nature computational science

Perspective

https://doi.org/10.1038/s43588-024-00642-3

Linguistics-based formalization of the antibody language as a basis for antibody language models

Received: 29 September 2022

Accepted: 13 May 2024

Published online: 14 June 2024

Check for updates

Mai Ha Vu [1,5] ✉, Philippe A. Robert [2,5], Rahmad Akbar [2,5], Bartlomiej Swiatczak [3], Geir Kjetil Sandve [4,6], Dag Trygve Truslew Haug [1,6] & Victor Greiff [2,6] ✉

Apparent parallels between natural language and antibody sequences have led to a surge in deep language models applied to antibody sequences for predicting cognate antigen recognition. However, a linguistic formal definition of antibody language does not exist, and insight into how antibody language models capture antibody-specific binding features remains largely uninterpretable. Here we describe how a linguistic formalization of the antibody language, by characterizing its tokens and grammar, could address current challenges in antibody language model rule mining.

The use of language as a metaphor for adaptive immune receptors reaches back several decades[1] (Box 1), most prominently discussed in Niels Jerne's 1984 Nobel lecture[2]. The immune system relies on highly diverse immune receptors (antibodies and T-cell receptors) to fight previously unseen infections, whose sequences are generated by complex genetic recombination mechanisms[3]. Immune receptors recognize structures called antigens by binding to them[4]. In the case of protein antigens, residues involved in the binding interface (usually defined as residues within a distance of <5 Å of one another[5,6]) are the paratope and the epitope on the antibody and antigen, respectively. One of the longest-standing problems in immunology is the antibody-specificity prediction problem[7,8], which aims to determine which antibody sequences bind a defined antigen and vice versa[9]. The affinity of the antibody–antigen interaction determines the physiological efficacy of an antibody, and studies typically separate binders and non-binders based on a selected affinity threshold. Consequently, antibody-binding prediction is a specialized case of protein–protein interaction prediction with different domain-specific binding motifs and amino acid usage[5], lack of evolutionary information, and low conservation or similarity between antibodies of the same function[9–12]. It has recently been shown that fine-tuning protein language models (LMs) with antibody sequence information improves performance on various tasks[13,14]. At the same time, domain-specific antibody LMs performed differently to protein LMs depending on specific immunological tasks[15].

The language metaphor implies the existence of a rule-based grammar system that can link immune receptor sequence to function, thereby enabling the designing and predicting of immune receptor and cognate antigen interaction. So far, there does not exist a rigorous linguistic formalization of an immune receptor language, as the first evidence that antibody specificity is predictable was only recently found with the availability of large-scale data[5,16], implying the existence of an antibody grammar. At the same time, the exclusive use of LMs borrowed from the field of natural language processing (NLP) without any grounding in domain-specific models is inadequate for mining antibody grammar rules, as the goals of NLP are to generate plausible data rather than to discover the principles that underlie that data[17,18]; deciphering language grammar by considering rare, out-of-distribution data remains in the realm of linguistic analysis[19]. Explicit knowledge of biological sequence rules would facilitate rational antibody design by building information-efficient machine learning tools, enhancing the interpretability of existing tools and thereby guiding the functional labeling of rare sequences[20]. Here we focus our discussion on antibody sequences, but most of our points apply to the entirety of adaptive immune receptors[21].

We propose a guideline for creating linguistic formalizations of the antibody language, given its particular characteristics that make

[1]Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo, Norway. [2]Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway. [3]Department of History of Science and Scientific Archeology, University of Science and Technology of China, Hefei, China. [4]Department of Informatics, University of Oslo, Oslo, Norway. [5]These authors contributed equally: Mai Ha Vu, Philippe A. Robert, Rahmad Akbar. [6]These authors jointly supervised this work: Geir Kjetil Sandve, Dag Trygve Truslew Haug, Victor Greiff. ✉e-mail: m.h.vu@iln.uio.no; victor.greiff@medisin.uio.no

BOX 1

Early endeavors in immunolinguistics

One of the earliest works connecting linguistics and immunology can be traced back to Burnet[1], who drew linguistic analogies to theorize about the origin of immune receptor variation. He compared immune receptors to short strings of letters randomized by a computer program[1]. He speculated that mechanisms might be in place to increase the frequency of meaningful combinations of stochastically generated letter-like gene fragments during development, which has been experimentally supported later[130].

A more elaborate version of a linguistic framework was provided by Niels Jerne in his 1984 Nobel lecture[2]. Jerne drew a parallel between the open-endedness of language, which refers to language's creative capacity to express any possible semantic meaning[40,131], and the completeness hypothesis of the antibody repertoire, which proposes that every possible foreign antigen can be recognized by an antibody sequence in the organism[132,133]. By comparing immune receptor specificity to linguistic meaning and highlighting the importance of innate learning skills in language acquisition, he alluded to universal combinatorial rules of immunity akin to those postulated by Chomsky in language[40].

These early insights were further developed by linguists who saw in the immune system an example of a system that, despite being creative and productive, is also innate and conserved, reliant on inborn rules and structures[134,135]. The semantic dimension of immunity was analyzed, in turn, jointly by semioticians (specialists in the study of meaning and signs) and immunologists, who attempted to frame intercellular immune exchanges in terms of symbolic communication and meaning-making[136]. As an extension of these early efforts, Atlan and Cohen considered the process of information processing in the immune system, suggesting that immune meaning unfolds in a complex process following receptor activation[137]. Although bearing promise to expand the conceptual framework of immunology and provide novel testable hypotheses in the field, these early immunolinguistic considerations remained largely theoretical, as supporting large-scale data were unavailable at that time.

it analogous to natural language and the antibody-specific challenges for current ML-based approaches (Fig. 1a). In particular, knowing which type of tokens or rules contain information on antibody binding may inform more efficient ML architectures by creating an interpretable information bottleneck. We further argue that any linguistic formalization of biological sequences needs to be custom-built for a particular research question to obtain concrete, symbolic rules that explicitly target the given problem. We here showcase a possible implementation of our proposal on formalizing the antibody language in the context of the antibody-specificity prediction problem.

Linguistic formalization can aid with antibody-specific challenges in antibody LMs

In recent years, LMs have become the state-of-the-art machine learning tool for processing linguistic data[22,23]. In a recent Perspective[24], we argued in favor of linguistically guided protein LMs to better integrate domain-specific knowledge into LMs and to build interpretable LMs that are more likely to learn domain-specific rules[24]. Given the linguistic metaphor for the antibody sequence universe, there has been a proliferation of LMs applied to antibody sequences to predict antibody structure[13,25–34] and properties[13,14,26,28,35–38]. Current antibody LMs[25–32] typically do not use domain-informed tokenization, focusing on prediction accuracy rather than interpretability and explainability. These models, due to their black-box nature, cannot yet directly aid in the discovery of a verifiable rule-based antibody grammar (1 in Fig. 2), or even serve as reliable tools for antibody-specificity prediction[20]. As such, they do not fulfill the premise of the antibody language metaphor, which promises the ability to decipher the antibody language grammar. Furthermore, a challenge to ML models is the accurate prediction of out-of-distribution data[39], whereas linguistic grammars aim to model all possible sequences regardless of statistical frequency[40].

We argue that a purpose-driven, linguistic formalization of the antibody language is a prerequisite to building interpretable antibody LMs that address antibody-specific challenges in machine learning. Without a linguistic formalization, it would be largely impossible to determine whether the statistical patterns extracted from an LM correspond to actual scientific principles or are merely spurious correlations[41,42]. Linguistics has aimed to provide symbolic, rule-based formalizations of natural language[40], in contrast to deep LMs, which are based on many-parameter statistical modeling of linguistics data[22,43–45]. A full linguistic formalization of natural languages consists of defining the basic units (tokens) of the language, syntactic rules that govern how these tokens combine into larger linguistic structures and semantic rules that map linguistic structures to meaning. These formalizations can provide grounding for natural language LMs, for example, by defining linguistically valid tokenization of the input[46,47]. Without a linguistically guided implementation, it is not guaranteed that LMs have learned scientifically valid rules[24]. For example, LMs for English can perform many tasks with high accuracy even when their input is not segmented into linguistically valid tokens[22,46,48]. Therefore, LMs without grounding in symbolic formalization, even when performing with high accuracy, are not guaranteed to operate on scientific principles and thus are largely unsuitable tools for discovering scientifically valid rules.

