Learning representations from mass spectra for peptide property prediction

Michael Murphy [1] **Kevin Yang** [2] **Stefanie Jegelka** [1] **Ernest Fraenkel** [1]

Abstract

Peptide molecules have long been viewed as promising candidates for design of novel therapeutics, and are likely to benefit from ML-guided approaches. However, learning to predict biological properties of peptides often suffers from a scarcity of labelled training data. We demonstrate mass spectrometry, which provides high-throughput, multidimensional biophysical measurements of peptides, can be used to learn effective representations for peptide property prediction. Specifically, our pretext task asks to identify masked residues of a peptide sequence using its mass spectrum. This yields an encoder that we can then apply to any peptide sequence, irrespective of whether we have spectra for it. Our approach is competitive with a state-of-the-art evolutionary pretext task on a number of downstream tasks, and requires orders of magnitude fewer pretraining examples.

1. Introduction

The functional diversity and ease of synthesis of peptide molecules makes them attractive candidates for drug development (Fosgerau & Hoffmann, 2015). However, an exponentially-large search space and often-difficult experimental protocols make it challenging to design peptides with specific properties via traditional laboratory techniques. Machine learning methods promise to fill this gap (Chen et al., 2021; Wan et al., 2022; Schissel et al., 2020): given a surrogate model mapping sequence to biological function, we can screen large numbers of peptides in-silico and prioritize candidates for experimental followup. Unfortunately, many property prediction tasks of therapeutic relevance suffer from a scarcity of labelled sequences (Chen et al., 2021). This calls for schemes such as multi-task learning (Qi et al., 2012), active learning-guided experimental design (Yang et al., 2019), or self-supervised pretraining from unlabelled protein sequences (Rives et al., 2021; Yang et al., 2022). The state-of-the-art in pretraining for amino acid sequences applies the *masked language modelling* task introduced by Devlin et al. (2018) to large protein databases: a subset of the residues in a protein sequence are masked, and the model is asked to predict these from the rest. The pretraining signal here is *evolutionary*: protein sequences are not random, but rather arise through natural selection to maximize fitness. The masking pretext can therefore be interpreted as predicting which residue(s) would maximize biological activity of the protein encoded by the sequence.

While evolutionary pretraining enjoys a biologically plausible objective for protein sequences, typically hundreds of amino acids long, it may not be optimal for *peptides*, which are at most tens of amino acids long, and are not usually evolutionarily selected for function in isolation from a larger protein. This precludes training a similar masking pretext on peptide-length sequences without additional context, as they lack sufficient information to identify the masked residue. Learning to extract this additional context from proteins incurs a high computational cost, requiring tens to hundreds of millions of long sequences (Yang et al., 2022; Rives et al., 2021). We resolve these issues by incorporating an information-rich, plentiful – yet to our knowledge, previously-unexploited – data modality for pretraining of peptide property prediction tasks: *mass spectrometry*.

A typical proteomics experiment produces mass spectra of tens of thousands of peptides (Steen & Mann, 2004). Following algorithmic annotation, each spectrum represents a histogram of ions formed by fragmentation of an ionized peptide, which describes the propensity of the peptide to cleave at each bond along the backbone. This depends on a number of related physical properties, including the identities of the amino acids, the distribution of charge along the sidechains and backbone (Paizs & Suhai, 2005), and the peptide's secondary structure (Tsaprailis et al., 1999). All these properties are also central to biological activities – and, we conjecture, are more directly captured by the *biophysical* signals measured in mass spectrometry.

In short, this work makes the following contributions:

- We identify mass spectrometry as a modality capturing information relevant to peptide property prediction;
- We develop a pretext task that uses mass spectra to

[1] MIT [2] Microsoft Research New England. Correspondence to: Michael Murphy <murphy17@mit.edu>.

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

guide learning of a representation that can be used for peptide sequences directly, and is hence applicable to peptides that lack spectra; and

- We show that representations learned with this pretext task are competitive with evolutionary pretraining on a number of peptide property prediction tasks, using far fewer pretraining examples.

