Communications Biology

https://doi.org/10.1038/s42003-026-09727-3

Article in Press

LICHEN enables light-chain immunoglobulin sequence generation conditioned on the heavy chain and experimental needs

Received: 15 August 2025

Accepted: 6 February 2026

Cite this article as: Capel, H.L., Ellmen, I., Murray, C.J. *et al.* LICHEN enables light-chain immunoglobulin sequence generation conditioned on the heavy chain and experimental needs. *Commun Biol* (2026). https://doi.org/10.1038/s42003-026-09727-3

Henriette L. Capel, Isaac Ellmen, Chris J. Murray, Giulia Mignone, Megan Black, Brendan Clarke, Conor Breen, Sean Tierney, Patrick Dougan, Richard J. Buick, Alexander Greenshields-Watson & Charlotte M. Deane

LICHEN enables Light-chain Immunoglobulin sequence generation Conditioned on the Heavy chain and Experimental Needs

Henriette L. Capel[1], Isaac Ellmen[1], Chris J. Murray[2], Giulia Mignone[2], Megan Black[2], Brendan Clarke[2], Conor Breen[2], Sean Tierney[2], Patrick Dougan[2], Richard J. Buick[2], Alexander Greenshields-Watson[1], and Charlotte M. Deane[1,✉]

[1]Oxford Protein Informatics Group, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB United Kingdom
[2]Fusion Antibodies plc, Springbank Industrial Estate, 1 Springbank Road, Dunmurry, Belfast, BT17 0QL United Kingdom

In developing therapeutic antibodies, the heavy chain is often prioritised due to its higher variability and its central role in antigen binding. An appropriate pairing of the light sequence is however important for antibody function. Here we present LICHEN, a heavy chain conditioned light sequence generation tool that enables collaborative light sequence design by leveraging computational capabilities alongside experimental expertise. LICHEN generates light sequences which are valid (antibody-like), diverse in sequence and structure, and conditioned on a specific heavy chain. LICHEN can also condition on germline and CDRs and automatically filter generated sequences for required properties. We carry out experimental validation of the method conditioning only on the heavy sequence and on the heavy sequence and binding information. Our *in vitro* results show that sequences created by LICHEN have effective expression yields and can retain antigen-binding. LICHEN can thus be used across multiple antibody engineering scenarios for efficient light-chain pairing.

Correspondence: ***deane@stats.ox.ac.uk***

Introduction

Antibodies are essential proteins of the adaptive immune system. The specific binding of the antigen with high affinity, coupled with the potential to further mutate in response to a target makes antibodies an excellent potential therapeutic class[1]. A monoclonal antibody consists of two identical heavy chains paired with two identical smaller light chains[2]. The complementarity determining regions (CDRs), located at the tips of the variable region (Fv), contain most variability and mediate antigen binding. The variable heavy (VH) and light (VL) chain contain three CDRs each, with CDRH3 being the most variable.

Diversity of the CDRH3 derives from the combinatorial and junctional diversity of the three genes (V, D, and J) encoding the full VH sequence. The VL sequence is encoded by only the V and J-gene. During B-cell development, the VH gene locus is first rearranged, after which the VL chain is rearranged to produce a suitable light chain that pairs with the heavy chain[3]. Upon encountering a specific antigen, these germline encoded "naive" VH-VL sequences undergo an affinity maturation process which creates sequences that bind strongly to that antigen[3].

Modern antibody discovery for therapeutic purposes often starts with library-based discovery[4]. Generative language models, such as IgLM[5] and p-IgGen[6], allow the generation of large synthetic antibody libraries. IgLM and p-IgGen are both decoder-only models, originally designed for text generation tasks[7]. IgLM is trained on single chains, species and chain type. p-IgGen is trained on paired data - after pre-training on unpaired data - and can therefore generate paired VH and VL sequences. Developability issues for antibodies generated by language models have been reported [8] and these models are unaware of the antigen. Thus, generic libraries designed by these methods require optimisation for affinity to the target antigen, and developability during downstream antibody design pipelines.

In the development of antibody therapeutics there has been a focus on the heavy sequence due to its greater variability[9] and its importance in binding[10]. This is exemplified in the availability of unpaired sequence data. For example, there are six times as many VH sequences as VL sequences in OAS[1][11]. However, it is known that the proper pairing of the light chain to a heavy sequence is essential for the functionality of an antibody[12–14].

