AAPS Open

PERSPECTIVE

Open Access

Development of human-machine language interfaces for the visual analysis of complex biologics and RNA modalities and associated experimental data

Roxanne K. Kunz[1*], Atipat Rojnuckarin[2], Christian Marc Schmidt[3] and Les P. Miranda[1]

Abstract

The advent of recombinant protein-based therapeutic agents in the 1980s and subsequent waves of innovation in molecular biology and engineering of biologics has permitted the production of an increasingly broad array of complex, high molecular weight constructs. While this has opened a powerful new toolbox of molecular scaffolds with which to probe and interdict biological processes, it also makes deciphering the architectural nuances between individual constructs intuitively difficult. Key to downstream data processes for the detection of data trends is the ability to unambiguously identify, compare, and communicate the nature of molecular compositions. Existing small molecule orientated software tools are not intended for structures such as peptides, proteins, antibodies, and RNA, and do not contain adequate atomistic or domain-level detail to appropriately convey their higher structural complexity. Similarly, there is a paucity of large molecule-focused data analysis and visualization tools. This article will describe four new approaches we developed for the graphical representation and analysis of complex large molecules and experimental data. These tools help fulfill key needs in scientific communication and structure-property analysis of complex biologics and modified oligonucleotide-based drug candidates.

Keywords Protein engineering, Biotherapeutics, Oligonucleotides, Peptides, RNA drugs, Antibodies, Bioinformatics, Molecular visualization, Data management, Data visualization, Scientific communication

Introduction

Central to drug development processes is the design, make, test, and analyze iterative scientific cycle (Fig. 1). Efficient data analysis and downstream information dissemination workflows surrounding this cycle necessitate standardized methods for the unambiguous identification and representation of unique molecular entities, regardless of the level of structural complexity. The advent of modern computation and data processing reduced to practice the precise characterization of small molecule structures through the ability to generate standardized line notations and accompanying structural depictions. Most notably, indexable formats such as SMILES notation (Weininger 1988) or InChI strings (Heller et al. 2015) encode comprehensive atomistic, structural, stereoisomeric, and tautomeric substance attributes of small molecules into compressed forms for efficient storage and searching within computerized systems. However, comparable methods for large molecules such as peptides, proteins, antibodies, and RNA (5000–>150,000 Da)

*Correspondence:
Roxanne K. Kunz
kunz@amgen.com
[1] Process Development, Amgen Operations, One Amgen Center Drive, Thousand Oaks, CA 91320, USA
[2] Research Informatics, Information Systems, 360 Binney Street, Cambridge, MA 02142, USA
[3] Schema Design LLC., 620 12th Ave Suite 201, Seattle, WA 98122, USA

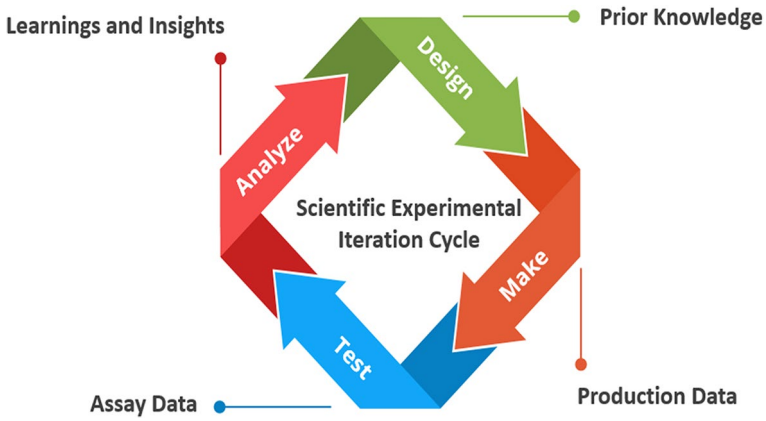

Fig. 1 Iterative scientific experimental cycle and accompanying information workflows

have remained largely inadequate and often rely on rudimentary text-based methods.

A contributing reason for capture and storage of large molecule entities as simple text strings containing only the primary sequence is the vastly higher structural complexity of such entities versus small molecules (Brinkmann and Kontermann 2017). Important identifying features of large molecule entities may be indicated through user-generated text labels, though such idiomatic names are not indexable for easy searching and hamper a common understanding of molecule composition. Additionally, attempts toward a universally recognized system for the nomenclature and terminology of engineered large molecules, such as IUPAC for small molecules (International Union of Pure and Applied Chemistry 2022), have not gained widespread adoption (Siani et al. 1994; Siani et al. 1995; Sweet-Jones et al. 2022; David et al. 2020). HELM, the Hierarchical Editing Language for Macromolecules, is arguably the most advanced entry in this area (Zhang et al. 2012; Pistoia Alliance 2023). While the HELM language has gone the furthest toward establishing a common standard notation, the accompanying graphical depictions are unwieldy for non-experts to interpret and do not intuitively reflect the two-dimensional compositional details of biomolecules. Thus, standardized, and extensible methods to encode large molecule compositions and subsequently generate graphical depictions of their two-dimensional architectures would be a valuable addition to biopharmaceutical drug development processes.

As such, we sought to develop four new informatic methods in this area: (1) a universal and lightweight solution to generate miniaturized versions of large molecule structures; (2) automated rendering of interactive graphics with depiction of key structural details for engineered large molecules; (3) a visual language to digitally generate schematics for the design of chemically modified siRNA; and (4) high-density data visualization interfaces for structure–property correlations of large molecule optimization data. These four informatic technologies facilitate the communication, interpretation, and analysis of highly engineered biotherapeutics among scientists and their external information consumers.

Method 1: BioFonts

A common feature in conventional representations of large molecule schematics is a lack of visual standardization. While there has been a recent notable entry in generating professional quality scientific figures (Biorender 2022), paywall restrictions and icon library limitations restrict broad access or free usage beyond educational purposes. An ability to dynamically generate uniform representations of large molecules without such limitations would thereby establish universal understanding between scientist inventors and their information partners. Consequently, we devised "BioFonts" as an OpenType font technology based on alphanumeric character strings that encode biomolecular domains into a suite of graphical objects (Fig. 2). The resulting glyphs form abstracted two-dimensional representations of large molecule structures.

