EProM - Exploration of Protein Modifications

Daniël M. Bot, Jannes Peeters, Jan Aerts

EProM - Exploration of Protein Modifications

Daniël M. Bot[§,*]
I-BioStat, Data Science Institute, Hasselt University, Hasselt, Belgium

Jannes Peeters[§,†]
I-BioStat, Data Science Institute, Hasselt University, Hasselt, Belgium

Jan Aerts[‡]
Amador Bioscience, Hasselt, Belgium; Hasselt University, Hasselt, Belgium

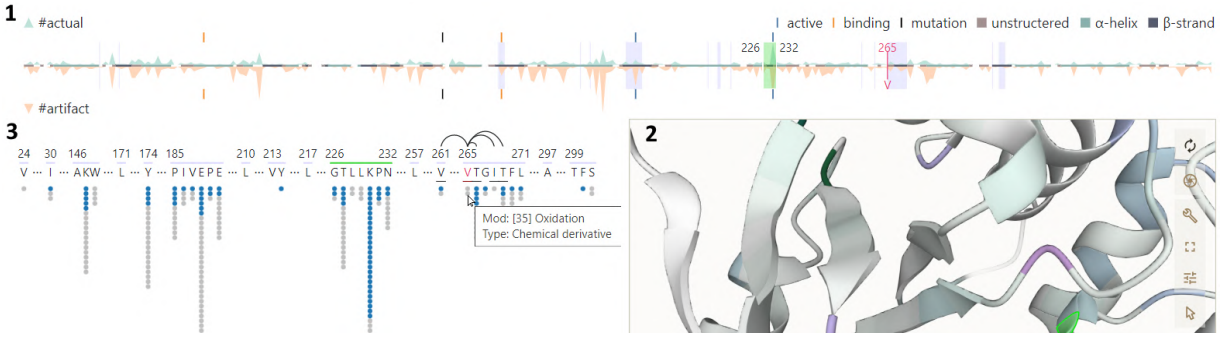

Figure 1: Main EProM visual analytics interface consisting of (1) a browser view used to navigate through the protein's primary structure; (2) a Mol* viewer showing the number of modifications in the protein's 3D structure; and (3) a detail plot visualizing modifications of interest measured in the selected area. Full version at Figure A1

Abstract

We present EProM—a visual analysis interface for the exploration of protein modifications—as a contribution to the IEEE VIS 2022 Bio+MedVis Challenge. The interface targets researchers in biochemistry, proteomics, and precision medicine as its primary users. Observed modifications can be inspected from the protein's primary, secondary, and tertiary structure, using a straightforward design and intuitive interactions. Modifications' measurement uncertainty and relation to residues with identified pathogenic mutations are considered.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics; Life and medical sciences—Computational biology—Computational proteomics

1 Introduction

Recent studies have revealed more protein modifications than initially expected [3, 6], to such extent that visualization becomes difficult. Hence, this year's Bio+MedVis Challenge (http://biovis.net/2022/biovisChallenges_vis) addresses the need to revise current protein modification visualization methods.

Relating modifications to a protein's 3D structure is essential for interpreting how these modifications alter the protein's behavior, especially when modifications occur near residues that are known to be involved in mutations related to rare diseases. This paper presents an intuitive visual interface designed to interpret such modifications within the context of a protein's structure and to assess their relation to residues with known pathogenic mutations.

*e-mail: jelmer.bot@uhasselt.be
†e-mail: jannes.peeters@uhasselt.be
‡e-mail: jan.aerts@uhasselt.be
§These authors contributed equally to this work.

1.1 Data

The data for this year's challenge was assembled by researchers in the CompOmics group at VIB and Ghent University (www.compomics.com) and includes information on three proteins related to rare diseases [1]: (1) Aldolase A (ALDOA); (2) Heterogeneous nuclear ribonucleoprotein A1 (HNRNPA1); and (3) Transforming growth factor beta 1 (TGFB1). A list of identified modifications with their position and classification is provided for each protein, alongside their AlphaFold structures [7, 9].

We have enriched the data by incorporating UniProt's protein descriptions, including their full name, function description, active sites, and binding sites [4]. In addition, we have approximated the modifications' position uncertainty by computing a list of residues at which each modification can occur, considering the amino acids that modification can bind to according to UniMod [5] located within five positions of the original residue. Several modifications occurred on amino acids not listed as possible sites in UniMod (Table A1).