2. Related work

There is substantial recent interest in evolutionary learning of protein representations using large language models, primarily transformers (Rives et al., 2021; Elnaggar et al., 2021; Rao et al., 2019), but also convolutional neural networks (Yang et al., 2022). These have recently been applied to peptide property prediction (Dee, 2022). Others pretrain for protein tasks by predicting other modalities from sequence, including 3D structure (Bepler & Berger, 2019; Zhang et al., 2022a) and functional annotations (Zhang et al., 2022b); but to our knowledge, mass spectrometry has not yet been used for this purpose. We also draw inspiration from successful application of deep learning for NLP to other tasks in mass spectrometry, including predicting spectra from sequence (Zhou et al., 2017; Gessulat et al., 2019) and sequence from spectra (Yilmaz et al., 2022; Qiao et al., 2021).

3. Methods

3.1. Pretext task

Our objective is similar to the masked language model in (Yang et al., 2022) and (Rives et al., 2021), in which we randomly mask a single residue from the peptide sequence and require our model to correctly impute it.[1] However, short peptide sequences alone do not provide sufficient context for this task. We therefore additionally condition on the entire observed mass spectrum in addition to the remainder of the sequence; which, as we later show, *does* contain sufficient information to identify the masked amino acid. By returning an intermediate representation of the peptide sequence prior to when it is merged with the spectrum, we can then apply the resulting model to *any* peptide sequence – not just those for which we have spectra.

3.2. Model architecture

Our architecture, shown in Figure 1, comprises two paired encoders: one for the sequence modality, and one for the spectrum modality. The sequence encoder follows the same architecture as the CARP model described in (Yang et al., 2022); briefly, it converts each amino acid symbol to an 8-dimensional embedding vector, and then passes this sequence of embeddings through a ByteNet dilated CNN as developed in (Kalchbrenner et al., 2016), which yields a *sequence-length* encoding at its output.

We represent a mass spectrum of a length-L peptide as an $(L-1) \times K$-dimensional table of probabilities of each of K ion types arising from cleavage of $L-1$ bonds, which we compute by normalizing the observed counts across annotated peaks to sum to 1. We also concatenate at each bond two scalar-valued properties of the spectrum as a whole: its observed electric charge prior to fragmentation, and the collision energy at which it was measured. This is passed into another ByteNet encoder, yielding at its output one embedding vector per bond. The outputs of the two encoders are concatenated feature-wise (padding the bond encodings to length L) and passed into a 2-layer ReLU classifier applied independently at each position, yielding a vector of logits per each of t amino acids. We then index into the prediction at the masked token and minimize cross-entropy loss against the true residue.

For both encoders we use the same depth and width as the smallest pretrained model in (Yang et al., 2022), CARP-600k, which uses $n = 16$ layers of $d = 128$-dimensional ByteNet blocks. We train using minibatches of 512 peptide-spectrum pairs with Adam (Kingma & Ba, 2014) (learning rate $= 5 \times 10^{-4}$), early-stopping on validation cross-entropy after 126 epochs.

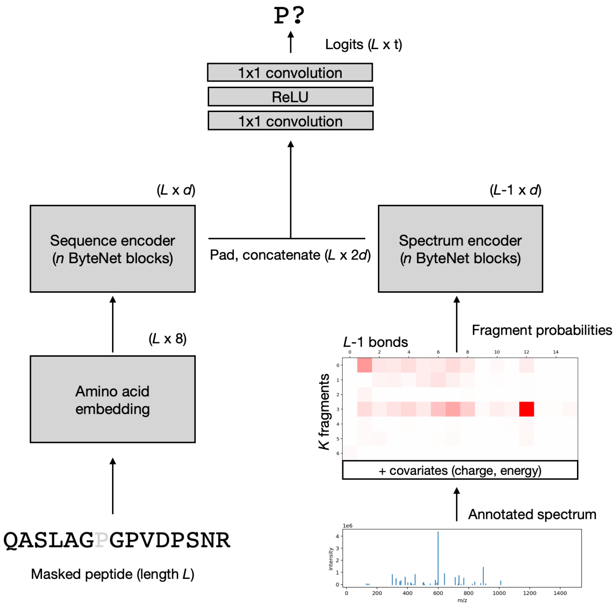


Figure 1. Our model's architecture. We return the $L \times d$ output of the sequence encoder as our embeddings. See Yang et al. (2022) for a more in-depth exposition of the ByteNet architecture.

[1] We initially tried a simpler pretext of predicting spectra from sequence: using only a sequence encoder, and yielding a logit per ion type instead of per amino acid in Figure 1. While this performed well on the pretext (mean test $R^2 = 0.93$), we found it less effective on downstream tasks than our masking objective.

