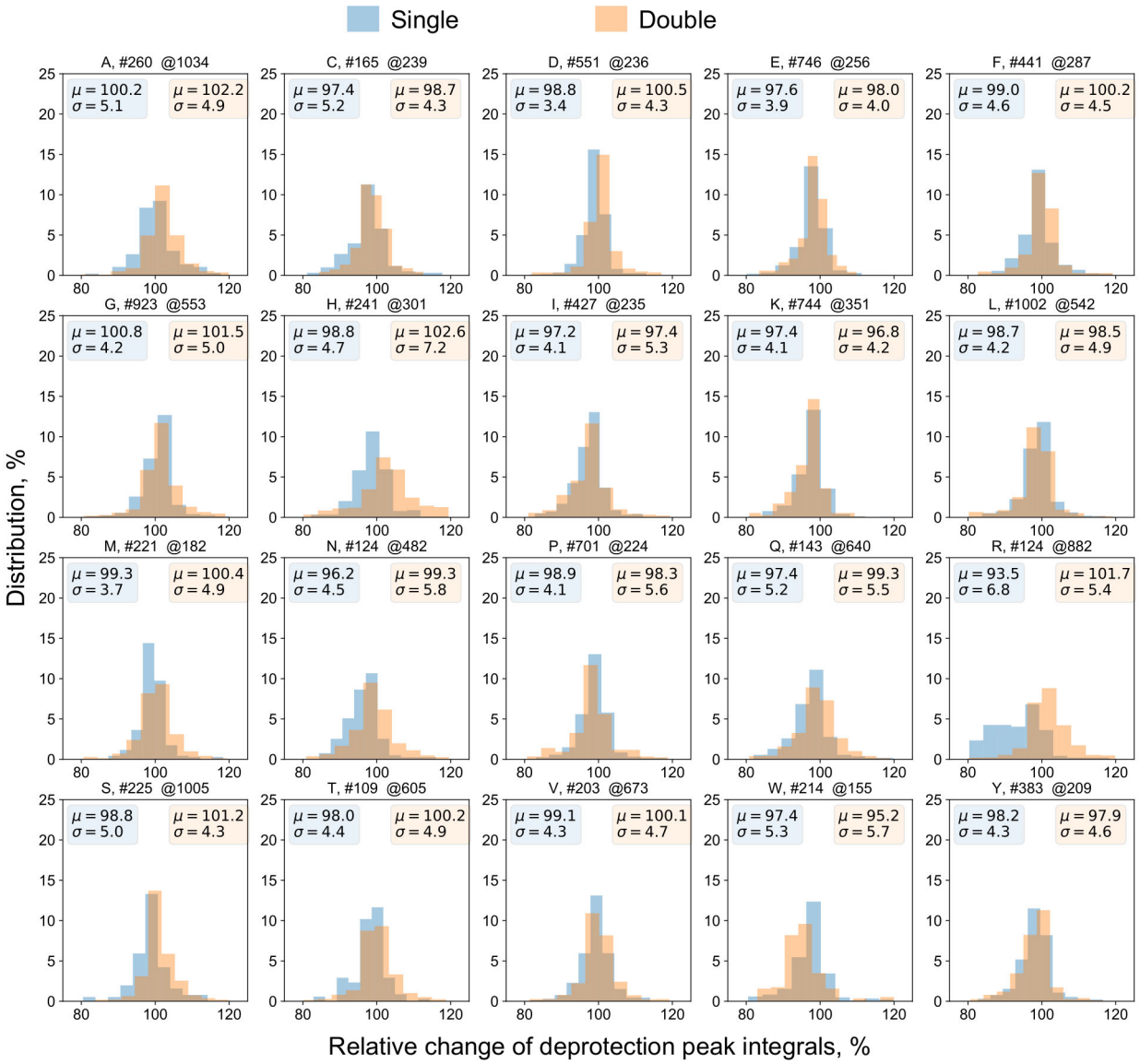
125

9 Appendix 2

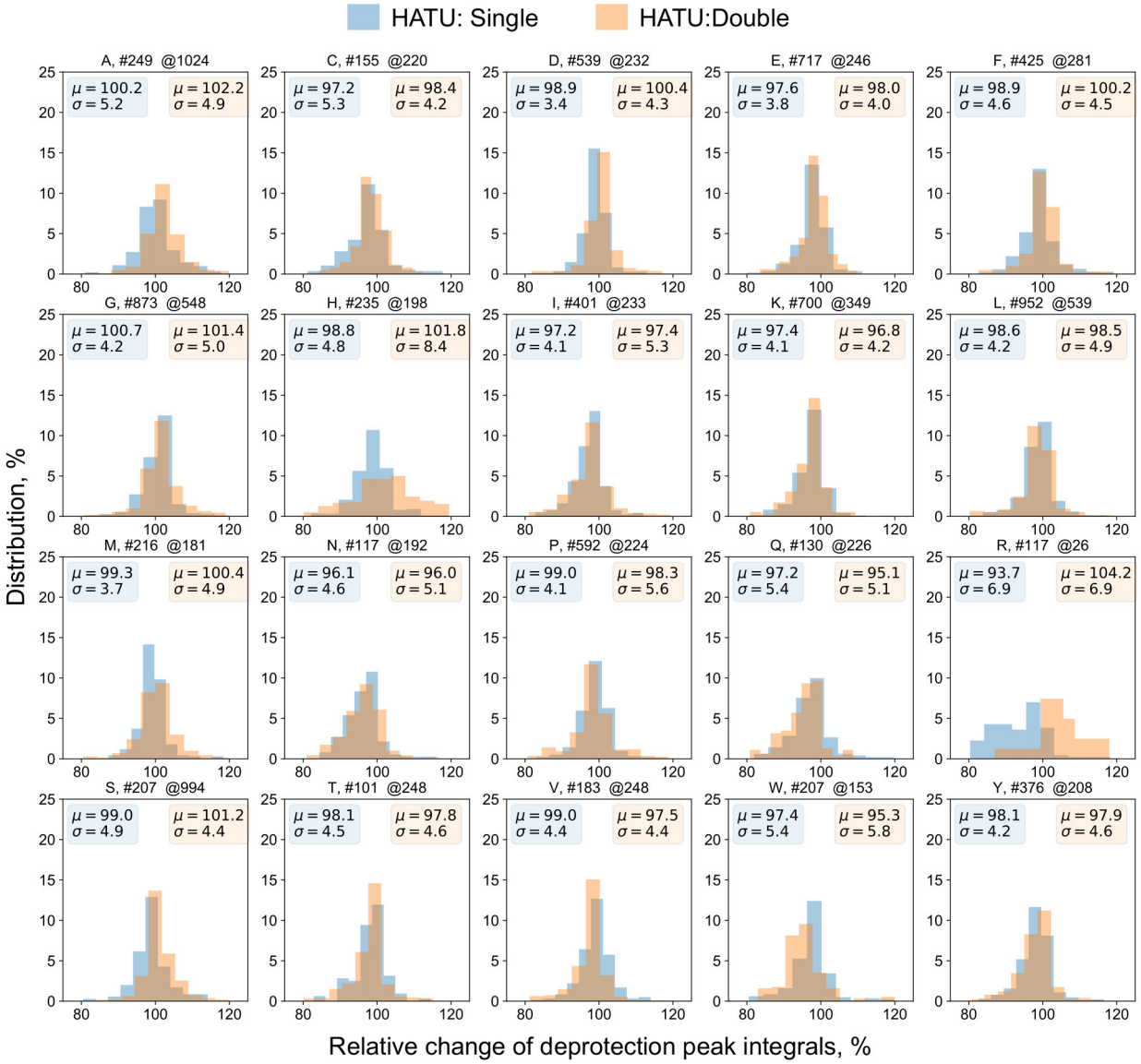
Histogram of relative change of deprotection peak integrals by amino acid, for different coupling agents – HATU (■) and PyAOP (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number of data 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.