Constructing the appropriate light sequence for a given heavy chain could be done using generative models, especially when combined with prior attained knowledge on key regions for antigen binding or germline usage. This maximises the potential of the computational tools to guide, and experimental knowledge to inform the antibody engineering workflow.

Here, we present LICHEN, a light sequence generation tool tailored to a given heavy sequence and (if desired and/or available) to experimental prior knowledge. LICHEN is a sequence-to-sequence model trained on natural paired human Fv sequences. We demonstrate that generated sequences are valid human light sequences, which are diverse and a fit for the heavy sequence, simulating patterns observed in nature. In addition to the heavy sequence, LICHEN flexibly supports conditioning on CDR sequences and germline information, and offers applications when only heavy sequence information is available (e.g. from unpaired heavy sequencing), as well as when additional binding information is available (e.g. sequences derived from animal immunisation).

[1]https://opig.stats.ox.ac.uk/webapps/oas, assessed on 5 June 2025

Germline guidance improves applicability to experimental settings (e.g. when specific V-gene usage is preferred). This is further enhanced by automatic sequence filtering options. *In vitro* we show that light sequences generated by LICHEN express well and maintain antigen-binding when CDR information is provided. LICHEN is available open source as both a python package (https://github.com/oxpig/LICHEN) and web application (https://opig.stats.ox.ac.uk/webapps/lichen).

Results

LICHEN is a sequence-to-sequence model trained on human Fv sequences that generates antibody light sequences conditioned on the heavy chain. We evaluated generated light sequences based on validity, diversity, and fitness for the given heavy sequence. We experimentally validated LICHEN on a pairing use case and two pairing while maintaining antigen-binding use cases.

LICHEN generates valid human light sequences. For the 20 generated light sequences for each of the 500 heavy sequences in the standard_ds we evaluated their antibody light chain likeliness. All generated sequences were identified by ANARCII as light sequences and could be numbered. Almost all sequences are human according to Hu-mAb (99.92%) and Humatch (99.73%) and could be structurally modelled with ABB2 (99.80%) (see Table S1).

The VL and CDR sequence lengths show similar distributions to those observed in native light sequences (see Figure S1).

Generated sequences are diverse. The diversity of sequences generated by LICHEN is in line with the variation observed in nature. According to germline assignment by ANARCI, generated sequences are closest to a diverse set of light chain V and J-genes and contain a varying number of mutations away from these germlines (see Figure S2). Light chain V-gene distributions when conditioned on the different heavy chain germlines in the genes_ds shows that LICHEN has a bias towards light V-genes more often observed in OAS [15] (see Figure S2). On average, generated sequences by LICHEN contain more mutations away from the germline V-gene compared to the native light sequences (see Figure 1). Positional mutation counts and amino acid frequencies, as well as edit distance and the percentage of mutations in the CDRs and framework regions (FRs) from the germline are shown in Figure S3. The canonical forms - the structural conformation - of the CDRs are equally diverse (see Figure S4).

Diversity is not only observed in light sequences generated for diverse heavy sequences, but also for a single heavy sequence (see Figure S5). We calculate the structural diversity of the light sequence CDRs using canonical form assignment, the Fv by the VH-VL orientation, and the unaltered CDRH3 sequence due to VL pairing by RMSD calculations. Again, the diversity is in line with the diversity of light sequences naturally pairing with the same heavy sequence (see Figure S5).

In general, LICHEN is more likely to generate light sequences of the most frequent light chain types in the training data (see Figure S2). However, these sequences are not just germline sequences and LICHEN is able to generate novel full VL, CDR and CDRL3 sequences compared to light sequences used for training and validation and does not just re-capitulate sequences observed during training (see Figure S6).

LICHEN has learned the co-evolutionary relationship between heavy and light sequences. It is known that during the antibody affinity maturation process sequences undergo somatic hypermutations in both chains. Optimised antibodies for the antigen are therefore expected to have a high number of mutations in both the heavy and light sequence. Figure 2 shows this co-evolutionary relationship between the VH and VL chain of the native sequences based on the average germline identity score of both the V and J-gene according to ANARCI. This relationship is not observed when randomly pairing the native heavy and light chains. Generated light sequences show a similarly strong correlation, indicating that LICHEN has captured this co-evolutionary relationship.