The miniaturized glyphs are created by typing in a keypress code sequence which converts to the corresponding expanded and condensed version using the appropriate font typeface (e.g., IgG or BiTE®-scFc-siRNA suite; Fig. 3).

The encoded keys can be combined in various ways to build new biological construct architectures and further customized via regular font formatting options such as different colorations or sizing for distinct bio-modules (Fig. 4). The BioFont typefaces are bundled into a standard Unicode (The Unicode Consortium 2022) font

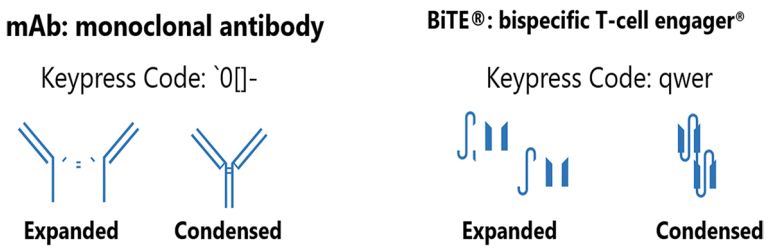

Fig. 2 Example Keypress codes for monoclonal antibody and BiTE® bispecific T-cell engager molecules and the resulting expanded and condensed glyphs created using BioFonts

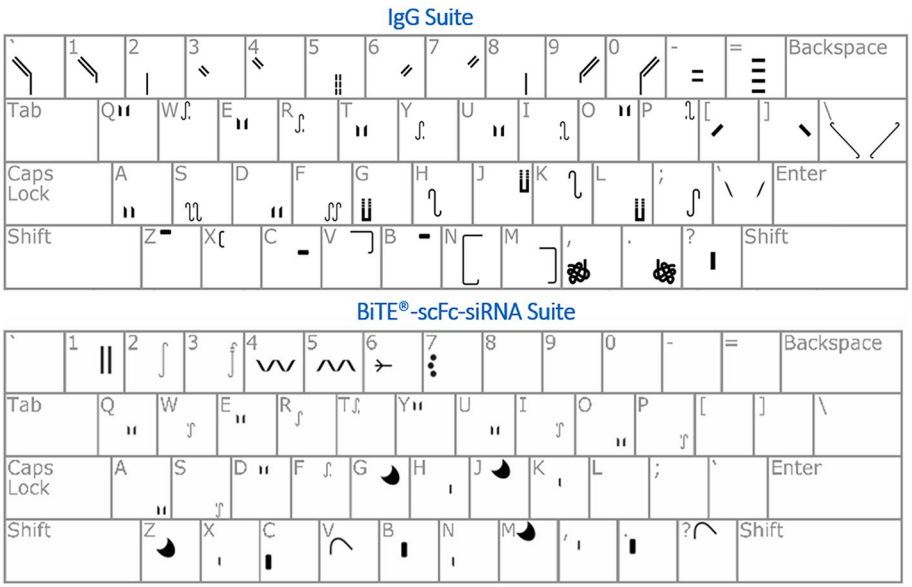

Fig. 3 Keyboard character mapping overlay for BioFonts IgG Suite and BiTE®-scFc-siRNA Suite

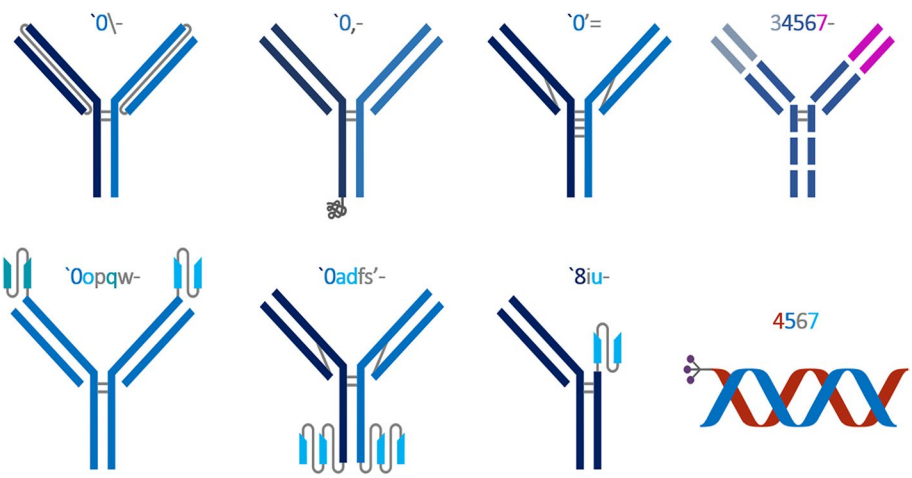

Fig. 4 Glyphs for select large molecule constructs created using BioFonts

package which can be installed on local desktops regardless of the operating system. The fonts are available in the font menu within standard business software applications without the need for any plug-in or integration. This method represents a lightweight and standardized method for rendering large molecule graphics into a high quality yet portable format and easily deployable to a broad set of users.

Method 2: BioMaps

Visually representing large molecules in a consistent and easy to interpret manner has been a long-standing challenge in bioinformatics. Historically, representation and comparison of macromolecular constructs generally did not extend past single-letter residue names, with important sequence features often indicated through human-generated text annotations (Chen et al. 2011). The imprecise and non-standardized nature of this process can cause ambiguity and lead to unintended misinterpretation of substances. Errors in sample identification may result and the ability to build a deeper understanding of associated optimization data and biophysical properties is undermined.

We utilized the principles of data reduction and compression to distill the higher complexity and size of large molecules into automated schematic representations that provide detailed and readily understandable visual renderings of two-dimensional architectures (Fig. 5).

The uniformly styled schematics are responsive to individual construct differences and algorithmically derived from a comprehensive sequence JSON file for each specific entity (Fig. 6).

The standardized nature of the schematics simplifies interpretation, facilitates understanding, and decreases confusion around the compositional differences of individual constructs (Fig. 7).

