A High-Throughput Platform for Efficient Exploration of Functional Polypeptides Chemical Space via Automation and Machine Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Rapid and in-depth exploration of the chemical space of high molecular weight synthetic polypeptides via the ring-opening polymerization (ROP) of N-carboxyanhydride (NCA) is a viable approach towards protein mimics and functional biomaterials. Here, we develop an efficient chemistry for the high throughput diversification of polypeptides based on a click-like reaction between selenolate and various electrophiles in aqueous solutions. With the assistance of automation and machine learning, iterative exploration of the random heteropolypeptides (RHPs) library efficiently and effectively identifies hit materials from a model system of which we have little prior knowledge. This automated and high-throughput platform provides a useful interface between wet and dry experiment, which would accelerate the discovery of new polypeptide materials for unmet challenges such as *de novo* design of artificial enzyme, biomacromolecule delivery, and understanding of intrinsically disordered proteins.

1 Introduction

Proteins are natural biopolymers with vast chemical space and sophisticated functions such as binding, catalysis, transportation and signaling. For decades, an overarching goal of polymer science is to create protein-like functional polymeric materials for not only fundamental understanding of proteins but also solving real-world challenges[1–5]. To this end, synthetic polypeptides prepared by the ring-opening polymerization (ROP) of N-carboxyanhydrides (NCA) have emerged as promising protein mimics with the potential to synergize the advantages of both peptides and synthetic polymers[6–11]. Specifically, polypeptides possess the same backbone and secondary conformations as protein and can be efficiently produced at up to kilogram scales in a high number-averaged molecular weight (M_n)[12]. Nevertheless, similar to other polymers, polypeptides are subjected to the curse of dimensionality, i.e., the combination of just a few residues can lead to a chemical space that is too large to be fully explored[13, 14]. To reach functional protein-mimicking polypeptides from the enormous chemical space, one needs to:

1. Facilely prepare polypeptides from the design space with high fidelity,
2. Establish an efficient strategy for effective exploration of the space at affordable labor and time cost.

To date, attempts to incorporate the NCA and polypeptide chemistry to high-throughput synthesis (HTS) workflow have been challenging and sparse[15]. This is partially attributable to the high moisture sensitivity of NCA, poor aqueous solubility of polypeptides, and laborious workup processes. Meanwhile, one underlying challenge for machine learning (ML)-assisted polymer design is the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

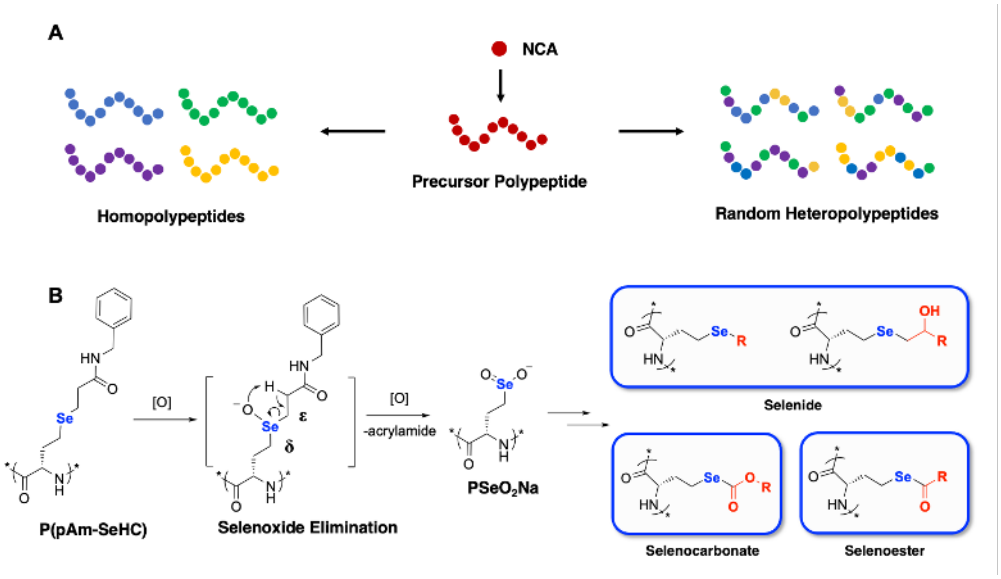

Figure 1: (A) Schematic illustration of the post polymerization modification strategy for making homopolypeptides (left) and random heteropolypeptides (RHP, right). (B) The selenopolypeptide $PSeO_2Na$, derived from the selenoxide elimination of P(pAm-SeHC), could be derivatized to various structures.

limited availability of high-quality data[16]. While most studies exploited data from literature and virtual experiments (e.g., electronic structure calculations or simulations), an ideal platform is capable of performing new experiments to support model training.