We next tested if LICHEN also prefers pairing of the optimised memory heavy sequence with the optimised memory light sequence over pairing with the germline-reverted light sequence using the conditioning_ds. We therefore extracted log likelihood scores of specific light sequences with existing heavy sequences. The log likelihood score, as well as perplexity scores are available from LICHEN (https://github.com/oxpig/LICHEN). Figure 3 shows native pairings are preferred over germline-reverted pairings with a stronger preference observed when the light sequence contains at least ten mutations away from the germline.

LICHEN balances validity and diversity of generated light sequences. Diversity from the germline and fitness for the heavy sequence suggests LICHEN provides a valuable alternative to the standard strategy of taking a common germline light sequence to partner a heavy chain. We next compared LICHEN against an alternative machine learning model: p-IgGen[6]. p-IgGen is a generative model for paired antibodies which can be seeded with initial sequential information; i.e. a light sequence can in principle be generated for a given heavy sequence. p-IgGen is not specifically trained for this task. Based on 10 generated light sequences for each of the 400 heavy sequences in the comparison_ds, p-IgGen is more likely to generate sequences that cannot be numbered and/or are not human according to *in silico* methods (see Table S2).

LICHEN can be tailored to maintain key CDRs and select specific germlines. Existing generative models, such as p-IgGen, are unable to incorporate such additional information obtained prior to sequence design. LICHEN is designed to be able to incorporate binding information by con-

ditioning on light chain CDRs. To accommodate for experimental choices on germline usage, e.g. when a specific V-gene is preferred based on prior knowledge of its characteristics or user expertise, LICHEN can be restricted to generation of a specific light chain type (kappa or lambda), V-gene family, or V-gene. For example, a kappa light chain might be preferred as a lambda light chains are in general at greater risk of developability issues[1]. We designed two *in vitro* case studies to demonstrate how LICHEN can be restricted to specific V-gene families and make use of CDR information for pairing.

Sequences generated by LICHEN express well. For both the VH sequence of therapeutics adalimumab and pembrolizumab we generated VL sequences for three use cases: pairing_CDRs, pairing_eCDR, and pairing (see Methods). The generated sequences were evaluated *in silico*, after which a diverse set of sequences was selected for *in vitro* validation. Although adalimumab and pembrolizumab are optimised therapeutics, the light sequences generated by LICHEN and paired with the therapeutic VH exceeded the expression yield in 48% and 65% cases, respectively (see Figure 4A and 4C, and Table S3). Only one antibody sequence generated by LICHEN for adalimumab, a IGKV2-24*01 light sequence, produced insufficient material. Reduced expression yields for pembrolizumab are observed when CDR information is incorporated for both LICHEN and baseline antibodies. The potential yield and monodispersity of all samples as well as the closest V- and J-gene of these sequences (according to ANARCI) are shown in Table S3. These results show that LICHEN can generate antibodies that are stable and express well.

CDR tailored sequences generated by LICHEN maintain antigen-binding. Conditioning LICHEN on the light CDRs of adalimumab resulted in highly diverse sequences (see Figure S7) with strong expression yields while maintaining binding to TNF-α (see Figure 4B). Providing only CDRL3 (pairing_eCDR) created some weak binders. Similarly for pembrolizumab (Figure 4D), generated antibodies with strong expression yield maintain binding ability to PD-1, and weak binders are observed among the paired_eCDR antibodies. Independent of expression yields, all germline sequences used as baselines maintain antigen-binding. The ELISA binding patterns are in line with BLI results, showing high on-rates and very slow dissociation for the high affinity binders (see Figure S8). Antigen-binding was not expected for the variants in the "pairing" case study (LICHEN-14 to LICHEN-23) as no additional CDR information was provided (see Figure S9). These results indicate that when conditioning on binding information, LICHEN generates diverse light sequences with sensible expression yields that maintain target binding.

Discussion

Generative machine learning models have shown success for antibody library design. However, methods are often explicitly designed for the initial exploration process and are purely computational. In practice, machine learning guidance and experimental design and validation go hand-in-hand.

To support the collaborative design of antibody therapeutics, we have developed LICHEN. It is a sequence-to-sequence model for light sequence generation. Light sequences are generated in a next-token fashion, conditioned on the heavy chain and preceding generated light chain residues. LICHEN can also be conditioned by the user on CDRs and germline usage. This allows for generation of light sequences with the key CDRs for antigen binding. The three CDRs can be provided in any combination (e.g. only CDRL3) and according to both the IMGT[16] and Kabat[17] definitions. Limiting LICHEN to a specific light chain type (kappa or lambda), V-gene family, or V-gene allows for experimental choices.