Sequence features and other metadata can be interactively displayed, including antibody sub-regions, disulfide bond pairing, variable fragment target associations, and germline information, by rendering the images as scaled vector graphic files. This technology eliminates the need for scientists to create ad hoc cartoons while also ensuring entities are uniformly represented, thereby removing sources of inconsistency and ambiguity around distinct molecular compositions in scientific communications.

Method 3: interactive design of chemically modified oligonucleotides

Recent advances in oligonucleotide-based therapeutics (e.g., siRNA, mRNA, anti-sense oligonucleotides, circular RNA) have been driven in part by the selected chemical modification of nucleotides. Moreover, siRNA candidates for therapeutics development are typically heavily chemically modified to impart stability and potency properties, and often contain no unmodified nucleotides (Khvorova and Watts 2017). The recent rapid growth (Setten et al.

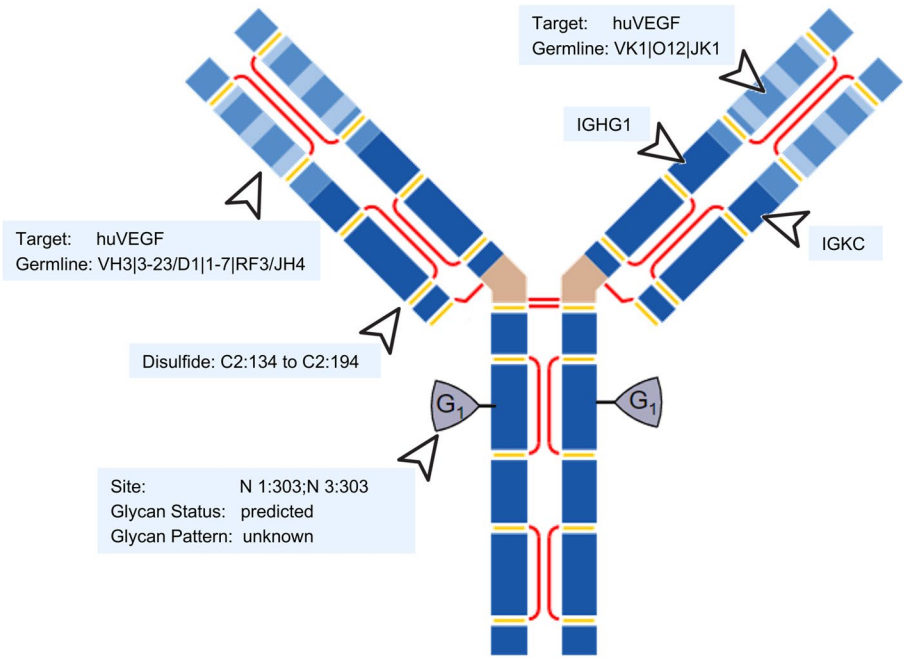

Fig. 5 An example of a IgG1 monoclonal antibody BioMap with an interactive display of sequence features and associated metadata

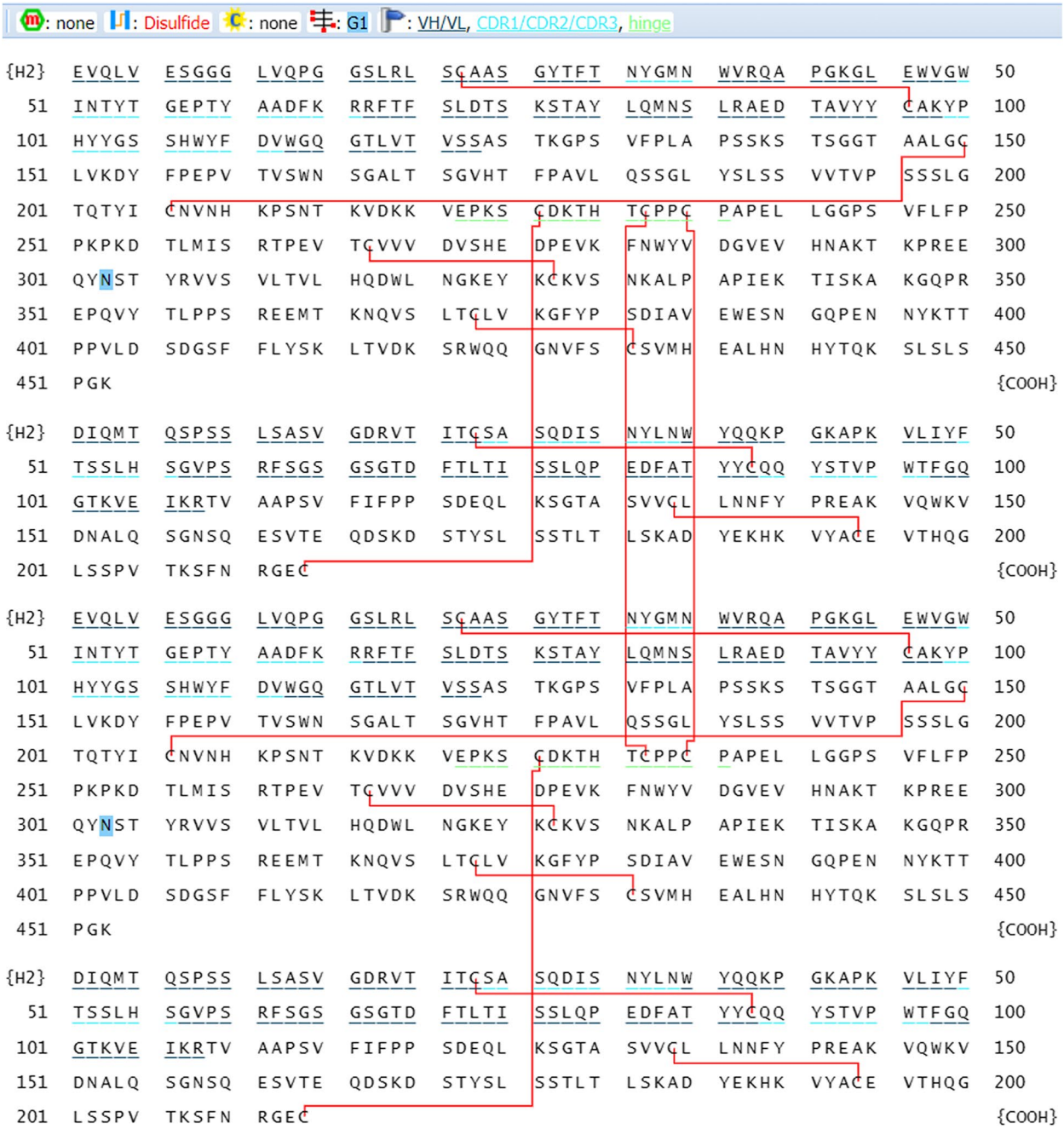

Fig. 6 BioMaps are dynamically rendered from a comprehensive sequence notation file for each molecule, in this case showing a monoclonal antibody