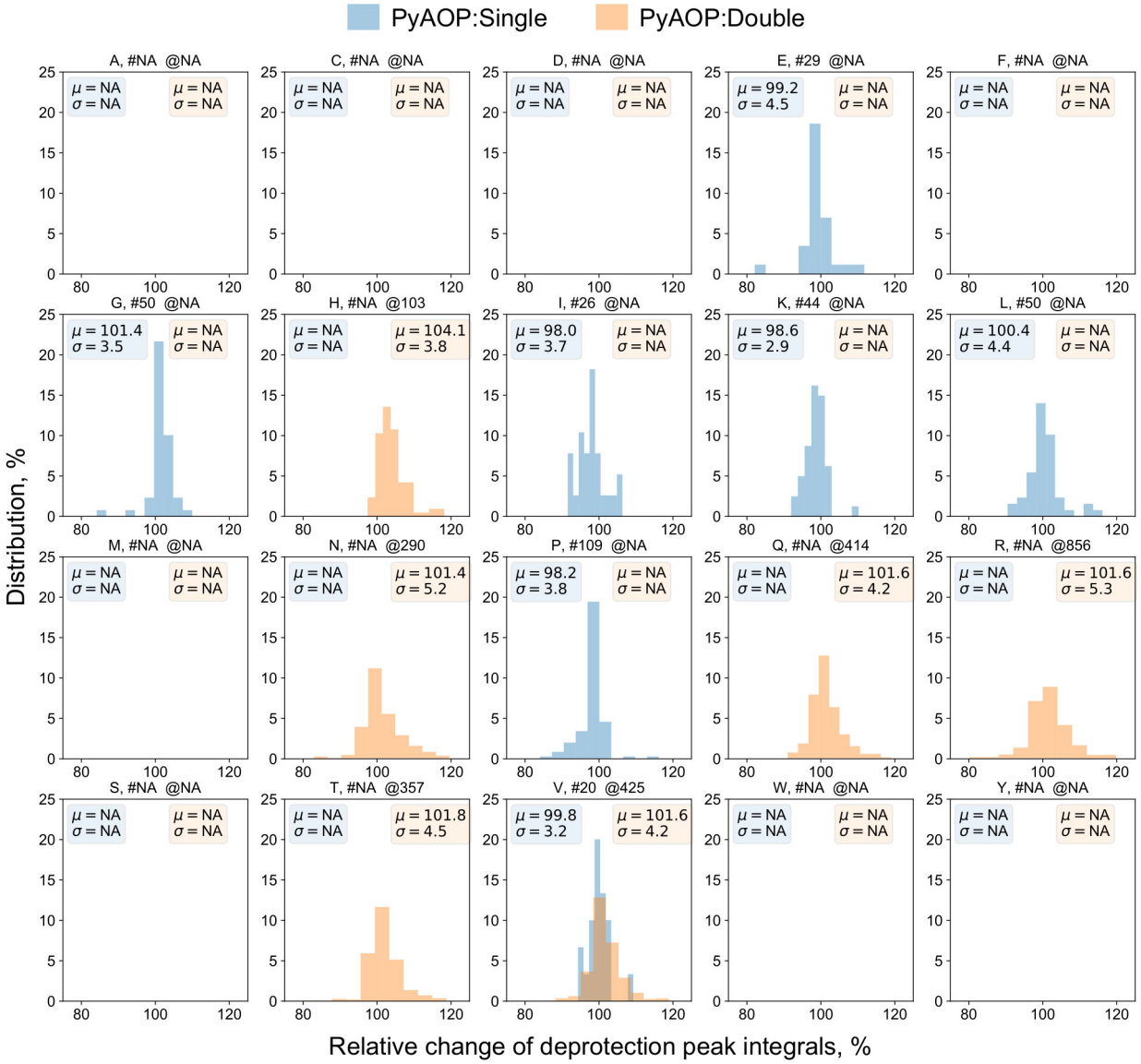
Histogram of relative change of deprotection peak integrals by amino acid, for different coupling strokes – Single (■) and Double (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. Distributions for which the number 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.