For antibody sequences, a full linguistic formalization remains elusive. Nevertheless, incorporating already known rules of antibody binding, such as specific motifs that play an important role in antigen recognition and defining the nature of the antibody language grammar and tokens at a higher level can further advance the quest of extracting a full antibody grammar from more interpretable antibody LMs. For example, establishing well-defined properties of antibody tokens could help limit the search space for valid sequence or structural motifs that could play a biological role in binding. More generally, a conceptual description of the antibody language can guide the interpretation of the probabilistic patterns extracted from antibody LMs.

A further challenge for applying machine learning tools to antibody sequences is the limited amount of available sequence data (2 in Fig. 2): there are only of the order of 10^9 publicly available immune receptor sequences[49,50] compared with more than 10^{14} biologically possible sequences[16,51]. Antibody–antigen structural binding data are even scarcer at $\sim 10^4$ binding pairs available[52]. Antibody LMs with built-in biological knowledge drawn from a linguistic formalization can alleviate this challenge (3 in Fig. 2), as natural language LMs of low-resource natural languages benefit from injecting linguistic information of the data first, such as morphological knowledge[47,53,54] or cross-linguistic data[55–57]. Both theoretical[5,10,58–60] and experimental (for example, cryogenic electron microscopy [61], deep mutational scanning[10], single-cell sequencing[62]) domain knowledge about antibody sequences is fast emerging, which creates a particularly golden opportunity now for implementing our proposal.

Finally, antibody-specificity prediction remains particularly challenging for machine learning methodologies (3 and 4 in Fig. 2), due to the unique properties of antibody sequences as opposed to non-immune protein sequences. Sequence similarity does not necessarily correlate with binding similarity: sequences with close edit

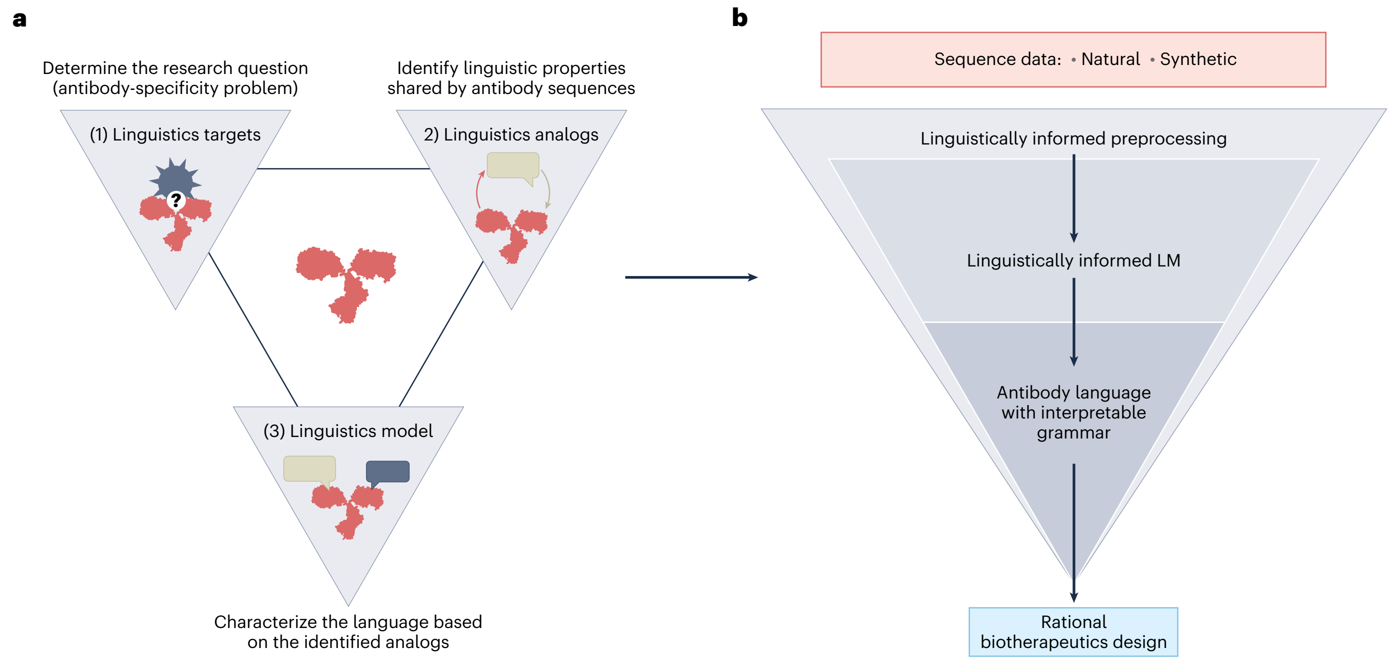
a
Determine the research question
(antibody-specificity problem)
(1) Linguistics targets
?
Identify linguistic properties
shared by antibody sequences
2) Linguistics analogs
(3) Linguistics model
Characterize the language based
on the identified analogs
b
Sequence data: • Natural • Synthetic
Linguistically informed preprocessing
Linguistically informed LM
Antibody language
with interpretable
grammar
Rational
biotherapeutics design

Fig. 1 | Integration of a linguistic formalization of antibody sequences into antibody LMs. a, Linguistic formalization of antibody sequences. We propose that a linguistic formalization requires the following steps: (1) determine the research question that the linguistic formalization aims to solve, which in our case is antibody-specificity prediction, (2) identify natural language properties shared by antibody sequences, which we illustrate in Fig. 3, and (3) characterize the language based on the identified analogies, which we illustrate in Fig. 4. **b**, Linguistically grounded antibody LM. The obtained linguistic formalization informs the input sequence data preprocessing and the design of the language model itself. Thus integrating a linguistic formalization of antibody sequences into current machine learning techniques can lead to better interpretability while maintaining the ability to statistically process large, unstructured data, resulting in a linguistically grounded deep antibody LMs. Easily interpretable LMs can aid with deciphering the rules of antibody specificity, which is crucial for rational and in silico antibody biotherapeutics design.

distance might bind different antigens, while dissimilar sequences can bind the same antigens[9,10,20,63] (3 in Fig. 2). Furthermore, antibody sequences show cross-reactivity: antibody–antigen binding is a many-to-many mapping where the same immune receptor may recognize multiple different antigens, and an antigen can be recognized by multiple antibody sequences[9–12], creating a complex many-to-many recognition network between antibodies and antigens (4 in Fig. 2). Our proposed linguistic formalization can provide a unique perspective on these particular challenges of antibody-specificity prediction, with practical implications for building an antibody LM (Fig. 2).

Formalizing the antibody language for antibody-specificity prediction

Here we provide a possible high-level formalization of the antibody language. Because linguistic formalizations of natural languages are well established, it is helpful first to pinpoint shared properties between natural language and antibody sequences to define a linguistic formalization for biological sequences (Fig. 3). We then build on these identified properties to provide a formalization targeted to antibody-specificity prediction (Fig. 4). This formalization can be a starting point for better antibody LM design (Fig. 1).

Shared properties between natural language and antibody sequences

Discreteness

Discreteness holds that linguistic sequences are built from a finite number of smaller units into a countably infinite number of possible combinations with the help of a finite set of rules[64]: sounds are combined into lexical items, lexical items are combined into phrases and phrases are built into sentences.

At every step, language-specific rules determine an infinite number of possible combinations due to the possibility of recursive rules[64]. Similarly, antibody sequences consist of amino acids at the lowest level, which combine into larger meaningful subunits similar to words (Fig. 3a). The existence of these antibody subunits is evident in the fact that encoding sequences as such subunits improved antibody-specificity prediction[5,9,65,66] and de novo antibody sequences could be designed by combining subunits[67,68].