3.3. Pretext dataset

We use a library of mass spectra generated in Part I of the ProteomeTools project (Zolg et al., 2017). This comprises 2,980,009 annotated mass spectra of 391,273 unique peptides derived from the human proteome, composed of the 20 canonical amino acids plus two post-translational modifications (methionine sulfoxide and carbamidomethyl-cysteine). The redundancy per peptide arises through different combinations of charge states, collision energies, and modifications, each of which generally yield distinct spectra. These spectra are provided as lists of (mass, intensity) tuples, each of which is annotated with the bond and one of K possible types of the respective fragment. We use an $85/5/10$ train/validation/test split; to avoid leakage, we cluster sequences using CD-HIT (Fu et al., 2012) (threshold 0.5, word length 3) and randomly assign entire clusters to each split.

3.4. Peptide property prediction datasets

We identified a number of datasets of peptide sequences, labelled either positive or negative for some biological property, as representative of potential objectives for peptide design. Where applicable, we only include sequences comprising the 20 canonical amino acids, with $5 \leq L \leq 100$:

- MITOCHONDRIAL TARGETING. Zarin et al. (2021) provide annotations of 5,348 N-terminal intrinsically-disordered regions (IDRs) identified in a screen for mitochondrial targeting in yeast. From these sequences we select 160 positive and 3,960 negative examples.
- CDC28 BINDING. Zarin et al. (2021) also indicate whether the same IDRs are substrates of the kinase Cdc28: this gives 80 positive examples and 4040 negative examples.
- SIGNAL PEPTIDE. SignalP (Teufel et al., 2022) is a high-quality repository of annotated signal peptide sequences derived from eukaryotes and prokaryotes. For simplicity we consider only the SignalP 6.0 training set and ignore type annotations for the prokaryotic sequences, resulting in a binary classification problem of 15,625 positives against 4,665 negatives.
- MHC BINDING. The Dana-Farber Repository for Machine Learning in Immunology (Zhang et al., 2011) provides peptide sequences that bind a number of human and mouse MHC-II complexes. We construct as our positive class the unique 'binding' peptide sequences across the union of training sets listed at `http://projects.met-hilab.org/DFRMLI/HTML/natural.php`, and the union of 'non-binding' peptides for the negative class; discarding sequences appearing in both to yield 9,720 positive examples and 6,945 negative examples.

Within each task we carry out 3-fold nested cross-validation, again splitting via CD-HIT clustering.

3.5. Downstream tasks

To apply our model to the downstream tasks, we discard the spectrum encoder and final classifier, evaluate the sequence encoder on the input peptide sequence, and learn a new classifier on the resulting embeddings. Because the encoder yields a sequence-length representation, our classifier pools across positions via an attention layer (Vaswani et al., 2017), then applies a 2-layer ReLU binary classifier.

We compare to four baselines. (1) `CARP-600k` (Yang et al., 2022) is employed as representative of the state-of-the-art in evolutionary pretraining. This model has essentially the same architecture as our sequence encoder – permitting direct comparison of pretext tasks without confounding from architecture – but is trained on far more data: 41.5 million full-length protein sequences from UniRef50 (Suzek et al., 2014). (2) We also include our model initialized from random without pretraining. Finally, we use two simple models: (3) a length-averaged prediction of a linear classifier applied individually to each amino acid; and (4) a 3-layer, 128-wide CNN with ReLU activations, kernel width of 5, and length-wise average-pooling, prior to a linear output layer.

For pretrained models, we test both freezing the encoder weights and training only the final classifier, and fine-tuning the entire network. All models use Adam (batch size 256, learning rate 5×10^{-4}), early-stopping on validation AUC.

4. Results and Discussion

4.1. Pretext accuracy

Our method identifies randomly-masked amino acids with 69.9% accuracy on the test set. In comparison, evaluating the pretrained evolutionary model on peptide sequences achieves only 10.4% accuracy: this is unsurprising, as the global protein context on which it depends is missing. To confirm the spectral information is actually used, we also tried solving our pretext using the sequence alone: this peaked at a maximal validation accuracy of 18% and then proceeded to overfit, indicating masking alone is insufficient for peptides and spectra are indeed necessary.