To address these challenges, we developed a high-throughput synthesis (HTS) platform in aqueous solutions for polypeptides based on a click-like reaction between selenolate and electrophiles (Figure 1). This quantitative chemistry gave accurate control of the structure and molecular composition of polypeptides. The process was amenable to off-the-shelf automated liquid handling platform and allowed efficient generation and purification of over 1200 polymers within one day, which greatly increase the synthesis capability of the materials. With the assistance of ML model-guided optimization, a closed-loop discovery workflow was established. We were able to perform iterative exploration of the RHP chemical space for enzyme mimics and identified candidates with improved activity in a more efficient and effective way.

2 Results

We designed and synthesized a selenopolypeptide, P(pAm-SeHC), whose pendant group is a latent selenolate, a highly nucleophilic species in organic chemistry. Oxidation of P(pAm-SeHC) leads to selenoxide elimination, generating $PSeO_2Na$ (Figure 1B). $PSeO_2Na$ could be reduced with $NaBH_4$ in water, affording the selenolate-bearing polypeptide (PSeNa) for further functionalization with organohalides (Figure 2). The modification of PSeNa showed remarkable tolerance to various functionalities. We were able to prepared many polypeptides with potential applications in biomedical engineering including those that are hard to be introduced directly through the ROP of NCA (Figure 2). When tried feeding with more than one organohalides, we found the molecular composition through the feeding volume ratio of the organohalides (Figure S1). A map from the feeding volume ratio to the molecular composition was thus directly created, which saves tremendous amount of times from additional characterization.

Based on these findings, the HTS of RHP was established with the assistance of a commercialized automated workstation for dispensing stock solutions of organohalides to plates. This semi-automated workflow greatly boosted the synthesis capability and enabled the parallel preparation of 400 RHP (4 plates) in one day. The throughput could be easily improved to 1200 RHP (12 plates) per day if only activated organohalides were used for modification. We applied a Bayesian optimization (BO)