2019) of this field highlighted the need for a modern software tool customized to the specialized nature of siRNA optimization approaches. Manual workarounds in the form of offline documents are often used to describe, analyze, and communicate such structures. These error-prone processes are sorely inadequate for defining, tracking, and communicating variations in siRNA molecular composition. In this method, we combine the flexible nature of BioFonts with the interactivity of BioMaps to digitally encode chemically modified siRNA structures into interactive schematics using a new visual language. This tool takes the further step of enabling the fully interactive design of new siRNA molecular compositions (Fig. 8).

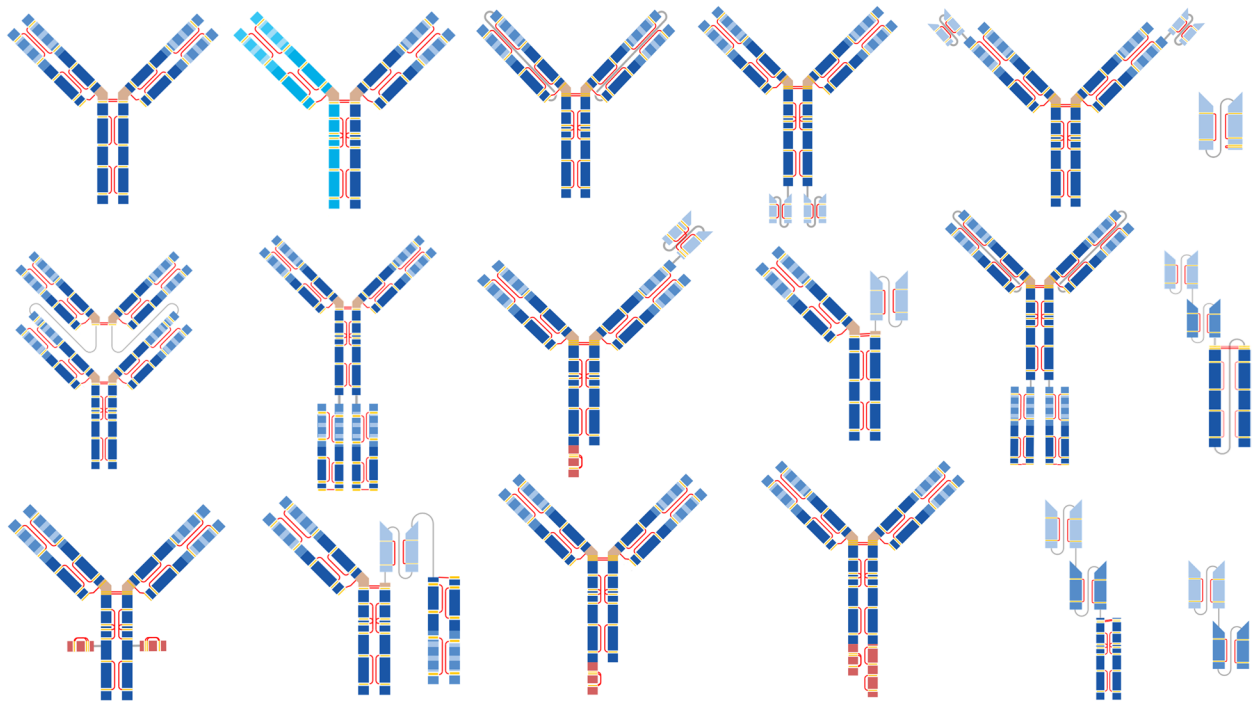

Fig. 7 Gallery of selected BioMaps for a variety of large molecule formats

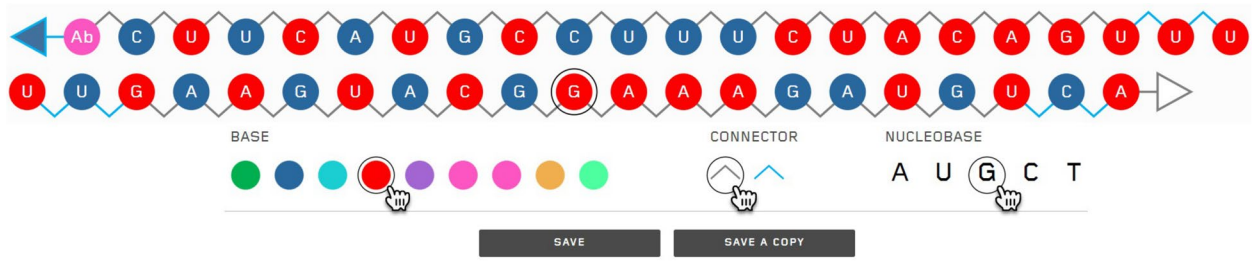

Fig. 8 An example of an siRNA duplex molecule with complete interactivity of each visual element

The complete atomic structure of chemically modified RNA molecules is captured via an internal structure dictionary of modified nucleoside base units. Various other structural elements, such as 2' chemical modifications, phosphorothioate vs. phosphodiester bonds, and glycan caps, are represented through a controlled vocabulary containing different shape and color specifications (Fig. 9).

Embedding the automated and fully interactive schematics within a dynamic web-based interface allows for the completely visual design of new siRNA therapeutic candidates through the chemistry menu selections (Fig. 10). With this tool, scientists can graphically design new siRNA sequences and export the accompanying images as informative and standardized scientific communication aids.