Hierarchical structure

Language is organized in hierarchical structures determined by syntactic rules and commonly represented as trees[69]. Similarly, antibodies have a three-dimensional (3D) structure beyond their linear sequence, similar to proteins (Fig. 3b). Protein sequences fold into a primary and secondary structure with local structural patterns defined by the local amino acid sequence, followed by the 3D conformation of longer-distance domains (tertiary and quaternary structure)[70,71]. The order of local folding events can, for instance, be represented in a lattice structure, as a tree of folding events[72–75]. Alternatively, it is also possible to encode trees and higher-dimensional structures as strings[9,76–79] by incorporating angles into the string encoding.

The antibody 3D structure is crucial for determining antibody-binding behavior[80–83]. The folding creates paratopes on the antibody[80–83], and antibody sequences might bind different antigens determined by their folded structure[81–83]. Of note, the same antibody sequence may bind different targets using a different set of paratope residues[81–83]. Consequently, non-paratope parts of antibodies also contribute to the specific binding by making the appropriate 3D structure possible[80].

Ambiguity

Linguistic sequences can be ambiguous, as they can map to multiple meanings. Linguistic ambiguity can be due to tokenization ambiguity, where a sequence can be subdivided into different sets of tokens[84], structural ambiguity, where different structures contribute to

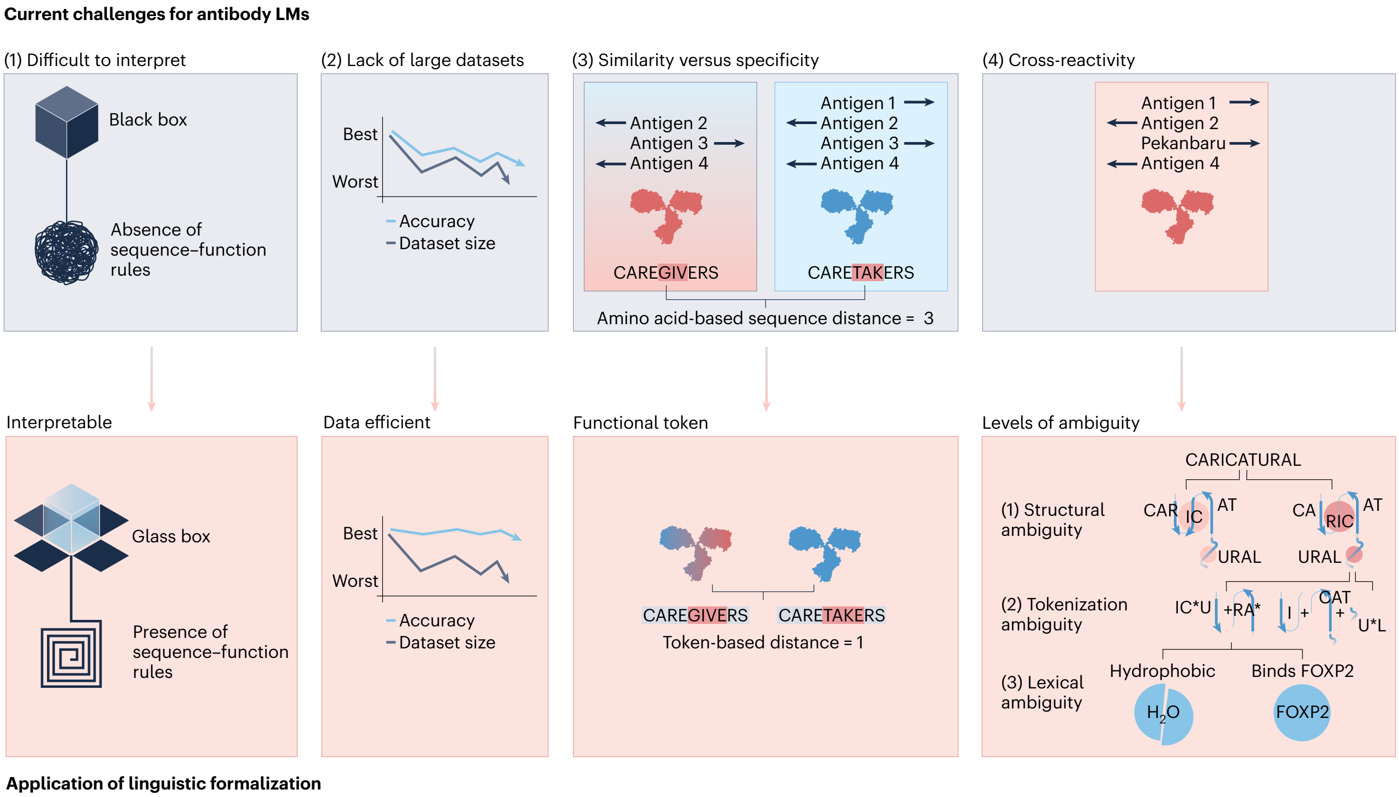

Fig. 2 | Linguistic formalization as a solution to current challenges in antibody-specificity prediction with LMs. Current antibody LMs are typically black-box models trained on inputs that are tokenized on an amino acid basis that remain difficult to interpret and face challenges for antibody-specificity prediction. A linguistic formalization can address challenges of antibody LMs. (1) Difficult to interpret: the formalization can guide the interpretation of these models, leading to the extraction of scientifically valid sequence–function rules. (2) Lack of large datasets: the lack of antibody data, which in general hinders machine learning success, can be alleviated by integrating domain-specific knowledge into antibody LMs, as similar approaches have proved to be successful for low-resource languages in NLP[47,53–57]. (3) Similarity versus specificity: amino acid-based sequence similarity lacks correlation with antibody specificity[9,10,20,63]. A linguistic formalization that provides functional tokens would allow for calculating token-based sequence similarity instead, which could lead to a better correlation between sequence similarity and specificity. (4) Cross-reactivity: a linguistic formalization recasts cross-reactivity as different types of linguistic ambiguity, which can help with a more precise understanding of antibody grammar. For example, the sequence 'CARICATURAL' can take up different structures (for example, it folds between the segments CAR+IC+AT+URAL versus CA+RIC+AT+URAL) (structural ambiguity), the same structure can be segmented in different ways, leading to different tokenizations (for example, IC*U + RA* versus I + CAT + U*L, where * stands for wild amino acids) (tokenization ambiguity), and the same token might match with different meanings (for example, it is hydrophobic versus binds the FOXP2 gene) (lexical ambiguity).

different meanings[85], or lexical ambiguity, where a token itself might have several different meanings. Antibodies similarly show ambiguity in their binding behavior as they are cross-reactive[7,11,81,86] (Fig. 3c). As with language, antibody sequence ambiguity could be due to structural ambiguity, where each different antibody fold associated with the same sequence binds different targets[87]. However, often the same antibody structure can also recognize several different antigens[11,81,86], where tokenization or lexical ambiguity might become relevant: either the same antibody sequence could be re-analyzed as containing different tokens (tokenization ambiguity) or the same token has multiple associated functions (lexical ambiguity) (4 in Fig. 2).

Compositionality of meaning

Every linguistic sequence maps to meaning, which is grounded outside of language. The meaning of sentences is compositional, as it can be derived from the meanings associated with the individual lexical items that built it and from the order they combined. Compositionality is key to semantic rules that link a linguistic sequence to its meaning: by combining the meanings of individual tokens in the sentence in the same order they build the linguistic structure, we can derive the meaning of the full sequence (Fig. 3d).

Antibody binding is governed by complex physicochemical laws, which in principle, makes antibody meaning non-arbitrary. Moreover, there is evidence that antibody specificity can also be derived from the compositional combination of its subunits, as antibody specificity is predictable from sequence alone[10,88,89] and binding motifs are shared across different antibody–antigen-binding complexes[5,66]. These works are complemented by first successes in paratope–epitope prediction[5,90–92]. As antigen specificity is mainly determined by ~5–20 amino acids in the complementarity-determining region 3 (CDR3), the recognition and classification of billions of different antigens is based on a feature space of merely ~5–20 dimensions. The low dimensionality implies the presence of strong high-order dependencies between amino acids. Consequently, antigen specificity may result from the conditional combination of components in the form of paratope subsequences (contiguous or gapped *k*-mers)[93], similar to natural language, where semantics arise from a combination of words according to a given grammar. Indeed, the study by Akbar et al. showed that a systematic subdivision of antibodies into interacting and non-interacting paratope motifs improved antigen-binding prediction[5], and, importantly, these motifs were shared across entirely different antibody–antigen-binding complexes, suggesting the existence of a generalizable, systematic antibody grammar. If antibody meaning can be derived compositionally, then identifying these compositional rules is the key to solving the antibody-specificity prediction problem.