2 Main Challenge

Our interface is designed for biochemistry, proteomics and precision medicine researchers, with the objective to (1) relate modifications to the protein's 3D structure, and (2) identify modifications that occur near residues with pathogenic mutations.

We settled on a design containing three views (Figure 1): (1) a protein browser, (2) a Mol* viewer [8], and (3) a detail view. The final design was implemented using Svelte (https://svelte.dev) and D3.js [2] and deployed with SvelteKit (https://kit.svelte.dev) at `https://biovis2022.vercel.app/`.

The remainder of this section describes each view in detail.

2.1 2D Plot: Protein Browser

The top-most view (Figure 1.1) represents the protein primary structure with annotations for the secondary structure: it can be used to select a residue sequence to focus on. This plot visualizes: (1) the number of modifications along the primary structure, distinguishing *artefacts* (i.e., "Artefact" or "Chemical derivative") from *actual* modifications; (2) the secondary structure—unstructured, α-helixes, and β-strands—along the x-axis; and (3) the residues related to pathogenic mutations, active sites, and binding sites.

2.2 3D Plot: Mol* Viewer

The second part (Figure 1.2) contains a Mol* viewer [8] showing the secondary and tertiary structures. The Mol* viewer was adapted to color residues by their number of modifications. The view automatically zooms in on the residues selected in the protein browser (Figure 1.1) and highlights these residues in green. In addition, hovering over a residue highlights it in red in all three views.

2.3 Detail Plot

The final view (Figure 1.3) contains a detail plot that facilitates exploring all modifications on the selected residues and residues in their 3D proximity. Residues are indicated by their amino acid's letter, emphasizing residues related to pathogenic mutations and separating non-consecutive residues by dots. The lines above the residue letters indicate whether those residues were selected (green) or in 3D-proximity to that selection (purple). The same colors are used to indicate the position of these residues in the 2D plot, making it easy to see whether active and binding sites are proximal to each other (see Figure A1). Hovering over a residue highlights it in red in all three views, and a mouse click opens a modal window that lists all observed modifications on that residue.

All modifications are represented by a circle stacked beneath the residue they were observed on. Hovering over a modification highlights all additional residues the modification can occur on (using arcs above the sequence), indicating its measurement uncertainty. In addition, hovering triggers a tree visualization showing which amino acids that modification can bind to according to UniMod [5]. The modifications can be colored interactively by selecting a combination of classification and modification name. The same combinations can be used to filter which modifications are shown.

3 Complementary Challenge: Redesign

The 2022 Bio+MedVis Challenge also includes the secondary task of revising and improving an existing visualization (see http://biovis.net/2022/biovisChallenges_vis/). We identified several aspects that can be improved in a redesign. Most importantly, the vertical lines clutter the view without providing information and the modification's circles overlap to an extent that it reduces legibility. In addition, the absence of a color legend makes it impossible to determine what the colors mean.

Our redesign was developed with the objective to provide an intuitive overview of the possible modifications for the entire protein and per residue. We settled on a polar plot that shows all modifications as circles colored by their classification, mapping the modification name to the angle and residue position to the radius (Figure 2). The color scheme assigns a monochromatic hue to the **-translational* and **-glycosylation* classes. Modifications classified as "Artefact" or "Chemical derivative" were given a neutral color and positioned behind the other modifications.

The visualization supports several interactions and is combined with a stripped-down vertical version of the protein browser. Both visualizations can select a section of the protein sequence through brushing. The selected section is automatically highlighted in the protein browser. In addition, the hovered residue is indicated by a red circle in the polar plot and a red line in the protein browser. Finally, the mouse's position is observed to draw a line from the center, through the mouse, to the border of the circle, indicating the modification name of modifications along that line.

4 Conclusion

In this paper, we presented an interface for exploring protein modifications in the context of the protein structures, and demonstrated how uncertainty in a modification's exact position could be incorporated.