Method 4: large molecule multivariate data visualization

The pre-clinical drug discovery process generates large volumes of multidimensional data on the properties of therapeutic candidate molecules. Well-informed decision-making must then ensue with the aim of selecting the best candidate for clinical advancement. Often, these data are in offline, manually prepared documents that do not allow for efficient interrogation. Identifying data trends and learnings around the characteristics of large molecules becomes difficult and misinterpretations can occur. Additionally, existing cheminformatic SAR tools for small molecules are not suitable in this context and lack the necessary customization to adequately handle the higher complexity of large molecules. Thus, the principles of data reduction and visualization are essential for the flexible analysis

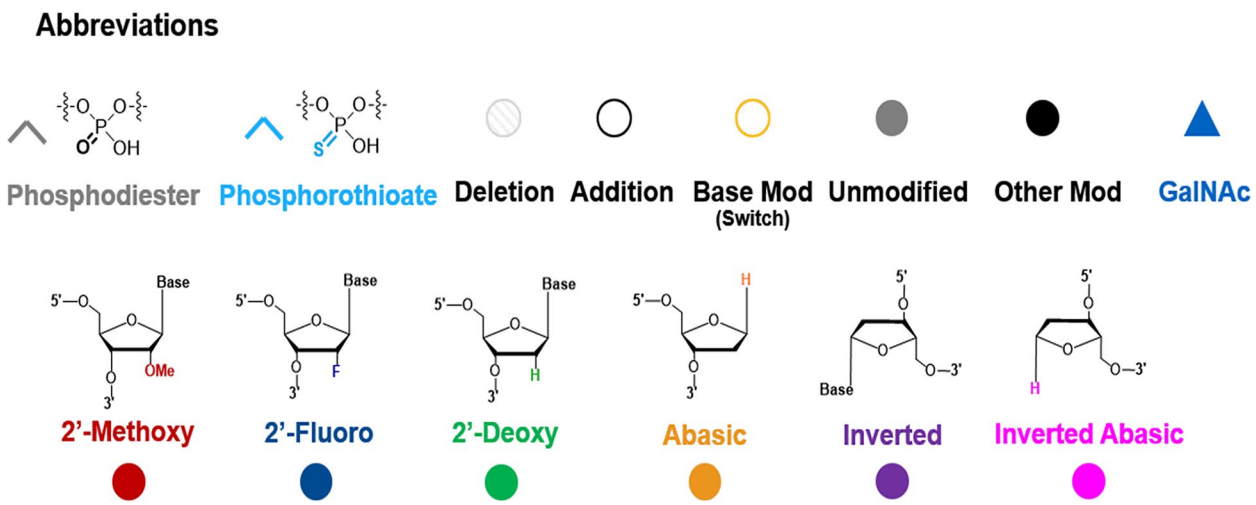

Fig. 9 Representative graphical encoding of siRNA structural units available through the internal controlled dictionary

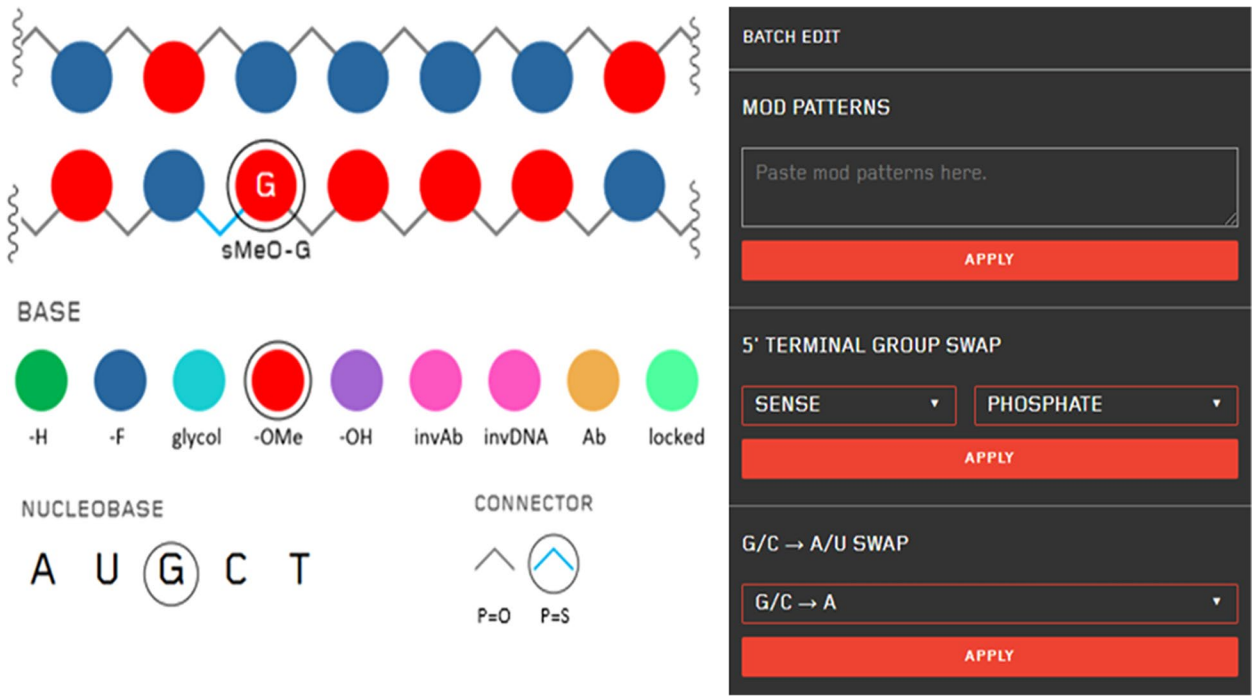

Fig. 10 Example chemistry menu selections available for users to build new siRNA constructs

of structure-property correlations in such high-density, multidimensional datasets.

To address this need, we sought to transition away from the paradigm of offline data tables into more compelling graphical interfaces for easier generation of insights (Hansen et al. 2013). Specifically, we sought an interface suitable for dynamic data visualization and developed three versions of a bespoke radial graph correlation chart customized for large molecule data. The radial graph charts enable the interactive analysis and exploration of multivariate candidate selection datasets through a fluid data visualization interface (Krzywinski et al. 2009). This allows scientists to seamlessly navigate sequence information and associated experimental measurement data in an intuitive manner.