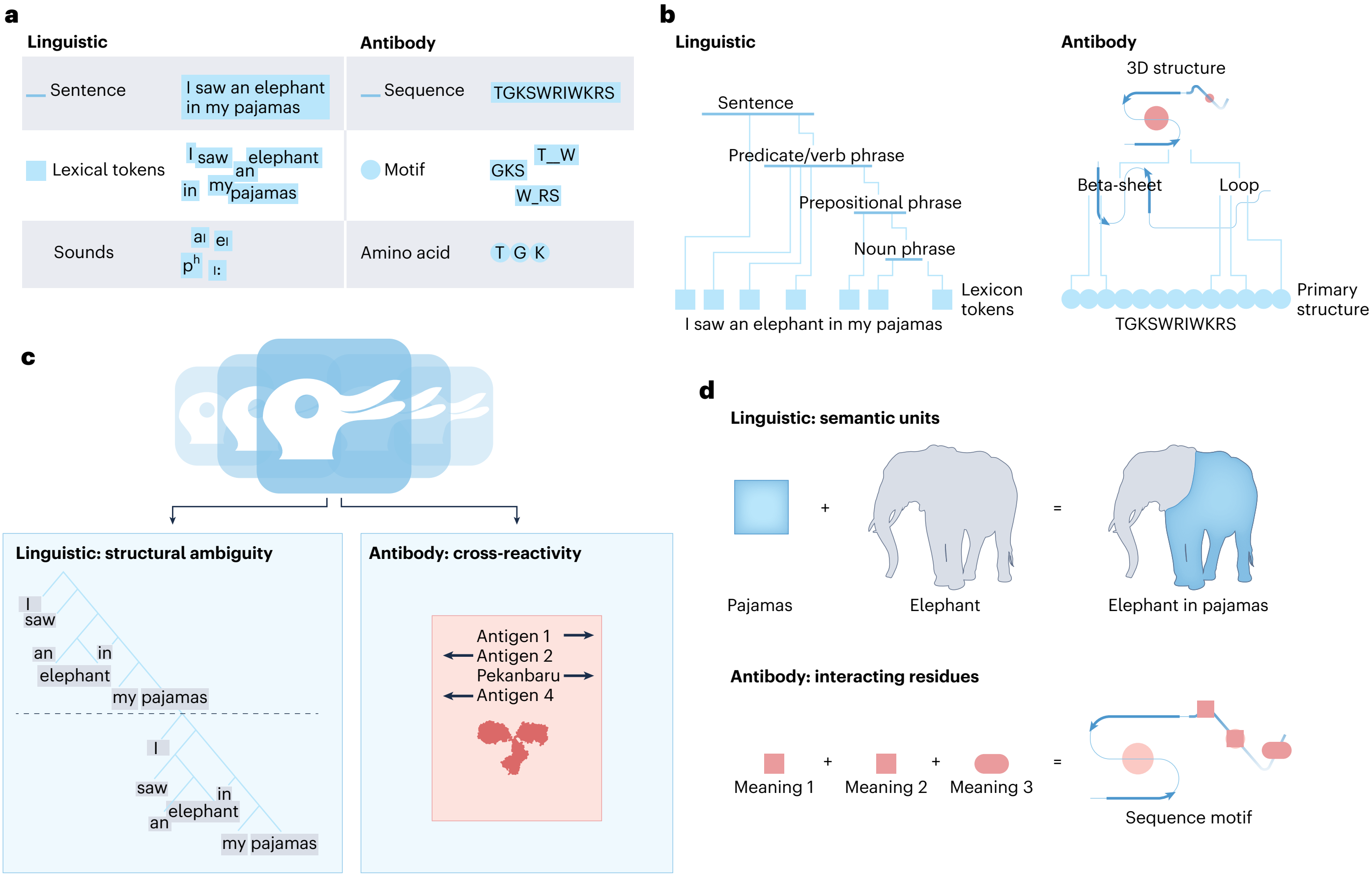

Fig. 3 | Shared properties between linguistic and antibody sequences. **a**, Discreteness: sentences are built from finite building blocks (sounds) via intermediary units (words, lexical items, tokens) that possess semantic meaning. Antibodies similarly are built from a finite set of amino acids and probably also possess intermediary units (motifs). **b**, Structure: linguistic strings possess a combinatorial building structure that can be represented as trees. Similarly, antibodies have a 3D structure, which results from the folding of subsequences into primary structure blocks that can form only a finite set of 3D structures according to physical constraints. **c**, Ambiguity: ambiguity, defined as the same sequence mapping to two distinct meanings (for example, the picture on the left can be either a rabbit or a duck), is found in both linguistic and antibody sequences. Natural language sentences can have multiple meanings, and antibody sequences can bind multiple targets (cross-reactivity). **d**, Compositionality: the meaning of a sentence is composed by combining the meaning of its lexical items that build it through the order they are combined in the sentence. Evidence suggests that antibody meaning can be similarly composed of individual motifs with associated meaning[10,88,89].

Defining the linguistic model

We recast antibody-specificity prediction as a formalized antibody language. Specifically, we define notions of well-formedness and meaning by characterizing antibody grammar and tokens in the context of antibody-specificity prediction.

Characterizing antibody grammar

For a natural language, a sequence or structure is well formed if it adheres to the language's syntactic rules, while compositional semantic rules map each syntactically well-formed sequence to meaning[94]. We propose that well-formed antibody sequences are equivalent to all observable sequences formed by V(D)J recombination that are in-frame and without stop codons. We distinguish between sequence-building syntactic rules (for example, V(D)J recombination[95]), which determine well-formed sequences on the one hand, and structure-building syntactic rules, which determine antibody folding on the other hand (Fig. 4). The meaning of a well-formed antibody sequence is the set of epitopes it binds (Fig. 4). Semantic rules map a folded antibody structure to its recognized epitopes and binding affinity for those epitopes.

If the input to antibody-specificity prediction is a well-formed antibody sequence, deciphering both the structural syntactic rules and the compositional semantic rules is necessary to predict antibody specificity. If the input is a well-formed antibody structure, then learning only the semantic rules is needed, and thus syntactic and semantic rules could be better teased apart.

Characterizing antibody tokens

Both syntactic and semantic rules operate over sets of discrete tokens. Semantic tokens must be items with functional meaning so that they can be combined into full sequence meaning, whereas syntactic tokens do not need to have functional meaning; they only need to facilitate the emergence of syntactic rules (Fig. 4). While in natural language, syntactic and semantic rules use the same set of tokens, we propose that the antibody language has separate syntactic and semantic tokens.

Antibody semantic tokens are discrete units that correspond to structural motifs with an identifiable functional meaning (Fig. 4 illustrates a few hypothetical examples, where motifs possibly correspond to physicochemical properties, shape and content of corresponding antigen epitope). Because the antibody semantic rules map from structure to meaning, the semantic tokens must already have structure (Fig. 4). As with linguistic lexical items, these motifs could be lexically ambiguous with multiple different meanings, and multiple motifs could be synonymous by mapping to the same meaning as well.

To identify semantic tokens, there needs to be an exhaustive list of relevant lexical meanings to antibody-specificity prediction and an analytical mapping between the structural motifs and corresponding lexical meaning. Compositional semantic rules map the combinations of functional motifs to a combined meaning, resulting in the features of a fully recognized epitope (Fig. 4). As the 'meaning' of antibody motifs is, to a large extent, affected by long-distance dependencies[5,10,96], the appropriate semantic rules for the antibody language must enable the combination of motifs that are distant from each other in the sequence. LMs can serve as a particularly powerful tool for discovering such rules, as they can find statistically significant long-distance dependencies[97,98] and calculate word embeddings[48] based on context in natural language.