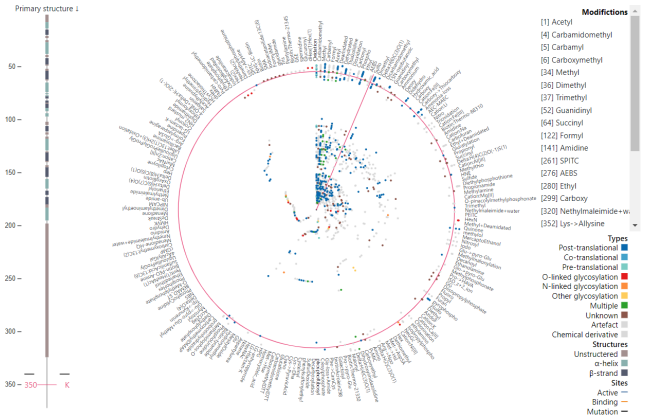

Figure 2: Our redesign. Modifications are shown as dots in a polar plot and colored by their classification. The red circle indicates the hovered residue's modifications, and the red line indicates the mouse position on the radial modification scale. Full version at Figure A2

Acknowledgments

This work was supported in part by Hasselt University BOF grants ADMIRE [BOF21GP17] and [BOF21DOC19], and the Flemish Government programme "Onderzoeksprogramma Artificiele Intelligentie (AI)". JA is supported by Amador Bioscience. We thank Lennart Martens for providing the domain background.

References

[1] National library of medicine. `https://www.ncbi.nlm.nih.gov/clinvar/`. Accessed: 2022/07/18.

[2] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185

[3] R. Bouwmeester, R. Gabriels, N. Hulstaert, L. Martens, and S. Degroeve. Deeplc can predict retention times for peptides that carry as-yet unseen modifications. *Nat Methods*, 18(11):1363–1369, 2021. doi: 10.1038/s41592-021-01301-5

[4] T. U. Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020. doi: 10.1093/nar/gkaa1100

[5] D. M. Creasy and J. S. Cottrell. Unimod: Protein modifications for mass spectrometry. *PROTEOMICS*, 4(6):1534–1536, 2004. doi: 10.1002/pmic.200300744

[6] S. Degroeve, R. Gabriels, K. Velghe, R. Bouwmeester, N. Tichshenko, and L. Martens. ionbot: a novel, innovative and sensitive machine learning approach to lc-ms/ms peptide identification. *bioRxiv*, 2022. doi: 10.1101/2021.07.02.450686

[7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[8] D. Sehnal, S. Bittrich, M. Deshpande, R. Svobodová, K. Berka, V. Bazgier, S. Velankar, S. K. Burley, J. Koča, and A. S. Rose. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, 05 2021. doi: 10.1093/nar/gkab314

[9] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. doi: 10.1093/nar/gkab1061