In the first adaption of the radial graph chart, experimental measurements across a set of assays are correlated to the respective molecule candidate (Fig. 11).

Abstraction of such highly dense data permits the rapid identification of candidates with the best performance profile. Assay thresholds can be adjusted using the filtering menu, which also allows individual assays to be hidden from the view, thereby reducing the data density down even further (Fig. 11).

Data trends across related sets of molecules can be more easily spotted as well. Time spent by scientists in the drug candidate selection process is significantly reduced by moving data analysis out of large, unwieldy spreadsheets, making it easier to quickly identify the most promising drug candidates for further study.

The second version of the radial graph realigns the experimental data from individual molecule candidates to a multi-candidate sequence comparison view. A sequence alignment is performed across all candidates and the radial graph chart is redrawn to map the experimental data to the aligned sequence positions (Fig. 12). Sequence domains, variable target associations, and sub-regions for each protein chain are interactively displayed around the outer rim of the sequence alignment segment of the chart. In this way, the structural and positional context for each point mutation can be understood. Key homologies between different sequence motifs and molecular domains are readily distinguished in this view, highlighting sequence engineering features that lead to more favorable molecule attributes.

In the final adaption of the radial graph, we imagined integrating the BioMap technology described earlier with the high-density data visualization capabilities of the radial graph. Here, the focus is on molecule stability data as measured by forced degradation chemical liability experiments. The chemical liability datapoints are mapped to the corresponding protein chain and positional location on the BioMap structure. The ability to navigate between one-dimensional sequence and two-dimensional structure is made possible by fully integrating the data across both graphs (Fig. 13). This

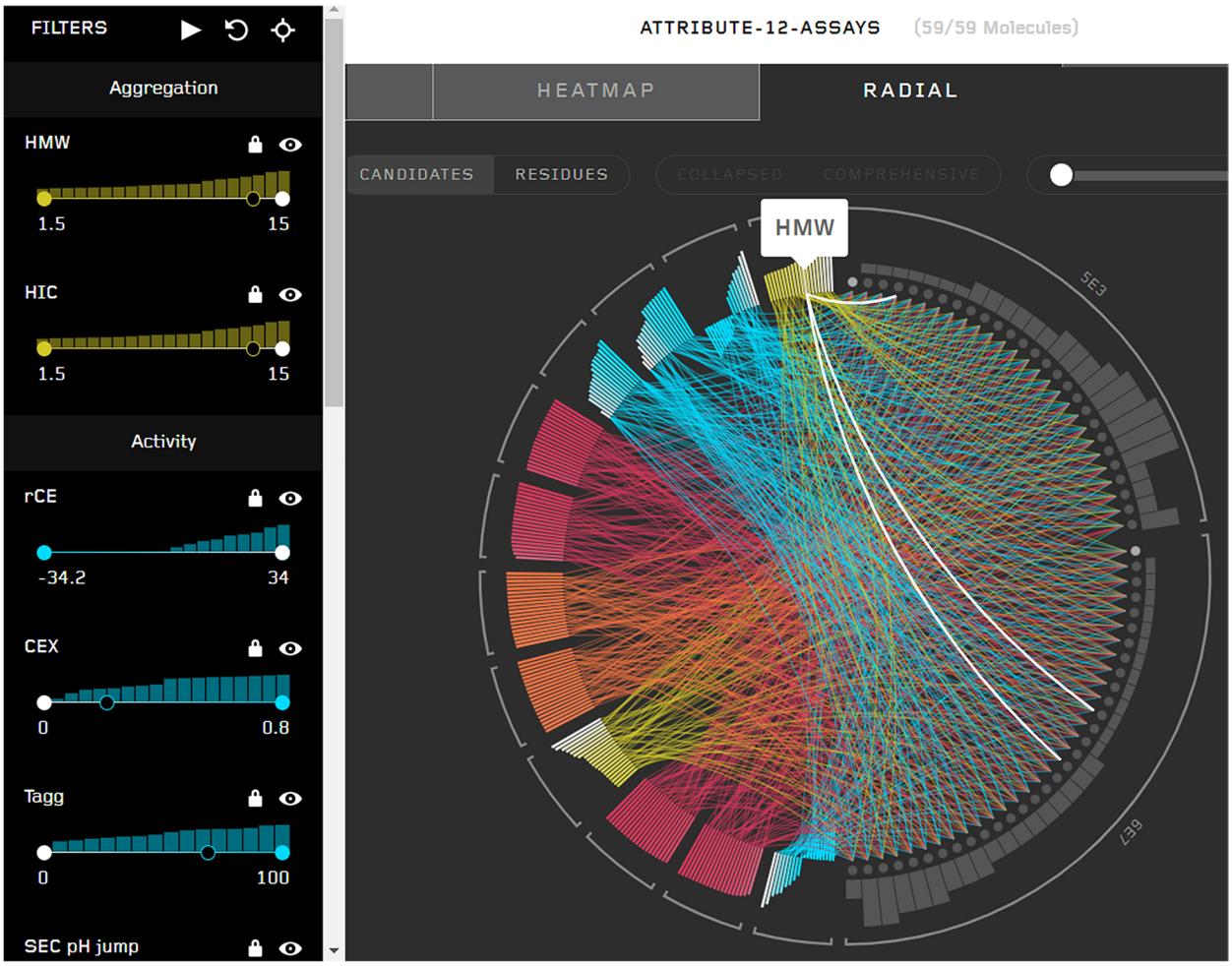

Fig. 11 Radial graph multivariate data correlation chart showing individual molecule candidate view