This tailoring to experimental needs offers more practical applications compared to previous methods. LICHEN can be used in settings where only VH information or both VH and binding information is available. For example, LICHEN can be used to restore missing light chains, generate libraries for light chain shuffling, humanise a light sequence by providing LICHEN the humanised heavy sequence and the parental light sequence CDR of non-human, and improve biophysical properties by exploring alternative light chain usage. Furthermore, LICHEN allows two heavy sequences as input to find a common light sequence for a bispecific antibody. The automatic validation of generated sequences by LICHEN reduces the amount of required manual downstream analysis of the outputs.

Publicly available paired antibody data used for training LICHEN is biased towards specific V-genes, particularly to IGHV3, IGKV1, and IGKV3. Sequences generated by LICHEN show similar light chain biases. This preference is also observed in available therapeutic antibodies and was previously indicated as systematic bias in drug discovery pipelines[1].

Evaluating antibody generative models *in silico* is challenging, especially for heavy and light chain pairing. Heavy sequences have been shown to pair with diverse light sequences and vice versa, making the definition of an appropriate pairing unclear. Here, we constructed tests evaluating conventional light chain and pairing features. *In vitro* validations on two therapeutic VH chains confirmed that LICHEN generates diverse light sequences that are a fit in terms of expression and stability for these heavy chains.

During evaluation, we focused on finding an optimal balance between correctness and diversity of generated light sequences. Residues were selected using a top-p sampling approach allowing for generating diverse light sequences. The sequence and structural diversity generated by LICHEN for a given heavy sequence offers a broad exploration of the potential design space.

No increase in model performance was observed when increasing the number of parameters. This might be related to the relatively small amount of available paired antibody data. Extending the training data with the vast amount of unpaired sequence data did not improve performance. As more data

became available after initially training LICHEN - resulting in OAS storing more than 3M paired human sequences (June 2025) - we retrained the model. Although the amount of data after filtering was increased (from 1.8M to 2.5M pairs), the data coverage is similar (see Figure S10) and no model improvement was observed.

To conclude, LICHEN generates human light sequences for a given heavy sequence. These light sequences are valid, diverse, and a fit for the heavy sequence. Additional tailoring towards key binding regions and restricting outputs to light chain types incorporates experimental needs. With experimental validations, we confirmed that the diverse light sequences created by LICHEN for two therapeutic VH chains form a stable and expressible antibody and, when conditioned on CDR information, maintain binding to the target. LICHEN thus offers a hybrid design strategy for diverse use cases at various stages in the antibody engineering workflow.

Methods

Dataset generation. Paired human VH-VL sequences were curated from the publicly available database OAS[11] in November 2023. Sequences missing the conserved cysteines in IMGT positions 23 and 104 were removed, as well as sequences with unknown residues and duplicate pairs. Deletions in framework residues, observed in around 8% of the antibody pairs, were restored with AbLang[18]. The remaining 1.8M paired VH and VL sequences were randomly split into 80% training, 10% validation, and 10% testing. Only exact duplicate heavy sequences were excluded from the test set. The filtered and cleaned data for training and testing have been deposited on Zenodo (https://doi.org/10.5281/zenodo.15917096). Training was performed on 1,452,229 paired sequences.

The performance of LICHEN was tested on a random subset of the test set, the "standard_ds". We selected 500 test set samples and generated 20 light sequences for each heavy sequence. When not specified otherwise, the standard_ds is used for evaluation. We evaluated the V-gene diversity of generated light sequences for different sets of heavy V-genes. For this "genes_ds" we selected 50 antibodies of all 47 heavy V-gene observed at least 50 times in the test set and generated ten light sequences for each of these heavy sequences. To explore sequence diversity when LICHEN is used to generate light sequences for heavy sequences on which the model is trained, we generated 20 light sequences for 500 heavy sequences in the training set, called the "observed_ds". For structural evaluation we selected five heavy sequences from five different V-gene families which all naturally paired with more than 20 light sequences, the "structural_ds". We sampled and generated 20 light sequences for these heavy sequences. The conditioning ability of LICHEN was evaluated on the "conditioning_ds". We selected 1000 test set samples each with between five and fifteen mutations away from the germline in both chains. Germline reverted light sequences were constructed by mutating the identified mutations back to the closest germline V-gene and J-gene (as determined by ANARCI[19]). For a fair comparison against p-IgGen[6], in which test sequences are selected on 95% concatenated CDR sequences across both chains, we created the "comparison_ds". We clustered our data on 95% concatenated heavy chain CDR sequences and sampled 400 sequences different from training and validation, and present in both the test set of LICHEN and p-IgGen.