A Appendix - Supplementary figures

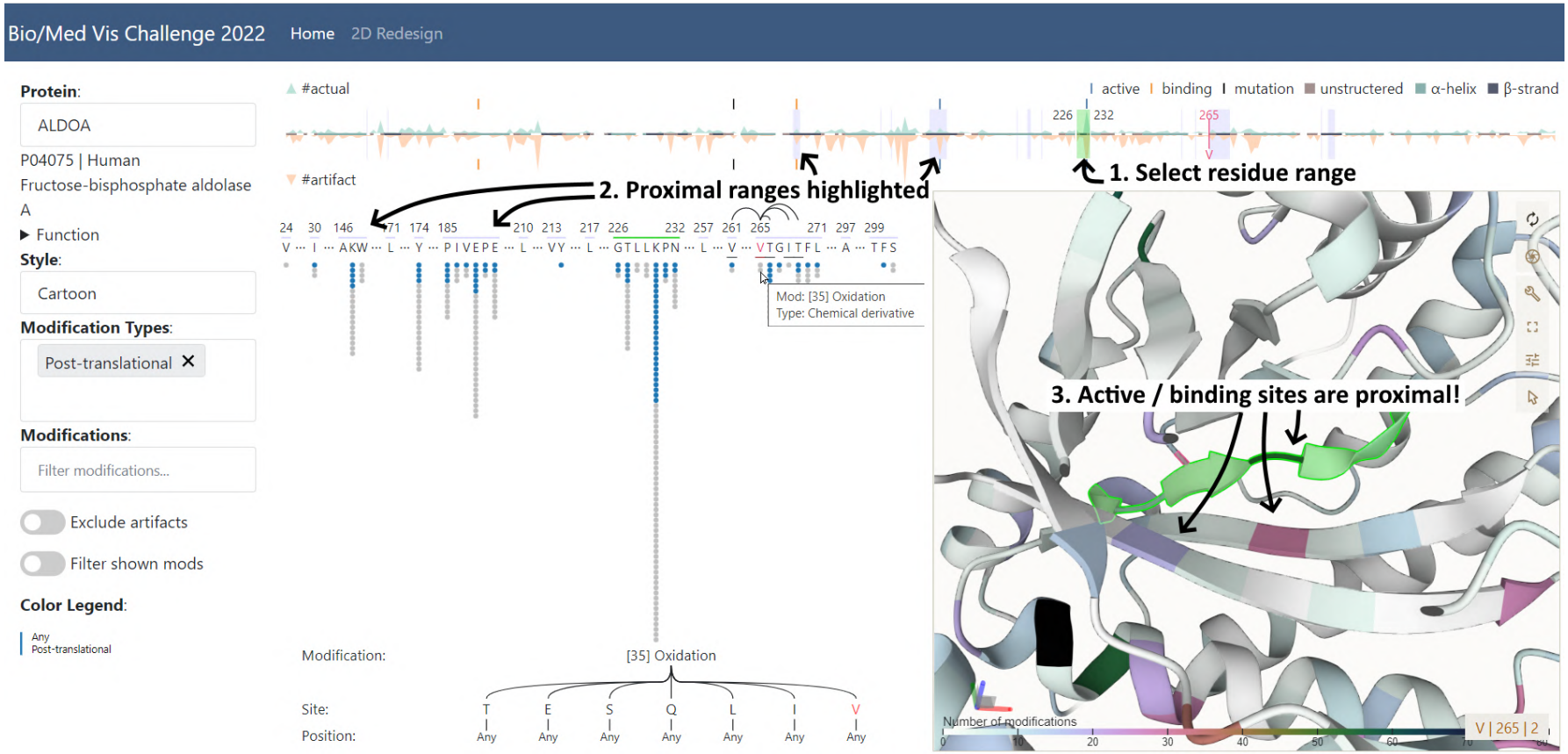

Figure A1: Full screenshot of the main interface.