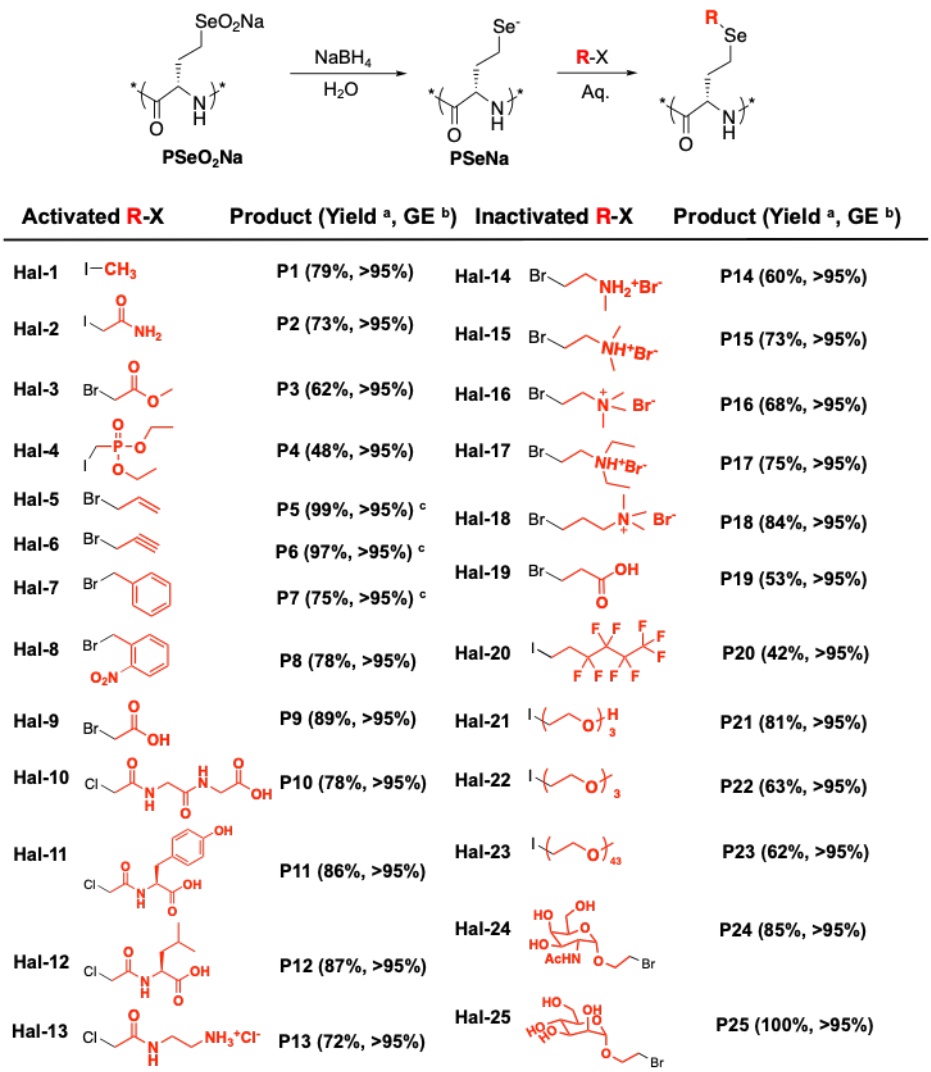

Figure 2: Post polymerization of $PSeO_2Na$ with activated and inactivated halides.[a] yield = purification yield, [b] GE: grafting efficiency based on the ^1H NMR spectra of the product, [c] Prepared from PEG-*b*-$PSeO_2Na$, a poly(ethylene glycol) block copolymer of $PSeO_2Na$.

framework based on BoTorch and Ax[17] and established a closed-loop material discovery workflow (Figure 3A).

For demonstration, we chose glutathione peroxidase-like (GPx-like) activity[18] of the materials as a optimization target (Figure 3A). Seven organohalides with different properties were selected for modification (Figure 3B). Within 4 days, four iterations comprising a total of 660 experiments were performed with 166 experiments per iteration. We did not find any similar system in the literature. So 166 RHPs were randomly chosen initially from the designed space to train a Gaussian process (GP) regression model. Candidates for successive iterations were chosen by selecting compositions that optimized an expected improvement (EI) acquisition function, subject to the constraint that total mole fractions equal 1. To avoid trapping into local minimums, random search and BO were performed simultaneously in each round. Both were used to select 83 RHP compositions to synthesize, evaluate, and retrain the surrogate GP before proposing the candidates for the next iteration. The results showed that while random search consistently found candidates with activities near a range of 150-200, BO efficiently found RHPs with substantially higher GPx-like activity, particularly in the third and fourth iteration (Figure 3C). T-distributed stochastic neighbor embeddings (t-SNE) showed that BO quickly identified an area in the design space that achieves higher activity (Figure S2). Two hits from BO

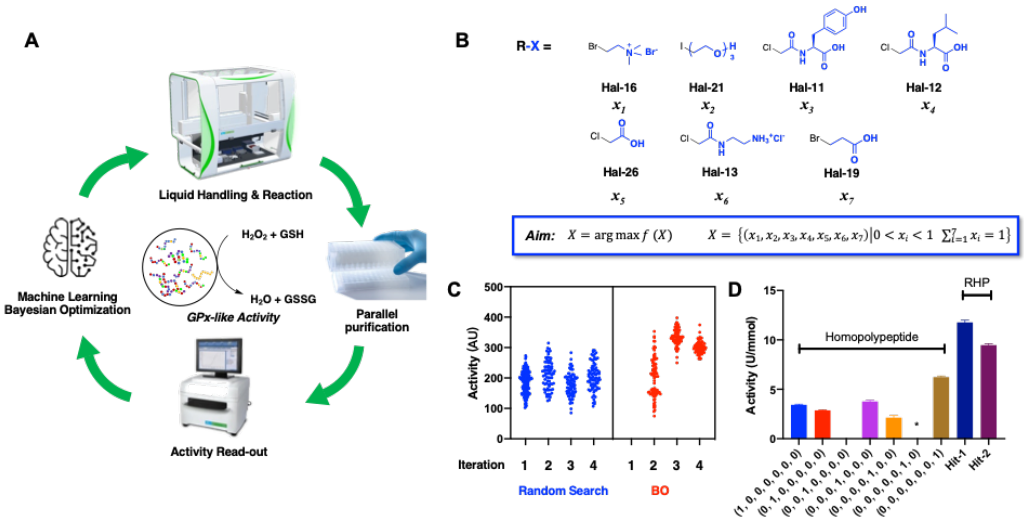

Figure 3: Closed-loop optimization of GPx activity of the RHPs via HTS and machine learning. (A) Cartoon illustration of the closed-loop workflow containing four modules, namely HTS, parallel purification, activity read-out, and Bayesian optimization. (B) Structure of the seven selected organohalides for RHP library generation and aim of optimization. The molecular composition of RHPs are descripted as seven-dimensional vectors to the program, $\boldsymbol{x} = (x_1, \ldots, x_7)$, where x_n (n = 1 to 7) is the relative volume ratio of the organohalide. (C) GPx-like activity of RHPs in each iteration via random searching (blue) or Bayesian optimization (red). (D) Comparison of the GPx-like activity of the two RHP hits with the seven homopolypeptides each modified with one individual organohalide used in HTS (n = 3). **Hit-1**: (0.12, 0.12, 0, 0, 0, 0, 0.76) and **Hit-2**: (0, 0.24, 0.22, 0, 0, 0, 0.54). All polymers are synthesized in flask and then purified for GPx activity. Error bars represent the standard deviation. *Activity of the homopolypeptide (0, 0, 0, 0, 0, 1, 0) cannot be measured properly owing to precipitation during testing.

were synthesized in flask and their GPx-like activity normalized to the amount of selenium was evaluated. **Hit-1** exhibited 2 times higher GPx-like activity than the most active homopolypeptide in the design space (Figure 3D), which meant that activity of RHPs is not merely the normalized average of the activity of each component.