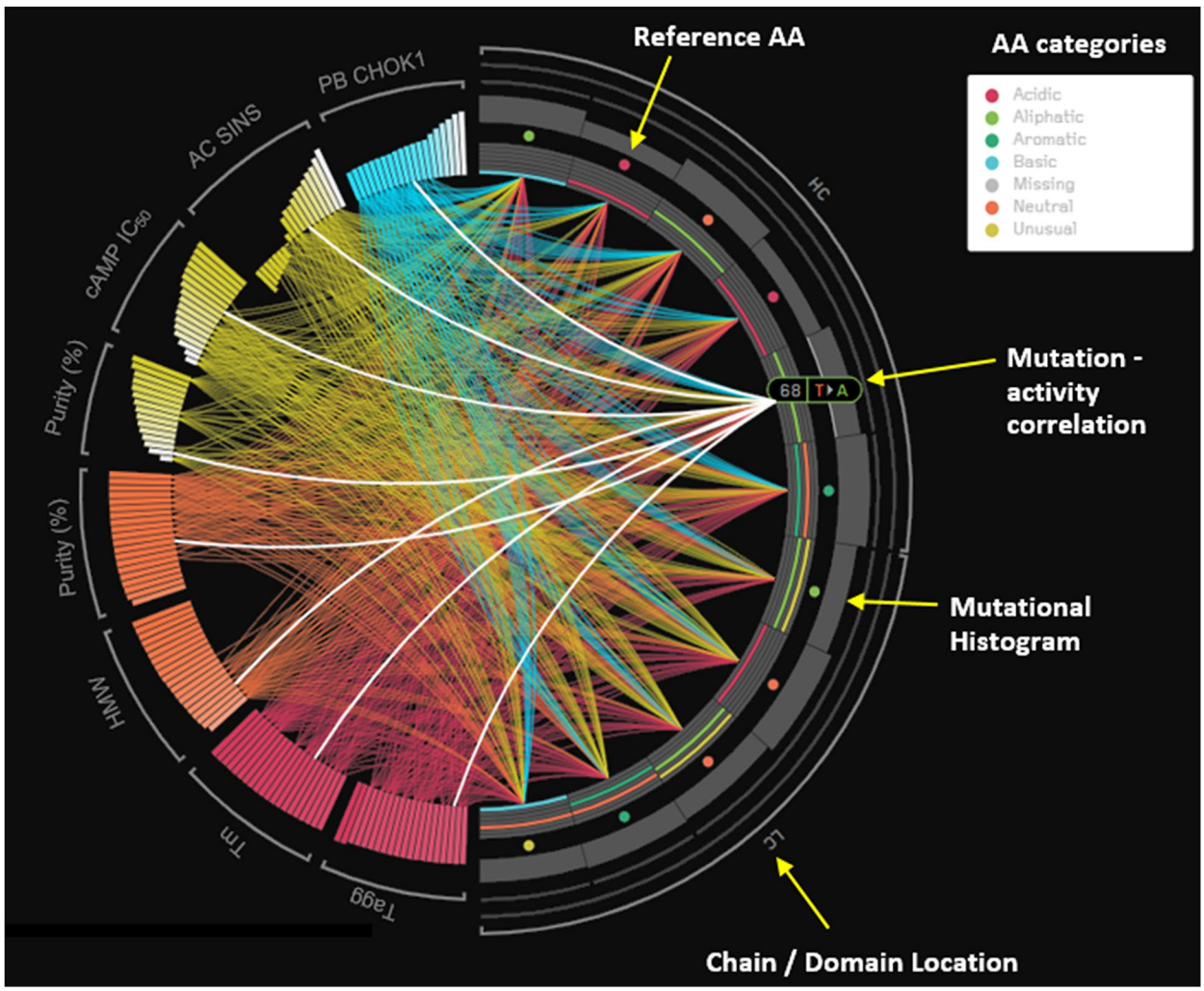

Fig. 12 Realignment of the radial graph showing multi-sequence alignment comparisons across all molecule candidates mapped to experimental measurement data

fully simultaneous interactivity offers facile navigation across thousands of measurement outputs for a range of experimental parameters. In this manner, sequence-related data can be understood within a structurally contextualized view of overall scaffold composition to inform potential product quality risks based upon stability stress conditions.

Conclusions and future directions

We have conceived and developed four new informatic methods for the visualization and analysis of engineered biologics and oligonucleotides and associated experimental data. These technologies support communications between scientists and their information consumers and represent an advance on existing techniques for the representation and analysis of biotherapeutics.

The concepts of data reduction and compression were applied to translate the structural complexities of complex biomolecular constructs into simplified, informative, and interactive visual schematics. Additionally, the innovative data visualization charts we developed enable more flexible interrogation and structurally informed interpretation of large-volume drug optimization datasets.

As a future step, we envision extending upon this work by utilizing a fully visual language to encode and compress biological supramolecular domains into a large molecule-focused syntax notation using characters, symbols, and logical operators. This concept has similarity with SMARTS, a molecular pattern language for small molecules (Daylight Inc. 4 2022). This would build upon the BioFont and BioMap technologies described herein and subsequently allow for automated