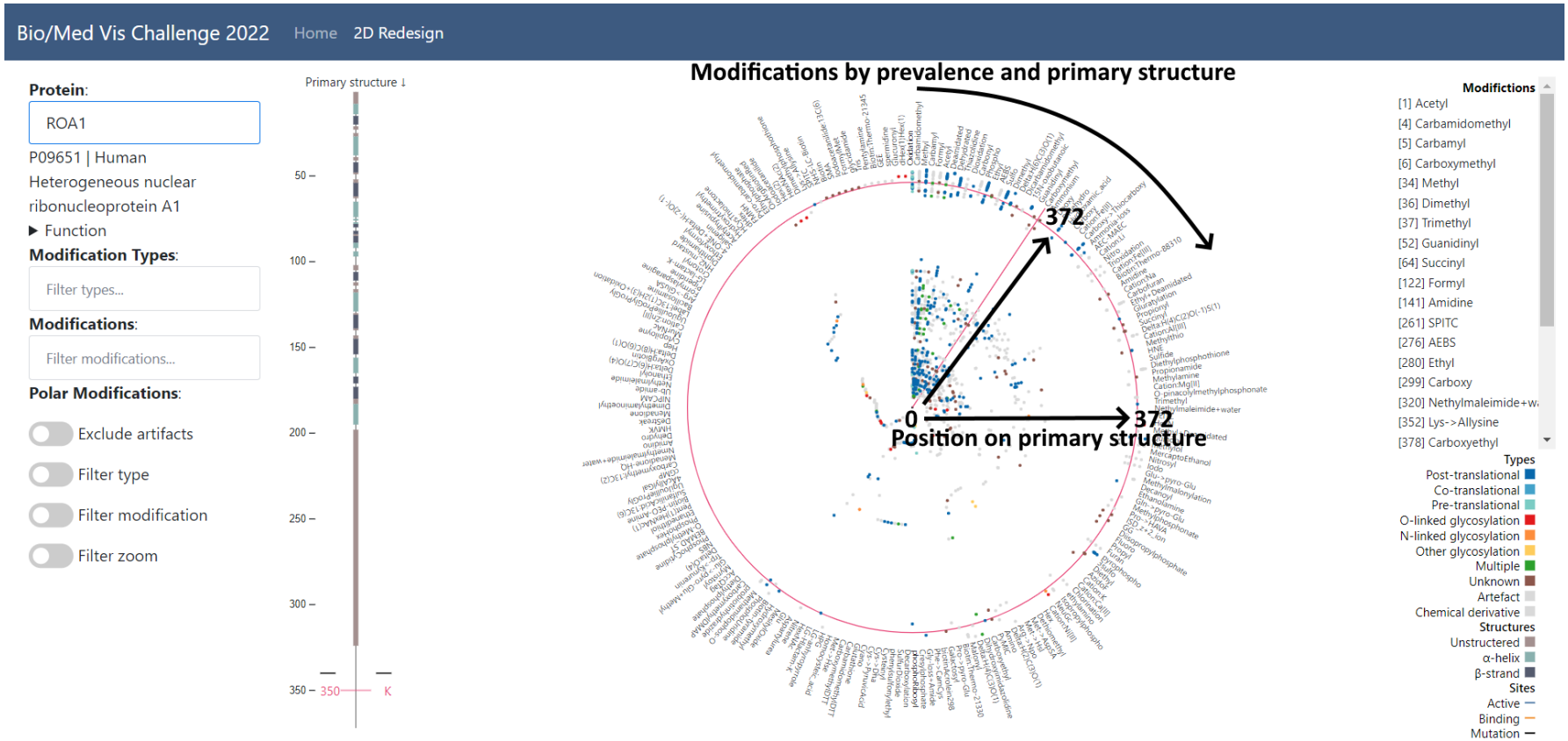

Figure A2: Full screenshot of the redesign.