To integrate binding affinity into the model, each semantically meaningful token would map to a probabilistic distribution of different meanings. Compositional semantic rules then would be functions that combine these semantic tokens to yield a probability that the full antibody binds a certain epitope. This probability would be analogous to binding affinity. Similar work exists already for capturing probabilistic meaning in natural language semantics[99], in particular, the use of probabilistic approaches to formalize lexical items[100] and calculate their combinatorial meaning in sentences[101–103].

In contrast, syntactic tokens do not need to be meaningful. For example, the rules of V(D)J recombination operate on the level of nucleotides, which do not have identifiable functional meaning that pertains to antibody specificity (Fig. 4). Syntactic tokens are relevant for antibody-specificity prediction if the structure needs to be predicted from sequence; if the structure is already given, then syntactic tokens become irrelevant.

Outlook

Our linguistic formalization of the antibody language provides practical implications for building interpretable antibody LM design that would be useful for scientific discovery in antibody-specificity prediction.

First, the linguistic perspective reveals that semantic rules are the key to antibody-specificity prediction. In practical terms, antibody LMs should be dedicated to learning only semantic rules for increased interpretability. To do so, the input data to LMs should already be encoded with structural information so that the LM does not need to learn syntactic rules, and the tokens should correspond to meaningful semantic tokens. There exist multiple verifiable design solutions for processing multi-dimensional structures with LMs. For one, structural information can be encoded into a string, either by using a lattice representation of the structure[9], or by incorporating more fine angles[77–79]. Alternatively, the LM architecture itself could be built so that it directly parses higher-dimensional structures, as has been done for linguistic trees[104,105], although this can become computationally costly for 3D representations. Existing studies suggest that encoding the input data with structure improves antibody-specificity prediction[9,106,107].

Furthermore, finding the correct semantic tokens addresses the lack of correlation between sequence similarity and specificity. Instead of calculating sequence similarity based on amino acid edit distance[9,10,20,63], we hold that semantic token-based similarity may be a more biologically founded predictor of antibody specificity (3 in Fig. 2).

Finally, cross-reactivity can be analyzed as various types of linguistic ambiguity, where the source of the ambiguity can be structural, tokenization or lexical (4 in Fig. 2). By understanding different sources of cross-reactivity through the lens of linguistic ambiguity, it becomes possible to finely control the type of ambiguity learned by the antibody LM by manipulating its input encoding. In the case of structured and semantically tokenized input, the only type of ambiguity would be lexical. For structured, untokenized input, both tokenization and lexical ambiguity are possible. Finally, for sequential inputs, structural ambiguities can also arise. Thus by choosing the input type, it becomes possible to interpret any extracted pattern as different parts of the antibody grammar, thereby making it possible to further decipher the full symbolic grammar of antibody-specificity prediction.

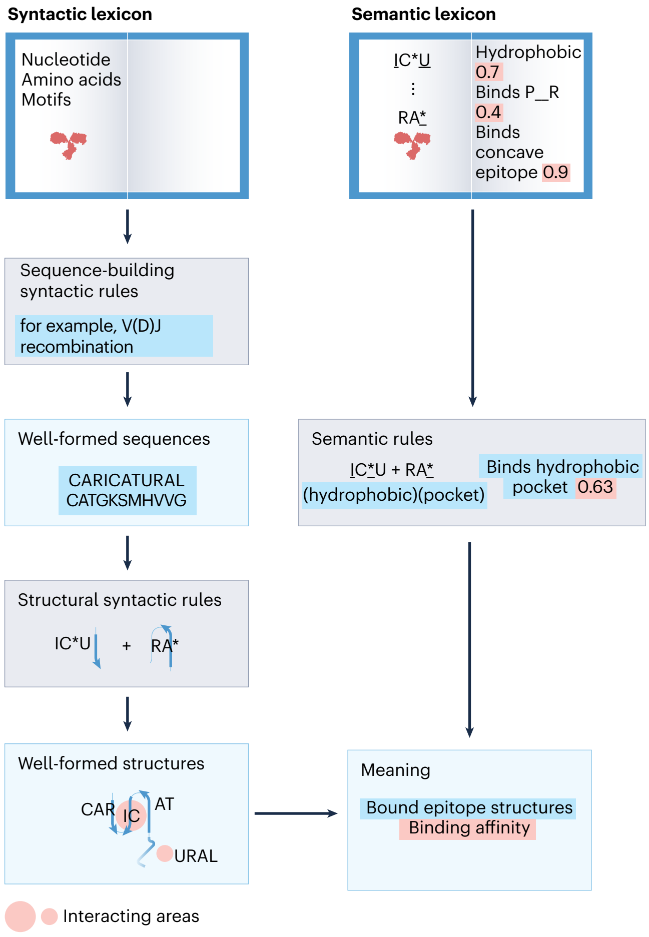

Fig. 4 | The formalization of the antibody language. In the antibody language sequence-building syntactic rules (for example, V(D)J recombination) build well-formed sequences from the syntactic lexicon (for example, nucleotides, amino acids, motifs), and structural syntactic rules map well-formed sequences to well-formed structures. Semantic compositional rules draw from a separate semantic lexicon (where each token, such as IC*U and RA*, has different meanings with certain probabilities, such as hydrophobic, binds sequence pattern P_R, or binds concave epitopes) to map folded antibody structures to meaning, that is bound epitopes. To capture binding affinity, the semantic lexicon links already folded tokens to their possible meanings together with a probability score for that meaning. The semantic rules then calculate a combined probability for the combined meaning of these tokens, yielding the binding affinity for a given well-formed antibody structure.

Of note, here we have defined the antibody language as the set of all well-formed antibodies whose meanings are the set of epitopes they recognize combined with the affinity for that recognition. There are other possible ways to define the language of antibodies. For instance, antibody sequences specific to the same antigen could be defined as their own language, where their meaning equals the different epitopes on that antigen. There could also be a separate antigen language, where each potential epitope is a string in that language. Antibodies with a certain range of affinity or developability parameters could also define separate languages. Still, our general formalizing guidelines remain relevant for all these alternative possibilities for defining the antibody language.

Current antibody LMs that receive only sequential input have to learn both syntactic and semantic rules to predict antibody specificity while for the most part using amino acid-based tokenization, which makes the statistical patterns extracted from them more challenging to interpret. In theory, information-efficient black-box ML designs with meaningful tokens and constraints reminiscent of syntactic rules may well reach as high accuracy as interpretable linguistic models. However, we believe that interpretability would be more successful in detecting or generating rare events (for instance, rare sequences following multiple rules). We thus call for prioritizing the search for meaningful structural tokens in antibody sequences that can be used for segmenting and encoding the input to antibody LMs.

At the same time, we emphasize that a purely analytical linguistic approach alone will not be adequate for finding new antibody-binding rules, given the vast space of still unknown antibody data. We believe that a linguistic formalization should be applied as grounding for statistical machine learning models and that both approaches are needed together to balance out each others' drawbacks and to achieve our goals of deciphering antibody grammar. Below, we further outline current bottlenecks in fully realizing the linguistic approach in terms of data, tokens, expressivity and framework.

Data

Structural data as input is necessary for delineating the semantic rules of the antibody language without the syntactic rules. On the one hand, we are accumulating large immune receptor sequence datasets at scale; on the other hand, experimentally solved structures to match the sequences remain prohibitively expensive to obtain[108]. Recent leaps in structure prediction, both multiple sequence alignment (MSA) based and MSA free, have begun to bridge the sequence–structure mismatch[108–111]. In particular, Fang et al.[112] have recently shown that an MSA-free structure prediction model dramatically reduced the time to model molecules (up to four orders of magnitude lower median time reduction) with minimal compromise on prediction accuracy across different protein classes (peptide, antibody and nanobody). Efficient structure prediction computation may now be combined with high-throughput sequencing experiments and information-driven docking to design scalable sequence–structure matched data on the variant level[113]. In the meantime, we advocate for supplementing the data synthetically, and for using simulation studies to evaluate the effectiveness of this approach[9,114].