Figure 2 shows a per-amino-acid confusion matrix. Reassuringly, errors tend to be structurally-similar amino acids: in particular the branched-chain amino acids (I = isoleucine, L = leucine, V = valine) and two of the three aromatic amino acids (F = phenylalanine, Y = tyrosine). We also see methionine (M) and its modified form (m) are *not* frequently confused. This suggests a potential blind spot of evolutionary pretraining, which does not separately represent modified and unmodified amino acids.

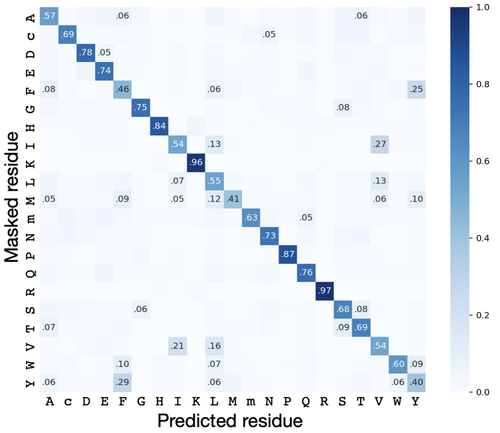
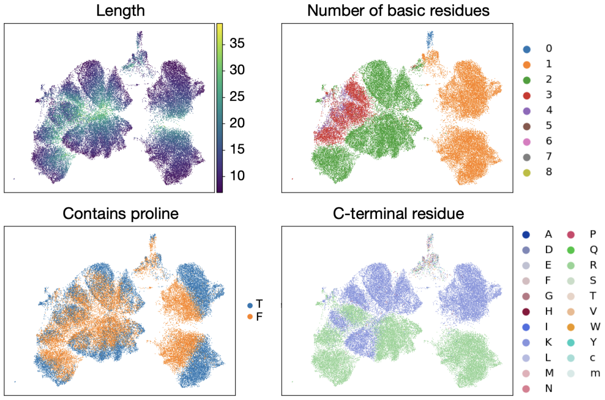


Figure 2. Confusion matrix for the pretext task between masked (rows) and predicted (columns) amino acids. Color indicates $P(\text{column} \mid \text{row})$. Entries > 0.05 are labelled.

4.2. Interpretation of embeddings

We embed each sequence in our test set, average-pool lengthwise, and visualize the result via UMAP (McInnes et al., 2018) in Figure 3. This indicates our sequence representation captures known determinants of peptide fragmentation. Length dependence is apparent, as is clustering according to: the number of basic amino acids, which determines the maximum charge a peptide can carry and where it localizes (Paizs & Suhai, 2005); the presence of proline (P), which bends the peptide (Vaisar & Urban, 1996); and the identity of the C-terminal amino acid, which reflects a selection bias from using trypsin to digest proteins into peptides.


Length

Number of basic residues

Contains proline

C-terminal residue

Figure 3. UMAP visualization of peptide embeddings, colored according to properties known to influence fragmentation.

4.3. Downstream performance

Test AUC of our model and the baselines for the four tasks considered are shown in Table 1. On the CDC28 BINDING and SIGNAL PEPTIDE tasks, a model pretrained on mass spectrometry is competitive with the state-of-the-art evolutionary approach. Our approach also outperforms state-of-the-art on MITOCHONDRIAL TARGETING. We suggest this is due to our choice of data: mass spectral peak intensities are substantially determined by the distribution of electric charge within the peptide, which is also known to determine mitochondrial targeting (Zarin et al., 2021). However, our approach fares poorly on MHC BINDING; the lower performance overall on that task may have resulted from our decision to pool sequences across different MHC complexes.

The evolutionary pretext, trained on proteins, proves effective for peptides. This might be due to the ByteNet architecture, whose first few layers of convolutional filters necessarily detect peptide-sized features. But this evolutionary pretext also benefits from a much larger corpus than our peptide task: `CARP-600k` is trained on about 100× more sequences than our model, each much longer than a peptide.

Model		Mito	Cdc28	SignalP	MHC
CARP-600k	FR	0.87	0.73	**0.99**	0.72
	FT	0.86	**0.78**	**0.99**	0.73
Linear		0.86	0.56	0.82	0.67
3-layer CNN		0.85	0.71	0.92	**0.75**
Random init.	FR	0.85	0.71	**0.99**	0.72
	FT	0.83	0.74	0.94	0.68
MS pretrained	FR	**0.89**	**0.78**	0.97	0.70
	FT	**0.89**	0.71	**0.99**	0.72

Table 1. Averaged 3-fold test AUC on downstream tasks, for: CARP-600K (SOTA); linear and CNN baselines; our model, initalized from random; and pretraining on the MS pretext. 'FR' only trains the final classifier; 'FT' additionally trains the encoder.