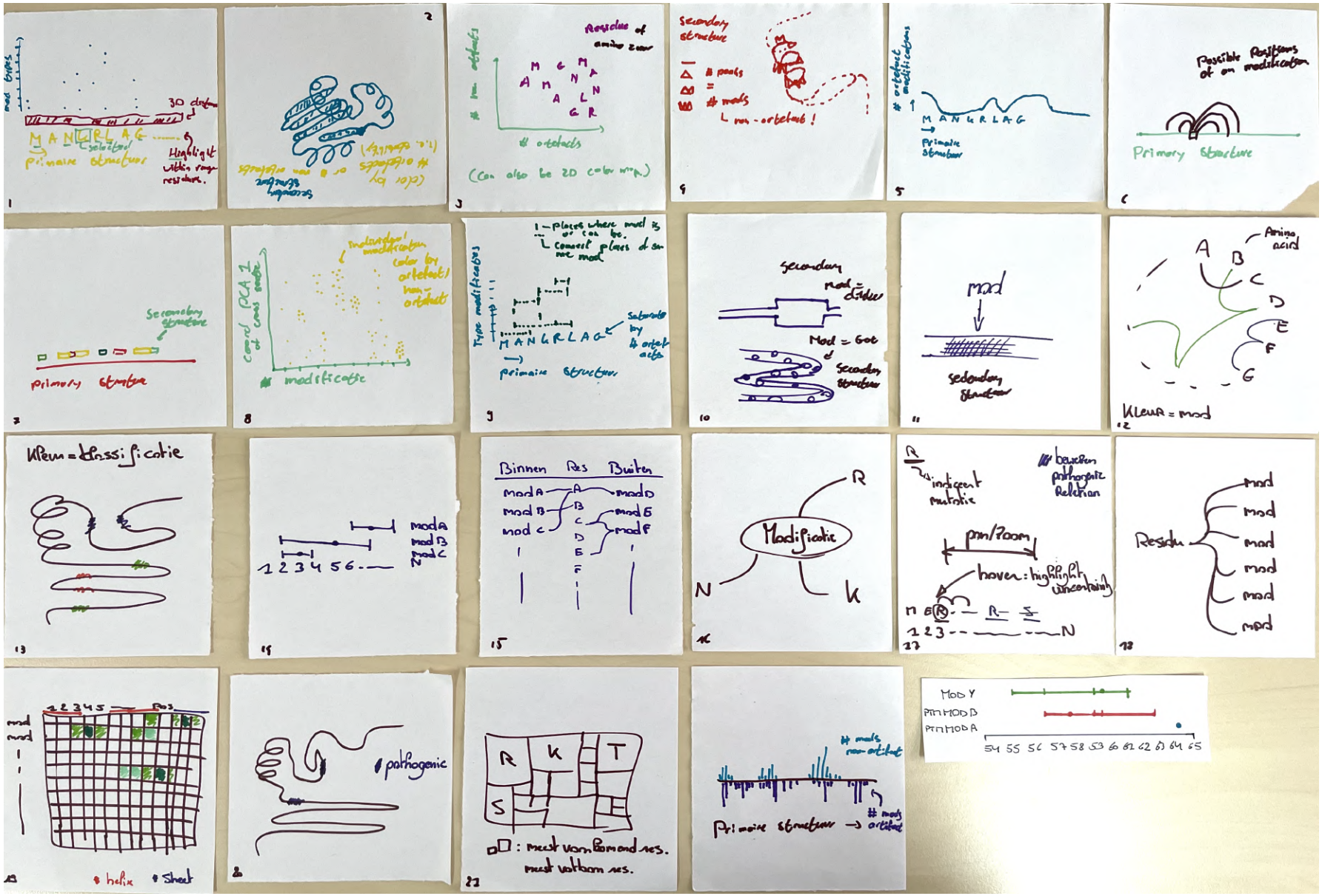

Figure A3: First sketches generated during the diverging phase.