Model architecture. LICHEN is a sequence-to-sequence model implemented according to the transformer from Vaswani et. al[20] in PyTorch[21]. LICHEN consist of 6 encoder and decoder layers, all with 8 heads. The embedding size and the number of hidden dimensions were both set to 512, resulting in a total number of parameters of 25,284,632. Parameters were initialised using a Xavier normal distribution (Glorot initialization)[22].

The antibody sequences were tokenised using one-hot encoding of the amino acids as used in AbLang[18] and a start and stop token.

Training protocol. LICHEN was trained using sinusoidal positional encoding for 4 epochs. The model was optimised using the Adam optimizer with default parameters except for the learning rate. We used a cosine scheduled learning rate with 5% warm-up steps and a maximum learning rate of 10^{-4}, and a batch size of 64, a dropout of 0.1, and cross entropy loss. Training was performed on a single Quadro RTX 6000 GPU.

Model inference. Light sequences were generated with a top-p parameter of 0.9 and a temperature parameter of 1.0. Minimal tuning of the two parameters showed that these values generate diverse and realistic light sequences.

LICHEN allows germline and CDR sequence information to be provided during inference. Users can restrict the model to generate a specific type (kappa or lambda), V-gene family (e.g. IGKV1), or V-gene (IGKV1-39). LICHEN is therefore seeded with either the first two (for type) or ten (for V-gene family and V-gene) residues. Initial residues are sampled based on observed frequency in germline sequences as stored in IMGT[2]. Users can also provide sequences of any length as a seed.

LICHEN can incorporate sequences of all three or a subset of the CDRs according to either the Kabat[17] or IMGT[16] definitions. As the architecture conditions the selection of the next residue on both the heavy sequence and previously generated light residues, LICHEN conditions on provided additional germline and CDR data. Correct CDR placement is established based on conserved residues and is checked by ANARCII[9]. When CDR1 and CDR2 are provided without a light chain type, LICHEN is automatically seeded with the kappa or lambda type based on the closest alignment score of the CDRs against germline CDR residues using BLOSUM62[23]. CDR1 is placed based on the conserved cysteine in IMGT position 23. Placing of conserved tryptophan in IMGT position 41 is forced after CDR1. CDR2 is

[2] https://www.imgt.org/genedb/, assessed in January 2025

placed based on the CDR1 or tryptophan 41, and the resulting sequence thus has a fixed FR2 length. CDR3 is placed according to the conserved cysteine in IMGT position 104 and subsequently a phenylalanine is forced after CDR3.
Generated light sequences for a specific heavy sequence can be automatically filtered by LICHEN. The filters implemented include: identification by ANARCII[9], human by Humatch[24], and most likely by AbLang2[25]. Note that filtering on AbLang2 will result in germline diverse sequences. Moreover, filtering can be applied to remove duplicate sequences and to select the most variable set of generated sequences. This diverse set of antibodies is efficiently selected by evenly spaced sampling of ordered AbLang2 confidence scores.

Model evaluation. Light sequences generated by LICHEN should be valid, diverse, and a fit for the given heavy sequence.
The validity of generated sequences was checked by the ability of ANARCI[19] and ANARCII[9] to number the sequence and classify the sequence as a light sequence. Humanness of the sequence was checked using Hu-mAb[26] and Hu-match[24]. The full VH and CDR sequence lengths were compared against the native paired light sequences. Combined CDR sequence length was compared against all native CDR lengths in the dataset. Structural diversity was analysed using canonical form assignment by SCALOP[27]. ABodyBuilder2 (ABB2)[28] was used to structurally model the sequences.
The diversity of the generated sequences was evaluated by germline assignment and germline identity score as given by ANARCI. Sequence-based similarity of generated sequences to the training and validation data was determined by KAsearch[29]. The structural diversity was evaluated on the structural_ds by canonical form assignment of CDR sequences using SCALOP, VH-VL orientation of ABB2 models by ABangle[30], and CDRH3 RMSD of ABB2 models. The latter was calculated by aligning the full VH sequence of the ABB2 antibody models with the 20 native or generated light sequences to their mean.
The fit of the generated light sequences for the heavy chain was evaluated based on the co-evolutionary relationship between the VH and VL arising from the affinity maturation process for the cognate antigen. This was checked based on the correlation between the number of mutations in the VH and VL, as well as the model preference for native pairing. For the former ANARCI germline identity scores were used. For the latter pairing probability scores were calculated on the conditioning_test. The conditional probability of the native light sequence given the native heavy sequences was normalised by the unconditional probability. This normalised probability was then compared to the normalised probability of the germline reverted light sequence given the native heavy sequence. Probability scores were calculated based on next token logits.
Model performance of LICHEN and p-IgGen were compared on the basis of ten generated light sequences for all heavy sequences in the comparison_ds. p-IgGen was therefore seeded with the heavy sequence.