3 Conclusion

In summary, we reported a robust, quantitative, and divergent strategy for the rapid expansion of polypeptide library based on a universal precursor selenopolypeptide. This post polymerization modification strategy avoided the laborious efforts of making a variety of NCA that are synthetically challenging. The potential of this modification chemistry was highlighted by the establishment of a closed-loop optimization workflow for the discovery of functional RHP. Enabled by the efficiency of the reaction, a map from the feeding volume ratio to the molecular composition was directly created. Because all polypeptides were derivatized from the same precursor, this strategy could be particularly useful to generate standardized dataset. Moreover, the HTS was performed in aqueous solutions and open air, which allowed convenient transferring of the resulting polymers to subsequent biological assays. As a proof-of-concept, we demonstrated a concise workflow enabling the rapid identification of RHPs with promising GPx-like activity. While detailed structure-activity relationship is still under investigation, these results underscored the power of this system in accelerating material discovery by exploring polypeptide chemical space of which people have little knowledge. We envision that the potential of this platform is far beyond artificial enzymes and can accelerate the discovery of antimicrobial agents, understanding of protein phase separation, and development of intracellular delivery systems for therapeutic biomacromolecules, *etc.*

References

(1) Cole, J. P.; Hanlon, A. M.; Rodriguez, K. J.; Berda, E. B. *Journal of Polymer Science Part a-Polymer Chemistry* **2017**, *55*, 191–206.

(2) Rothfuss, H.; Knofel, N. D.; Roesky, P. W.; Barner-Kowollik, C. *Journal of the American Chemical Society* **2018**, *140*, 5875–5881.

(3) Bonduelle, C. *Polymer Chemistry* **2018**, *9*, 1517–1529.

(4) Varanko, A. K.; Su, J. C.; Chilkoti, A. *Annual Review of Biomedical Engineering, Vol 22* **2020**, *22*, 343–369.

(5) Callmann, C. E.; Thompson, M. P.; Gianneschi, N. C. *Accounts of Chemical Research* **2020**, *53*, 400–413.

(6) Song, Z. Y.; Tan, Z. Z.; Cheng, J. J. *Macromolecules* **2019**, *52*, 8521–8539.

(7) Song, Z. Y.; Han, Z. Y.; Lv, S. X.; Chen, C. Y.; Chen, L.; Yin, L. C.; Cheng, J. J. *Chemical Society Reviews* **2017**, *46*, 6570–6599.

(8) Zhou, X. F.; Li, Z. B. *Advanced Healthcare Materials* **2018**, *7*, e1800020.

(9) Deng, C.; Wu, J. T.; Cheng, R.; Meng, F. H.; Klok, H. A.; Zhong, Z. Y. *Progress in Polymer Science* **2014**, *39*, 330–364.

(10) Hou, Y. Q.; Lu, H. *Bioconjugate Chemistry* **2019**, *30*, 1604–1616.

(11) Deming, T. J. *Progress in Polymer Science* **2007**, *32*, 858–875.

(12) Liu, Y.; Li, D.; Ding, J. X.; Chen, X. S. *Chinese Chemical Letters* **2020**, *31*, 3001–3014.

(13) Taylor, S. V.; Walter, K. U.; Kast, P.; Hilvert, D. *Proceedings of the National Academy of Sciences* **2001**, *98*, 10596–601.

(14) Reis, M.; Gusev, F.; Taylor, N. G.; Chung, S. H.; Verber, M. D.; Lee, Y. Z.; Isayev, O.; Leibfarth, F. A. *Journal of the American Chemical Society* **2021**, *143*, 17677–17689.

(15) Wyrsta, M. D.; Cogen, A. L.; Deming, T. J. *Journal of the American Chemical Society* **2001**, *123*, 12919–12920.

(16) Patel, R. A.; Borca, C. H.; Webb, M. A. *Molecular Systems Design and Engineering* **2022**, DOI: `10.1039/D1ME00160D`.

(17) Balandat, M.; Karrer, B.; Jiang, D. R.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. In *Advances in Neural Information Processing Systems 33*, 2020.

(18) Wikipedia contributors Glutathione peroxidase — Wikipedia, The Free Encyclopedia, [Online; accessed 24-September-2022], 2022.