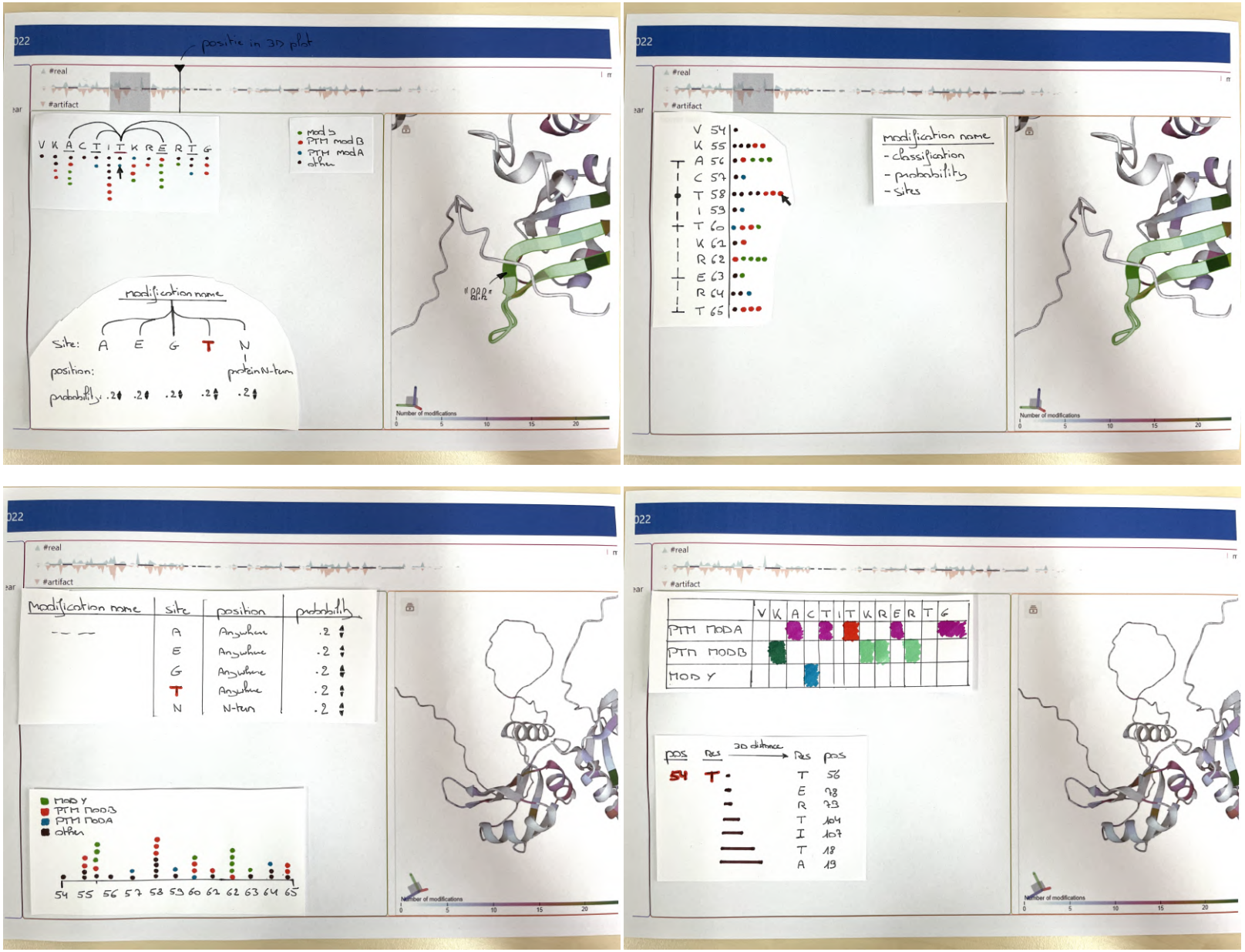

Figure A4: Pictures of the emerged designs.