Plastic tokens

The conformational flexibility of CDRs is a hallmark of antibody specificity[115]. Our understanding of an antibody molecule as a static entity struggles to attribute the breadth of its recognition capacity; thus, the view that an antibody molecule exists as an ensemble of different conformations has now taken precedence[87]. This crucial bit of information necessitates tokens that take into account the flexibility space (plastic token). The plasticity of the tokens could mimic the lexical ambiguity of linguistic lexical items and could attribute the (poly)specificity of antibodies. However, incorporating flexibility into sequence–structure data comes with a substantial computational tax. Well-accepted molecular simulation technology such as molecular dynamics simulations, is notoriously computationally costly and the appropriate timescale to sufficiently incorporate CDR loop flexibility can be lengthy[87]. Scalable solutions, either through abstraction (accelerated molecular simulation) or machine learning[116–121], will be needed for the incorporation of flexibility and the discovery of plastic antibody-specific tokens.

Expressivity

A key aspect of linguistic formalization is to describe the precise expressivity of the rules that are expected to govern antibody-binding behavior. It is important that the model should not be either overly powerful or not powerful enough. If the model is too weak in expressivity, it might be limiting for discovering more complex antibody rules. At the same time, a model with unlimited expressivity is not always desirable because it might result in rules that generate illegal sequences beyond those seen in a dataset, leading to overfitting the data and low precision[122,123]. Knowing precisely the expressivity of these rules would help narrow down the search space to only plausible antibody rules and avoid overgeneration. It would also help with choosing the most appropriate LM architecture for statistically finding these rules, as different architectures have different levels of expressivity[124,125].

Framework

Current LMs learn the statistical distributions in the datasets and infer, admittedly, complex short and long-range interrelationships. With the aid of random noise (as in denoising diffusion models), the models have expanded on the original boundaries of the training datasets[126]. Arguably the addition of noise can be seen as incorporating evolution as a separate component to the model (not to be confused with token plasticity). However, the incorporation of random noise expands the boundaries uniformly[127], whereas the rational design of antibodies may demand the ability to conditionally (non-uniformly) expand the boundaries, which in turn resembles a directed evolution. Similar to (human) languages that continue to evolve[128,129], the underlying principles for the evolution of antibody language are still to be discovered.

A new architectural paradigm, which accommodates sequence, structure, plastic tokens and linguistic evolution, such as the one outlined in Fig. 1, might become the centerpiece for the rational design of therapeutic antibodies. We foresee that the confluence of different branches of science, as seen here (linguistic and immunology), will continue to be favored over an isolated single-domain approach. While one branch, or the other, lacks the precise definition of the problem and the tools to observe beyond each respective horizon, together they might take further steps than separately.

In conclusion, a linguistic formalization, which rigorously defines biological sequences in terms of a natural language system, provides more explicit guidance on how specific LM design choices can affect the types of rules latently learned by the LM. Equipped with this finer understanding, it becomes possible to better interpret extracted statistical patterns from LMs. Although we have here shown only a formalization of the antibody language, similar formalizations could prove invaluable for other interpretable biological sequence modeling and point to new insights into existing biological questions.

References

1. Burnet, M. *Auto-Immunity and Auto-Immune Disease* (Springer, 1972); https://doi.org/10.1007/978-94-011-8095-5
2. Jerne, N. K. The generative grammar of the immune system. *Science* **229**, 1057–1059 (1985).
3. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).
4. Landsteiner, K. *The Specificity of Serological Reactions* (Harvard Univ. Press, 1945).
5. Akbar, R. et al. A compact vocabulary of paratope–epitope interactions enables predictability of antibody–antigen binding. *Cell Rep.* **34**, 108856 (2021).
6. Guest, J. D. et al. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure* **29**, 606–621.e5 (2021).
7. Rappazzo, C. G. et al. Defining and studying B cell receptor and TCR interactions. *J. Immunol.* **211**, 311–322 (2023).
8. Talmage, D. W. Immunological specificity. *Science* **129**, 1643–1648 (1959).
9. Robert, P. A. et al. Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. *Nat. Comput. Sci.* **2**, 845–865 (2022).
10. Mason, D. M. et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-021-00699-9 (2021).

11. Robert, P. A., Marschall, A. L. & Meyer-Hermann, M. Induction of broadly neutralizing antibodies in germinal centre simulations. *Curr. Opin. Biotechnol.* **51**, 137–145 (2018).
12. Greiff, V., Yaari, G. & Cowell, L. G. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr. Opin. Syst. Biol.* **24**, 109–119 (2020).
13. Burbach, S. M. & Briney, B. Improving antibody language models with native pairing. Preprint at https://arxiv.org/abs/2308.14300 (2023).
14. Singh, R. et al. Learning the language of antibody hypervariability. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.26.538476 (2023).
15. Deutchmann, N. et al. Do domain-specific protein language models outperform general models on immunology-related tasks? *ImmunoInformatics* **14**, 100036 (2024).
16. Greiff, V. et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep.* **19**, 1467–1478 (2017).
17. Min, B. et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput. Surv.* **56**, 1–40 (2023).
18. Li, J., Tang, T., Zhao, W. X., Nie, J.-Y. & Wen, J.-R. Pre-trained language models for text generation: a survey. *ACM Comput. Surv.* https://doi.org/10.1145/3649449 (2024).
19. Linzen, T. What can linguistics and deep learning contribute to each other? Response to pater. *Language* **95**, e99–e108 (2019).
20. Akbar, R. et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *mAbs* **14**, 2008790 (2022).
21. Mhanna, V. et al. Adaptive immune receptor repertoire analysis. *Nat. Rev. Methods Primer* **4**, 6 (2024).
22. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019); https://doi.org/10.18653/v1/N19-1423
23. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
24. Vu, M. H. et al. Linguistically inspired roadmap for building biologically reliable protein language models. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-023-00637-1 (2023).
25. Leem, J., Mitchell, L. S., Farmery, J. H. R., Barton, J. & Galson, J. D. Deciphering the language of antibodies using self-supervised learning. *Patterns* **3**, 100513 (2022).
26. Olsen, T. H., Moal, I. H. & Deane, C. M. AbLang: an antibody language model for completing antibody sequences. *Bioinform. Adv.* **2**, vbac046 (2022).
27. Ruffolo, J. A., Gray, J. J. & Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning. *Machine Learning for Structural Biology Workshop* (NeurIPS, 2021).
28. Shuai, R. W., Ruffolo, J. A. & Gray, J. J. IgLM: infilling language modeling for antibody sequence design. *Cell Syst.* **14**, 979–989. e4 (2023).
29. Ruffolo, J. A., Sulam, J. & Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns* **3**, 100406 (2022).
30. Prihoda, D. et al. BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs* **14**, 2020203 (2022).
31. Ostrovsky-Berman, M., Frankel, B., Polak, P. & Yaari, G. Immune2vec: embedding B/T cell receptor sequences in RN using natural language processing. *Front. Immunol.* **12**, 680687 (2021).
32. Chandra, A., Tünnermann, L., Löfstedt, T. & Gratz, R. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife* **12**, e82819 (2023).
33. Barton, J., Gaspariunas, A., Galson, J. D. & Leem, J. Building representation learning models for antibody comprehension. *Cold Spring Harb. Perspect. Biol.* **16**, a041462 (2024).
34. Dounas, A., Cotet, T.-S. & Yermanos, A. Learning immune receptor representations with protein language models. Preprint at https://arxiv.org/abs/2402.03823 (2024).
35. Hie, B. L. et al. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**, 275–283 (2024).
36. Zhao, Y. et al. SC-AIR-BERT: a pre-trained single-cell model for predicting the antigen-binding specificity of the adaptive immune receptor. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbad191 (2023).
37. Wang, Y. et al. An explainable language model for antibody specificity prediction using curated influenza hemagglutinin antibodies. *Biophys. J.* **123**, 3 (2024).
38. Barton, J., Galson, J. D. & Leem, J. Enhancing antibody language models with structural information. In *Machine Learning for Structural Biology Workshop* (NeurIPS, 2023).
39. Teney, D., Oh, S. J. & Abbasnejad, E. ID and OOD performance are sometimes inversely correlated on real-world datasets. In *37th Conference on Neural Information Processing Systems* (NeurIPS, 2023).
40. Chomsky, N. in *The Structure of Language: Readings in the Philosophy of Language* (eds Fodor, J. A. & Katz, J. J.) 50–118 (Prentice-Hall, 1964).
41. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
42. Chen, V. et al. Best practices for interpretable machine learning in computational biology. Preprint at *bioRxiv* 10.1101/2022.10.28.513978 (2022).
43. Sundermeyer, M., Schlüter, R. & Ney, H. LSTM neural networks for language modeling. In *Proc. Interspeech 2012* 194–197 (ISCA, 2012); https://doi.org/10.21437/Interspeech.2012-65
44. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst* **33**, 1877–1901 (2020).
45. Church, K. & Liberman, M. The future of computational linguistics: on beyond alchemy. *Front. Artif. Intell.* **4**, 625341 (2021).
46. Mielke, S. J. et al. Between words and characters: a brief history of open-vocabulary modeling and tokenization in NLP. Preprint at https://arxiv.org/abs/2112.10508 (2021).
47. Kutuzov, A. & Kuzmenko, E. To Lemmatize or not to lemmatize: how word normalisation affects ELMo performance in word sense disambiguation. In *Proc. First NLPL Workshop on Deep Learning for Natural Language Processing* 22–28 (Linköping Univ. Electronic Press, 2019).
48. Peters, M. E. et al. Deep contextualized word representations. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* 2227–2237 (Association for Computational Linguistics, 2018); https://doi.org/10.18653/v1/N18-1202
49. Olsen, T. H., Boyles, F. & Deane, C. M. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* **31**, 141–146 (2022).
50. Corrie, B. D. et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* **284**, 24–41 (2018).
51. Elhanati, Y. et al. Inferring processes underlying B-cell repertoire diversity. *Phil. Trans. R. Soc. B* **370**, 20140243 (2015).