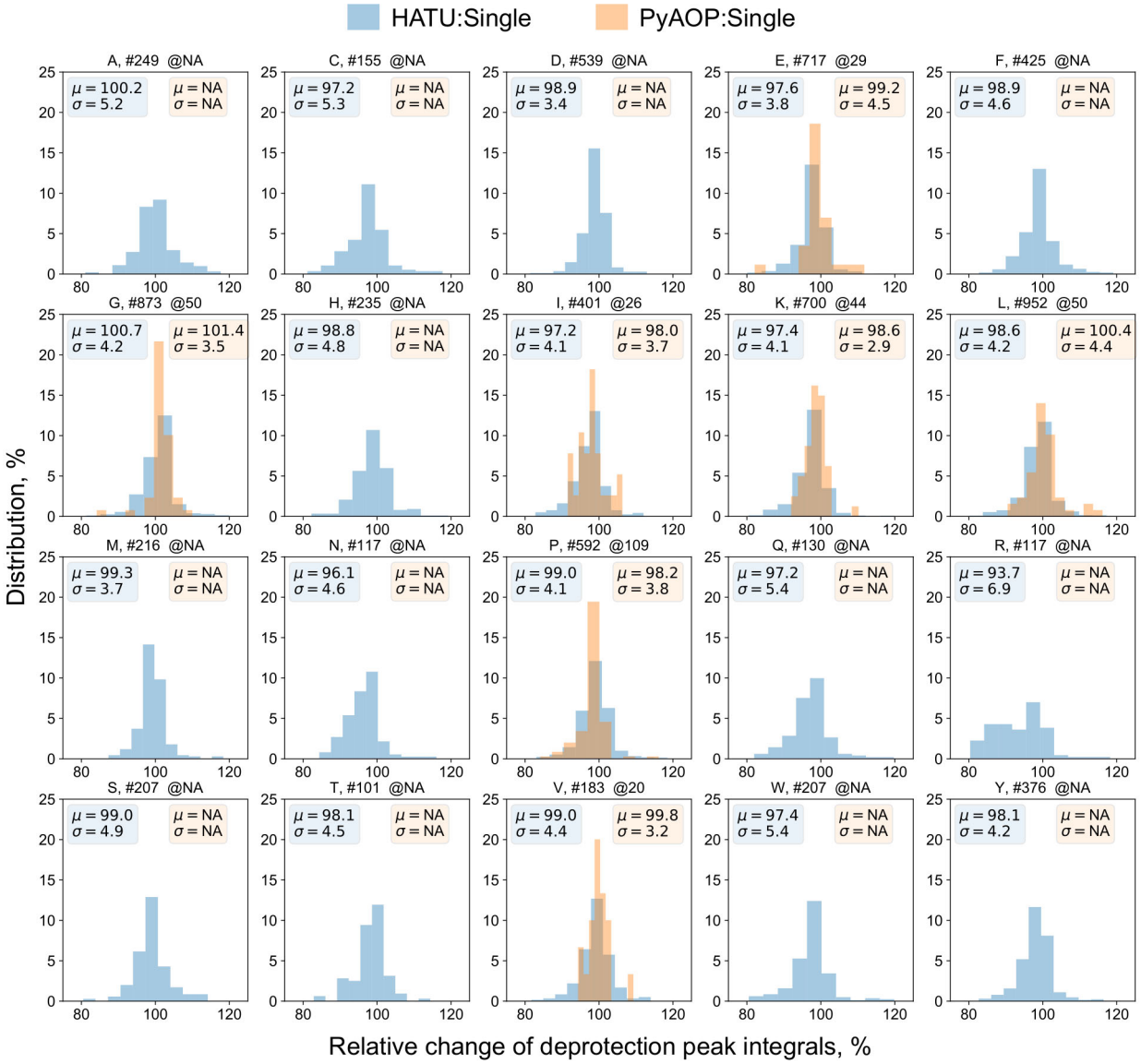
Histogram of relative change of deprotection peak integrals by amino acid, for different combinations of coupling agent:coupling strokes – HATU:Single (■) and HATU:Double (■). The number of data points are noted above the specific distribution after the symbol notation - # and @ for respective conditions. 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.