5. Conclusion

Here we show existing published mass spectrometry data can be used to derive representations for peptide property prediction that are competitive with a state-of-the-art evolutionary pretext, while using far fewer sequences. Evolutionary and mass-spectral pretraining need not be mutually exclusive: both enjoy plentiful data and may offer complementary views of peptide and protein structure – particularly for modified amino acids, which are not explicitly represented in evolutionary data, yet strongly influence protein structure and function (Mann & Jensen, 2003) and are represented with greater diversity in other ProteomeTools releases (Zolg et al., 2018). Integration of these two modalities is a promising avenue for further exploration, especially for peptide design tasks in which modifications or peptidomimetics (Gatto et al., 2021) are included in the sequence search space.

References

Bepler, T. and Berger, B. Learning protein sequence embeddings using information from structure. 2019. doi: 10.48550/ARXIV.1902.08661. URL https://arxiv.org/abs/1902.08661.

Chen, X., Li, C., Bernards, M. T., Shi, Y., Shao, Q., and He, Y. Sequence-based peptide identification, generation, and property prediction with deep learning: a review. *Molecular Systems Design &amp Engineering*, 6(6):406–428, 2021. doi: 10.1039/d0me00161a. URL https://doi.org/10.1039/d0me00161a.

Dee, W. LMPred: predicting antimicrobial peptides using pre-trained language models and deep learning. *Bioinformatics Advances*, 2(1), January 2022. doi: 10.1093/bioadv/vbac021. URL https://doi.org/10.1093/bioadv/vbac021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/tpami.2021.3095381. URL https://doi.org/10.1109/tpami.2021.3095381.

Fosgerau, K. and Hoffmann, T. Peptide therapeutics: current status and future directions. *Drug Discovery Today*, 20(1):122–128, January 2015. doi: 10.1016/j.drudis.2014.10.003. URL https://doi.org/10.1016/j.drudis.2014.10.003.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, October 2012. doi: 10.1093/bioinformatics/bts565. URL https://doi.org/10.1093/bioinformatics/bts565.

Gatto, A. D., Cobb, S. L., Zhang, J., and Zaccaro, L. Editorial: Peptidomimetics: Synthetic tools for drug discovery and development. *Frontiers in Chemistry*, 9, November 2021. doi: 10.3389/fchem.2021.802120. URL https://doi.org/10.3389/fchem.2021.802120.

Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B., and Wilhelm, M. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509–518, May 2019. doi: 10.1038/s41592-019-0426-7. URL https://doi.org/10.1038/s41592-019-0426-7.

Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., and Kavukcuoglu, K. Neural machine translation in linear time. *CoRR*, abs/1610.10099, 2016. URL http://arxiv.org/abs/1610.10099.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Mann, M. and Jensen, O. N. Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21(3): 255–261, March 2003. doi: 10.1038/nbt0303-255. URL https://doi.org/10.1038/nbt0303-255.

McInnes, L., Healy, J., Saul, N., and Großberger, L. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, September 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Paizs, B. and Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*, 24(4): 508–548, 2005. doi: 10.1002/mas.20024. URL https://doi.org/10.1002/mas.20024.

Qi, Y., Oja, M., Weston, J., and Noble, W. S. A unified multitask architecture for predicting local protein properties. *PLoS ONE*, 7(3):e32235, March 2012. doi: 10.1371/journal.pone.0032235. URL https://doi.org/10.1371/journal.pone.0032235.

Qiao, R., Tran, N. H., Xin, L., Chen, X., Li, M., Shan, B., and Ghodsi, A. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3(5):420–425, March 2021. doi: 10.1038/s42256-021-00304-3. URL https://doi.org/10.1038/s42256-021-00304-3.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, 2019.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), April 2021. doi: 10.1073/pnas.2016239118. URL https://doi.org/10.1073/pnas.2016239118.

Schissel, C. K., Mohapatra, S., Wolfe, J. M., Fadzen, C. M., Bellovoda, K., Wu, C.-L., Wood, J. A., Malmberg, A. B., Loas, A., Gómez-Bombarelli, R., and Pentelute, B. L. Interpretable deep learning for de novo design of cell-penetrating abiotic polymers. *bioRxiv*, 2020. doi: 10.1101/2020.04.10.036566. URL https://www.biorxiv.org/content/10.1101/2020.04.10.036566v1.