52. Ferdous, S. & Martin, A. C. R. AbDb: antibody structure database—a database of PDB-derived antibody structures. *Database* **2018**, bay040 (2018).
53. Pan, Y., Li, X., Yang, Y. & Dong, R. Morphological Word Segmentation on Agglutinative Languages for Neural Machine Translation. Preprint at http://arxiv.org/abs/2001.01589 (2020).
54. Schwartz, L. et al. Neural polysynthetic language modelling. Preprint at https://arxiv.org/abs/2005.05477 (2019).
55. Adams, O., Makarucha, A., Neubig, G., Bird, S. & Cohn, T. Cross-lingual word embeddings for low-resource language modeling. In *Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 937–947 (Association for Computational Linguistics, 2017); https://doi.org/10.18653/v1/E17-1088
56. Agić, Ž., Hovy, D. & Søgaard, A. If all you have is a bit of the Bible: learning POS taggers for truly low-resource languages. In *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. (*Volume 2: Short Papers)* 268–272 (Association for Computational Linguistics, 2015); https://doi.org/10.3115/v1/P15-2044
57. Fang, M. & Cohn, T. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (eds Barzilay, R. & Kan, M.-Y.) 587–593 (Association for Computational Linguistics, 2017); https://doi.org/10.18653/v1/P17-2093
58. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, 561 (2018).
59. Dong, Y. et al. Structural principles of B cell antigen receptor assembly. *Nature* **612**, 156–161 (2022).
60. Wong, W. K. et al. Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. *mAbs* **13**, 1873478 (2021).
61. Antanasijevic, A. et al. From structure to sequence: antibody discovery using cryoEM. *Sci. Adv.* **8**, eabk2039 (2022).
62. Abu-Shmais, A. A. et al. Convergent sequence features of antiviral B cells. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.06.556442 (2023).
63. Sangesland, M. et al. Allelic polymorphism controls autoreactivity and vaccine elicitation of human broadly neutralizing antibodies against influenza virus. *Immunity* **55**, 1693–1709.e8 (2022).
64. Hauser, M. D., Chomsky, N. & Fitch, W. T. The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579 (2002).
65. Pantazes, R. J. et al. Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. *Sci. Rep.* **6**, 30312 (2016).
66. Shrock, E. L. et al. Germline-encoded amino acid–binding motifs drive immunodominant public antibody responses. *Science* **380**, eadc9498 (2023).
67. Aguilar Rangel, M. et al. Fragment-based computational design of antibodies targeting structured epitopes. *Sci. Adv.* **8**, eabp9540 (2022).
68. Zhou, J., Panaitiu, A. E. & Grigoryan, G. A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures. *Proc. Natl Acad. Sci. USA* **117**, 1059–1068 (2020).
69. Chomsky, N. Three models for the description of language. *IRE Trans. Inf. Theory* **2**, 113–124 (1956).
70. Rossmann, M. G. & Argos, P. Protein folding. *Annu. Rev. Biochem.* **50**, 497–532 (1981).
71. Qing, R. et al. Protein design: from the aspect of water solubility and stability. *Chem. Rev.* https://doi.org/10.1021/acs.chemrev.1c00757 (2022).
72. Searls, D. B. A primer in macromolecular linguistics. *Biopolymers* **99**, 203–217 (2013).
73. Hockenmaier, J., Joshi, A. K. & Dill, K. A. Routes are trees: the parsing perspective on protein folding. *Proteins Struct. Funct. Bioinform.* **66**, 1–15 (2006).
74. Hockenmaier, J., Joshi, A. K. & Dill, K. A. Protein folding and chart parsing. In *Proc. 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06* 293–300 (Association for Computational Linguistics, 2006); https://doi.org/10.3115/1610075.1610117
75. Dill, K. A. et al. Computational linguistics: a new tool for exploring biopolymer structures and statistical mechanics. *Polymer* **48**, 4289–4300 (2007).
76. Thellmann, K.-D., Stadler, B., Usbeck, R. & Lehmann, J. Transformer with tree-order encoding for neural program generation. Preprint at https://arxiv.org/abs/2206.13354 (2022).
77. AlQuraishi, M. End-to-end differentiable learning of protein structure. *Cell Syst.* **8**, 292–301.e3 (2019).
78. Zhang, L. *et al*. AnglesRefine: refinement of 3D protein structures using Transformer based on torsion angles. Preprint at *bioRxiv* https://doi.org/10.1101/2023.07.25.550599 (2023).
79. Malliavin, T. E., Mucherino, A., Lavor, C. & Liberti, L. Systematic exploration of protein conformational space using a distance geometry approach. *J. Chem. Inf. Model.* **59**, 4486–4503 (2019).
80. Sela-Culang, I., Kunik, V. & Ofran, Y. The structural basis of antibody–antigen recognition. *Front. Immunol.* **4**, 302 (2013).
81. Boughter, C. T. et al. Biochemical patterns of antibody polyreactivity revealed through a bioinformatics-based analysis of CDR loops. *eLife* **9**, e61393 (2020).
82. Bunker, J. J. et al. Natural polyreactive IgA antibodies coat the intestinal microbiota. *Science* **358**, eaan6619 (2017).
83. Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S. & Dimitrov, J. D. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. *Mol. Immunol.* **112**, 338–346 (2019).
84. Guo, J. Critical tokenization and its properties. *Comput. Linguist.* **23**, 569–596 (1997).
85. Hindle, D. & Rooth, M. Structural ambiguity and lexical relations. *Comput. Linguist.* **19**, 103–120 (1993).
86. Cunningham, O., Scott, M., Zhou, Z. S. & Finlay, W. J. J. Polyreactivity and polyspecificity in therapeutic antibody development: risk factors for failure in preclinical and clinical development campaigns. *mAbs* **13**, 1999195 (2021).
87. Fernández-Quintero, M. L. et al. Characterizing the diversity of the CDR-H3 loop conformational ensembles in relationship to antibody binding properties. *Front. Immunol.* **9**, 3065 (2019).
88. Bachas, S. et al. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.16.504181 (2022).
89. Makowski, E. K. et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat. Commun.* **13**, 3788 (2022).
90. Pittala, S. & Bailey-Kellogg, C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics* **36**, 3996–4003 (2020).
91. Jespersen, M. C., Mahajan, S., Peters, B., Nielsen, M. & Marcatili, P. Antibody specific B-cell epitope predictions: leveraging information from antibody–antigen protein complexes. *Front. Immunol.* **10**, 298 (2019).
92. Del Vecchio, A., Deac, A., Liò, P. & Veličković, P. Neural message passing for joint paratope-epitope prediction. In *2021 ICML Workshop on Computational Biology* (2021).