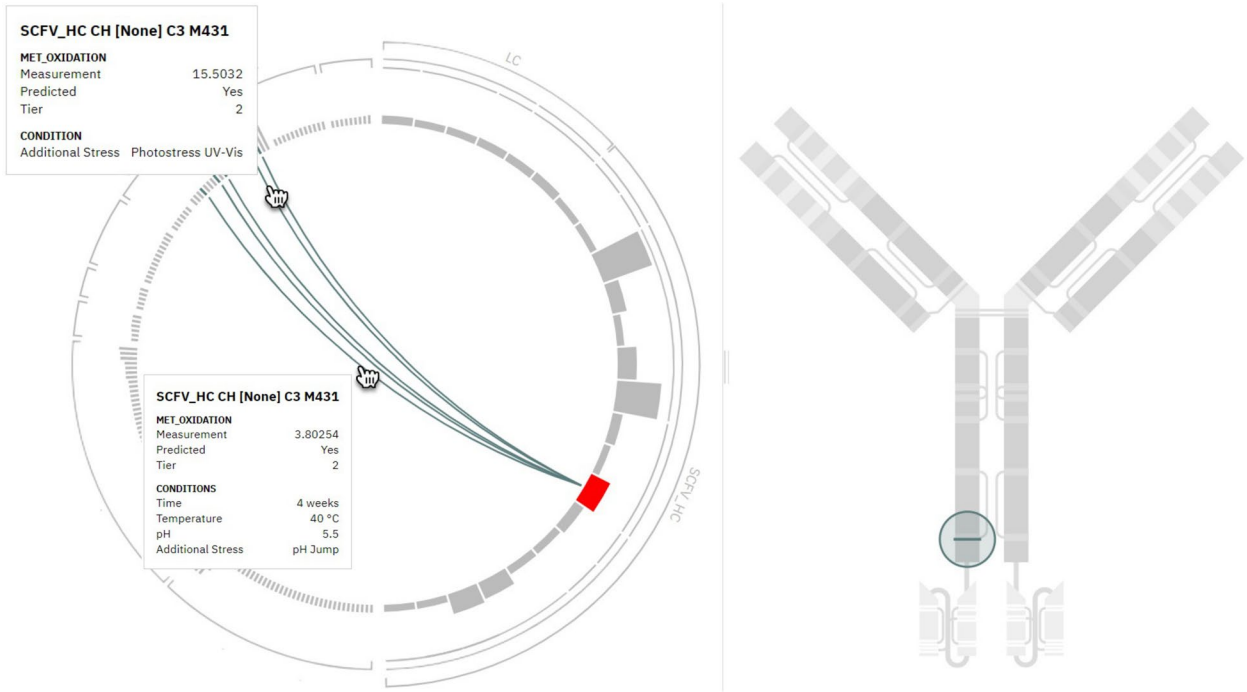

Fig. 13 Experimental property data are correlated to one-dimensional amino acid sequence positions in the radial graph and the two-dimensional location on BioMaps

conversion of such compressed 'Bio-SMARTS' notations into standardized graphical representations. Such a solution would provide a future-proof and extensible method for depicting complex large molecule constructs, which to the author's knowledge, does not exist currently in any available bioinformatics tool.

Acknowledgements
The authors thank Justin K. Murray and Bin Wu for helpful discussions around design and synthesis of chemically modified siRNA. We also thank Kenneth Walker, Michelle Hortter, and Marissa Mock for their feedback on large molecule schematic representation and optimization data. Finally, we thank members of Amgen Therapeutic Discovery, Information Systems, and Process Development for their support in building the tools described here.

Authors' contributions
RKK and LPM conceptualized and created the designs for BioMaps, RNA visual designer, all versions of the multivariate radial graph data visualization charts, and assisted throughout the development process for all tools described herein. AR devised and developed the comprehensive sequence notation, JSON files, and implemented BioMaps. Amgen partnered with Schema Design to build BioFonts, RNA visual designer, and radial graph data visualizations; CMS conceptualized BioFonts. RKK wrote the manuscript. The author(s) read and approved the final manuscript.

Funding
All work was funded by Amgen Inc.

Availability of data and materials
BioFonts and other applicable software code will be made available by depositing at the following open-source repository on Github, kunzrk/LM-Data-Viz (github.com). All experimental data from Amgen is proprietary.

Declarations

Competing interests
CMS is the Founder and a Partner at Schema Design and received funding from Amgen to develop aspects of this work, as outlined in the authors' contributions statement, and declares no other direct financial relationship with Amgen. All other authors declare they have no competing interests.

Received: 9 December 2022 Accepted: 25 January 2023
Published online: 01 March 2023

References

Biorender. https://biorender.com/. Accessed 18 Nov 2022
Brinkmann U, Kontermann RE (2017) The making of bispecific antibodies. MAbs 9(2):182–212. https://doi.org/10.1080/19420862.2016.1268307
Chen WL, Leland BA, Durant JL, Grier DL, Christie BD, Nourse JG et al (2011) Self-contained sequence representation: bridging the gap between bioinformatics and cheminformatics. J Chem Inf Model. 51(9):2186–2208. https://doi.org/10.1021/ci2001988/
David L, Thakkar A, Mercado R, Engkvist O (2020) Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminform. 2020(12):56. https://doi.org/10.1186/s13321-020-00460-5
Daylight Inc. SMARTS—a language for describing molecular patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed 18 Nov 2022
Hansen MR, Villar HO, Feyfant E (2013) Development of an Informatics Platform for Therapeutic Protein and Peptide Analytics. J Chem Inf Model. 53(10):2774–2779. https://doi.org/10.1021/ci400333x

Heller SR, McNaught AD, Pletnev I et al (2015) The IUPAC International Chemical Identifier (InChI). J Cheminform. 7:23–34. https://doi.org/10.1186/s13321-015-0068-4
International Union of Pure and Applied Chemistry: Nomenclature. https://iupac.org/what-we-do/nomenclature/. Accessed 18 Nov 2022
Khvorova A, Watts J (2017) The chemical evolution of oligonucleotide therapies of clinical utility. Nat Biotechnol. 35:238–248. https://doi.org/10.1038/nbt.3765
Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D et al (2009) Circos: An information aesthetic for comparative genomics. Genome Res. 19(9):1639–1645. https://doi.org/10.1101/gr.092759.109
Pistoia Alliance. https://www.pistoiaalliance.org/helm-notation/. Accessed 19 Jan 2023
Setten RL, Rossi JJ, Han SP (2019) The current state and future directions of RNAi-based therapeutics. Nature Rev Drug Discov. 18(6):421–446. https://doi.org/10.1038/s41573-019-0017-4
Siani MA, Weininger D, Blaney JM (1994) CHUCKLES: a method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. J Chem Inf Comput Sci. 34:588–593. https://doi.org/10.1021/ci00019a017
Siani MA, Weininger D, James CA, Blaney JM (1995) CHORTLES: a method for representing oligomeric and template-based mixtures. J Chem Inf Comput Sci 35(6):1026–1033. https://doi.org/10.1021/ci00028a012
Sweet-Jones JM, Ahmad M, Martin ACR (2022) Antibody markup language (AbML) - a notation language for antibody-based drug formats and software for creating and rendering AbML (abYdraw). MAbs 14(1):2101183–2101187. https://doi.org/10.1080/19420862.2022.2101183
The Unicode Consortium. http://www.unicode.org/press/seachange.html. Accessed 18 Nov 2022
Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 28(1):31–36. https://doi.org/10.1021/ci00057a005
Zhang T, Li H, Xi H, Stanton RV, Rotstein SH (2012) HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. J Chem Inf Model. 52(10):2796–2806. https://doi.org/10.1021/ci3001925

Publisher's Note