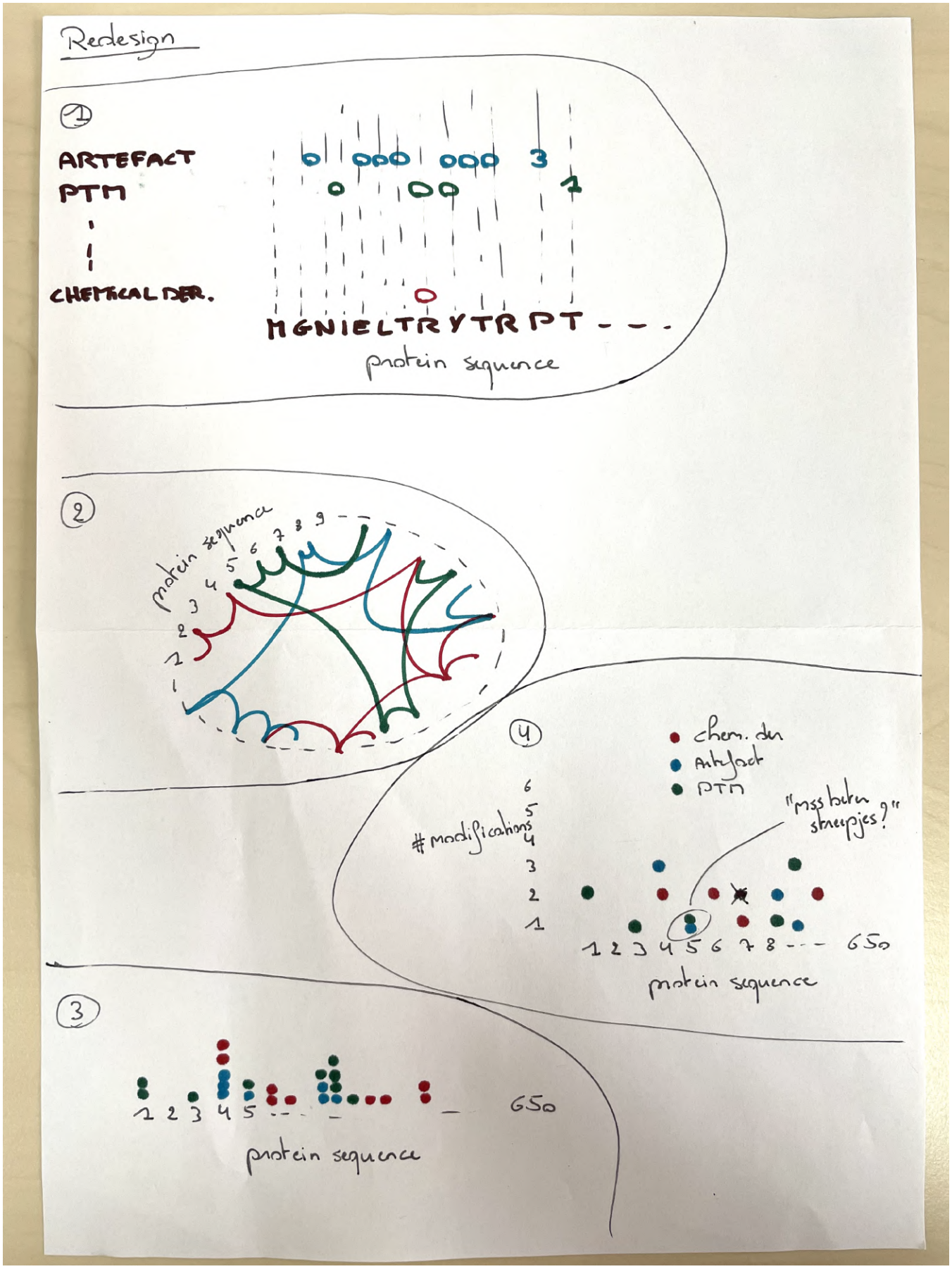

Figure A5: First sketches generated for the redesign exercise.

Table A1: UniMod Modification Mismatch.

Modification	Observed Sites	UniMod Sites	Not in UniMod
Acetyl	M,Y,K,S,T,R,H,C	R,Y,H,K,T,S,C	M
Thiazolidine	M,Y,K,R,H,C,F,W	W,Y,H,R,K,C,F	M
glycidamide	K,R	K	R
FormylMet	R		R
GG	R,M,K	C,T,S,K	R,M
Formyl	M,T,K,S,R	T,K,S	M,R
Ub-amide	C		C
Dicarbamidomethyl	M,K,R,C,H	K,H,C,R	M
Delta:H(6)C(3)O(1)	M,K,R,C,H	K,H,C	M,R
Delta:H(8)C(6)O(1)	R,K	K	R
biotinAcrolein298	R,K,H	H,K,C	R
CarboxymethylDMAP	R,K		R,K
Dansyl	K,R	K	R
Iodoacetanilide	K,R,C	K,C	R
Iodoacetanilide:13C(6)	K,C,R	K,C	R
Amidine	K,R	K	R
Biotin:Thermo-21330	M,K,R	K	M,R
Biotin:Thermo-21328	K,R	K	R
MesitylOxide	K,R,H	K,H	R
Delta:H(3)C(3)O(2)	K		K
AccQTag	K,R	K	R
Propionamide	M,R,K,C	C,K	M,R
SPITC	K,R	K	R
AEBS	K,Y,H,S,M,R	Y,S,K,H	M,R
Ethyl	K,D,E,R	E,D,K	R
SMA	R,K	K	R
Biotin	K,R	K	R
Methyl	M,T,E,K,I,L,D,S,R,Q,N,H	E,D,L,I,R,Q,N,K,H,C,S,T	M
Diisopropylphosphate	K,T,S,Y,R	K,Y,T,S	R
Methylthio	D,R,N,C,K	K,C,N,D	R
Carbamidomethyl	M,Y,E,K,D,H,R,T,S,C	Y,T,S,E,D,H,K,C,U,M	R
Phenylisocyanate	K		K
Phenylisocyanate:2H(5)	K,R		K,R
SPITC:13C(6)	K,R	K	R
PyMIC	R		R
LG-lactam-K	K,R	K	R
LG-Hlactam-K	R,K	K	R
Diethyl	K,R	K	R
Guanidinyl	M,K,R	K	M,R
Piperidine	K,R	K	R
Sulfo-NHS-LC-LC-Biotin	R	K	R
Propionyl	M,K,R,S	S,T,K	M,R
Carboxymethyl	M,K,R,C,W	K,C,W,U	M,R
Succinyl	K,R	K	R
Diethylphosphate	C,H,R,K	H,C,K,Y,T,S	R
Ethylphosphate	T,K,R,Y,S	K,Y,T,S	R
NO_SMX_SMCT	C		C
3sulfo	K,R		K,R
TNBS	R,K	K	R
Galactosyl	K,R	K	R
Ethoxyformyl	H		H
NHS-LC-Biotin	R	K	R
LG-anhydrolactam	K,R	K	R
LG-pyrrole	R,K	C,K	R
LG-anhyropyrrole	R,K	K	R
PEITC	K,C,R	K,C	R
ISD_z+2_ion	R,K		R,K
Hex-N-acetyl-D-glucosamine	N		N