93. Brown, A. J. et al. Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. *Mol. Syst. Des. Eng.* **4**, 701–736 (2019).
94. de Saussure, F. *Course in General Linguistics* (Open Court, 1986).
95. Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl Acad. Sci. USA* **73**, 3628–3632 (1976).
96. Adams, R. M., Kinney, J. B., Walczak, A. M. & Mora, T. Epistasis in a fitness landscape defined by antibody–antigen binding free energy. *Cell Syst.* **8**, 86–93.e3 (2019).
97. Linzen, T., Dupoux, E. & Goldberg, Y. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguist.* **4**, 521–535 (2016).
98. Goldberg, Y. Assessing BERT's syntactic abilities. Preprint at https://arxiv.org/abs/1901.05287 (2019).
99. Erk, K. The probabilistic turn in semantics and pragmatics. *Annu. Rev. Linguist.* **8**, 101–121 (2022).
100. Sutton, P. R. Towards a probabilistic semantics for vague adjectives. In *Bayesian Natural Language Semantics and Pragmatics* (eds Zeevat, H. & Schmitz, H.-C.) 221–246 (Springer, 2015); https://doi.org/10.1007/978-3-319-17064-0_10
101. Baroni, M. & Zamparelli, R. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proc. 2010 Conference on Empirical Methods in Natural Language Processing* 1183–1193 (Association for Computational Linguistics, 2010).
102. Clark, S., Coecke, B. & Sadrzadeh, M. A compositional distributional model of meaning. in *Proceedings of the Second Symposium on Quantum Interaction* (eds Bruza, P. et al.) 133–140 (Oxford, 2008).
103. Sadrzadeh, M. & Kartsaklis, D. Compositional distributional models of meaning. In *Proc. COLING 2016 26th International Conference on Computational Linguistics: Tutorial Abstracts* (eds Matsumoto, Y. & Prasad, R) 1–4 (2016).
104. McCoy, R. T., Frank, R. & Linzen, T. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Trans. Assoc. Comput. Linguist.* **8**, 125–140 (2020).
105. Harer, J., Reale, C. & Chin, P. Tree-Transformer: a transformer-based method for correction of tree-structured data. Preprint at https://arxiv.org/abs/1908.00449 (2019).
106. Akbar, R. et al. In silico proof of principle of machine learning-based antibody design at unconstrained scale. *mAbs* **14**, 2031482 (2022).
107. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. in *The Twelfth International Conference on Learning Representations* (2024).
108. Varadi, M. et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
109. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
110. Abanades, B. et al. ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. *Commun. Biol.* **6**, 575 (2023).
111. Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat. Commun.* **14**, 2389 (2023).
112. Fang, X. et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-023-00721-6 (2023).
113. Ambrosetti, F., Jiménez-García, B., Roel-Touris, J. & Bonvin, A. M. J. J. Modeling antibody–antigen complexes by information-driven docking. *Structure* **28**, 119–129.e2 (2020).
114. Sandve, G. K. & Greiff, V. Access to ground truth at unconstrained size makes simulated data as indispensable as experimental data for bioinformatics methods development and benchmarking. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac612 (2022).
115. Fernández-Quintero, M. L. et al. Challenges in antibody structure prediction. *mAbs* **15**, 1 (2023).
116. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).
117. Wang, Y., Lamim Ribeiro, J. M. & Tiwary, P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **61**, 139–145 (2020).
118. Doerr, S. et al. TorchMD: a deep learning framework for molecular simulations. *J. Chem. Theory Comput.* **17**, 2355–2363 (2021).
119. Jackson, N. E., Savoie, B. M., Statt, A. & Webb, M. A. Introduction to machine learning for molecular simulation. *J. Chem. Theory Comput.* **19**, 4335–4337 (2023).
120. Yang, Y. I., Shao, Q., Zhang, J., Yang, L. & Gao, Y. Q. Enhanced sampling in molecular dynamics. *J. Chem. Phys.* **151**, 070902 (2019).
121. Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).
122. Heinz, J. in *The Oxford Handbook of Developmental Linguistics* Vol. 1 (eds Lidz, J. L. et al.) 633–663 (Oxford Univ. Press, 2016).
123. Wilson, M., Petty, J. & Frank, R. How abstract is linguistic generalization in large language models? Experiments with argument structure. *Trans. Assoc. Comput. Linguist.* **11**, 1377–1395 (2023).
124. Delétang, G. et al. Neural networks and the Chomsky hierarchy. In *11th International Conference on Learning Representations, ICLR 2023* (2023).
125. Bhattamishra, S., Ahuja, K. & Goyal, N. On the ability and limitations of transformers to recognize formal languages. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 7096–7116 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.emnlp-main.576
126. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
127. Luo, S. et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Adv. Neural Inf. Process. Syst.* **35**, 9754–9767 (2022).
128. Keidar, D., Opedal, A., Jin, Z. & Sachan, M. Slangvolution: a causal analysis of semantic change and frequency dynamics in slang. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Muresan, S. et al.) 1422–1442 (Association for Computational Linguistics, 2022); https://doi.org/10.18653/v1/2022.acl-long.101
129. Kutuzov, A., Øvrelid, L., Szymanski, T. & Velldal, E. Diachronic word embeddings and semantic shifts: a survey. In *Proc. 27th International Conference on Computational Linguistics*. (eds Bender, E. M. et al.) 1384–1397 (Association for Computational Linguistics, 2018).
130. Krovi, S. H., Kappler, J. W., Marrack, P. & Gapin, L. Inherent reactivity of unselected TCR repertoires to peptide-MHC molecules. *Proc. Natl Acad. Sci. USA* **116**, 22252–22261 (2019).
131. Chomsky, N. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought* (Cambridge Univ. Press, 2009).
132. Perelson, A. S. Immune network theory. *Immunol. Rev.* **110**, 5 (1989).
133. Coutinho, A. The self-nonself discrimination and the nature and acquisition of the antibody repertoire. *Ann. Immunol.* **131D**, 235–253 (1980).

134. Piattelli-Palmarini, M. The rise of selective theories: A case study and some lessons from immunology. In *Language Learning and Concept Acquisition* (ed. Demopoulos, W.) Ch. 5 (Ablex, 1986).
135. Piattelli-Palmarini, M. & Uriagereka, J. The immune syntax: The evolution of the language virus. In *Variation and universals in biolinguistics* (ed. Jenkins, L.) 341–377 (Brill, 2004).
136. *The Semiotics of Cellular Communication in the Immune System* (Springer, 1988); https://doi.org/10.1007/978-3-642-73145-7
137. Atlan, H. & Cohen, I. R. Immune information, self-organization and meaning. *Int. Immunol.* **10**, 711–717 (1998).

Acknowledgements

Supported by the Leona M. and Harry B. Helmsley Charitable Trust (number 2019PG-T1D011, to V.G.), UiO World-Leading Research Community (to V.G.), UiO: LifeScience Convergence Environment Immunolingo (to V.G. and G.K.S.), EU Horizon 2020 iReceptorplus (number 825821) (to V.G.), a Norwegian Cancer Society Grant (number 215817, to V.G.), Research Council of Norway projects (numbers 300740, and 331890 to V.G.), a Research Council of Norway IKTPLUSS project (number 311341, to V.G. and G.K.S.), and Stiftelsen Kristian Gerhard Jebsen (K. G. Jebsen Coeliac Disease Research Centre, to G.K.S.).

Author contributions

All authors contributed to the conceptualization, writing and editing of this paper.

Competing interests

V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Specifica Inc, Adaptyv Biosystems, EVQLV, Omniscope, Diagonal Therapeutics and Absci. V.G. is a consultant for Roche/Genentech, immunai, Proteinea and LabGenius.

Additional information

Correspondence should be addressed to Mai Ha Vu or Victor Greiff.

Peer review information *Nature Computational Science* thanks Sheng-ce Tao and Hao Zhou for their contribution to the peer review of this work. Primary Handling Editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team.

Reprints and permissions information is available at www.nature.com/reprints.