Steen, H. and Mann, M. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5 (9):699–711, September 2004. doi: 10.1038/nrm1468. URL https://doi.org/10.1038/nrm1468.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and and, C. H. W. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, November 2014. doi: 10.1093/bioinformatics/btu739. URL https://doi.org/10.1093/bioinformatics/btu739.

Teufel, F., Armenteros, J. J. A., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, January 2022. doi: 10.1038/s41587-021-01156-3. URL https://doi.org/10.1038/s41587-021-01156-3.

Tsaprailis, G., Nair, H., Somogyi, Á., Wysocki, V. H., Zhong, W., Futrell, J. H., Summerfield, S. G., and Gaskell, S. J. Influence of secondary structure on the fragmentation of protonated peptides. *Journal of the American Chemical Society*, 121(22):5142–5154, May 1999. doi: 10.1021/ja982980h. URL https://doi.org/10.1021/ja982980h.

Vaisar, T. and Urban, J. Probing proline effect in CID of protonated peptides. *Journal of Mass Spectrometry*, 31(10):1185–1187, October 1996. doi: 10.1002/(sici)1096-9888(199610)31:10⟨1185::aid-jms396⟩3.0.co;2-q. URL https://doi.org/10.1002/(sici)1096-9888(199610)31:10<1185::aid-jms396>3.0.co;2-q.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Wan, F., Kontogiorgos-Heintz, D., and de la Fuente-Nunez, C. Deep generative models for peptide design. *Digital Discovery*, 2022. doi: 10.1039/d1dd00024a. URL https://doi.org/10.1039/d1dd00024a.

Yang, K. K., Wu, Z., and Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, July 2019. doi: 10.1038/s41592-019-0496-6. URL https://doi.org/10.1038/s41592-019-0496-6.

Yang, K. K., Lu, A. X., and Fusi, N. Convolutions are competitive with transformers for protein sequence pretraining. In *ICLR2022 Machine Learning for Drug Discovery*, 2022. URL https://openreview.net/forum?id=3i7WPak2sCx.

Yilmaz, M., Fondrie, W. E., Bittremieux, W., Oh, S., and Noble, W. S. ide novo/i mass spectrometry peptide sequencing with a transformer model. February 2022. doi: 10.1101/2022.02.07.479481. URL https://doi.org/10.1101/2022.02.07.479481.

Zarin, T., Strome, B., Peng, G., Pritišanac, I., Forman-Kay, J. D., and Moses, A. M. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *eLife*, 10, February 2021. doi: 10.7554/elife.60220. URL https://doi.org/10.7554/elife.60220.

Zhang, G. L., Lin, H. H., Keskin, D. B., Reinherz, E. L., and Brusic, V. Dana-farber repository for machine learning in immunology. *Journal of Immunological Methods*, 374(1-2):18–25, November 2011. doi: 10.1016/j.jim.2011.07.007. URL https://doi.org/10.1016/j.jim.2011.07.007.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022a.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining, 2022b. URL https://arxiv.org/abs/2203.06125.

Zhou, X.-X., Zeng, W.-F., Chi, H., Luo, C., Liu, C., Zhan, J., He, S.-M., and Zhang, Z. pDeep: Predicting MS/MS spectra of peptides with deep learning. *Analytical Chemistry*, 89(23):12690–12697, November 2017. doi: 10.1021/acs.analchem.7b02566. URL https://doi.org/10.1021/acs.analchem.7b02566.

Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H.-C., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., Deutsch, E. W., Aebersold, R., Moritz, R. L., Wenschuh, H., Moehring, T., Aiche, S., Huhmer, A., Reimer, U., and Kuster,

B. Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods*, 14(3):259–262, January 2017. doi: 10.1038/nmeth.4153. URL https://doi.org/10.1038/nmeth.4153.

Zolg, D. P., Wilhelm, M., Schmidt, T., Médard, G., Zerweck, J., Knaute, T., Wenschuh, H., Reimer, U., Schnatbaum, K., and Kuster, B. ProteomeTools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Molecular &amp Cellular Proteomics*, 17(9):1850–1863, September 2018. doi: 10.1074/mcp.tir118.000783. URL https://doi.org/10.1074/mcp.tir118.000783.