Experimental validation design. We also evaluated LICHEN experimentally by using it to predict sets of light chains for the therapeutics adalimumab[31] and pembrolizumab[32]. Adalimumab is a fully human antibody (IGHV3-IGKV1) targeting TNF-α, a pro-inflamatory cytokine. Pembrolizumab is a humanised antibody (IGHV1-IGKV3) targeting programmed death receptor 1 (PD-1). Three different use cases of LICHEN were evaluated: a case study with known CDRs (pairing_CDRs), a case study with the essential CDRs (pairing_eCDR), and a pairing case without CDR knowledge (pairing). The CDRL3 was assumed to be an essential CDR, due to its importance in antigen-binding compared to other light sequence regions [33]. For pembrolizumab the CDRL1 was also classified as essential because of its IMGT sequence length of 10, which is rarely observed in natural sequences [27]. For all cases, LICHEN was conditioned to make solely IGKV1, IGKV2, and IGKV3 light sequences for the native heavy sequences as these are most common.
Generated sequences were filtered on humanness by Humatch[24], ANARCI[19], and Hu-mAb[26]. Sequences with known canonical forms according to SCALOP[27], and that could be modelled with ABB2[28], were selected. Developability was tested using TAP[34] and only sequences with five green flags were selected for pembrolizumab. For adalimumab four green flags and a minimum score for the structural Fv charge symmetry parameter SFvCSP equal to native adalimumab (-19.5) were allowed. For both therapeutic antibodies five light sequences were selected for the "pairing_CDRs" case, eight for the "pairing_eCDR" case, and ten for the "pairing" case (see Figure S7). For all three cases we selected diverse sequences.
The native light sequence was used as a positive control. Germline encoded V-gene sequences combined with J-gene IGKJ1*01 were used as baseline. IGKV1-39*01 and IGKV1D-33*01 were chosen as baseline for adalimumab IGKV3-20*01 and IGKV3-11*01 for pembrolizumab based on known binding preferences[35]. The more rarely observed germline sequence IGKV7-3*01 was additionally used as a baseline. Baseline sequences were tested with and without native CDRs grafted for comparison to the CDR informed and standard pairing cases.

Experimental validation. DNA coding for the amino acid sequence of each antibody was synthesised and cloned into the mammalian transient expression plasmid pETE V3 (property of Fusion Antibodies plc). Antibodies were expressed in IgG1 format using a CHO-based transient expression system and the resulting antibodies were clarified by centrifugation and filtration. Antibodies were purified (using state-of-the-art AKTA chromatography equipment) from cell culture supernatants via affinity chromatography and analysed via size exclusion chromatography (SEC). The purity of the antibodies was determined to be >95%, as judged by reducing and denaturing Sodium Dodecyl Sulfate Polyacrylamide gels. SEC was run using a Superdex 200 Increase 10/300 GL

column. Antibody concentration was determined by measuring absorbance at 280 nm and calculated using the standard extinction coefficient 205,500 $M^{-1}cm^{-1}$ (or 1.0 mg/ml = A280 of 1.37 [assuming a MW = 150,000 Da]) for an antibody.

A screening ELISA was prepared for all antibody variants created by LICHEN while conditioning on CDR data and controls that expressed and purified well. All antigens used in ELISA screening were purchased from Acro Biosciences. Purified TNF-α (Cat #TNA-H82E3, Lot #BV2043-241DF1-1R8) was used to test adalimumab and Human PD-1 (Cat #TNA-H82E3, Lot #BV2043-241DF1-1R8) was used to test pembrolizumab. A negative antigen (SARS-CoV-2 Spike RBD protein, Cat #SPD-C82E9, Lot #BV3541b-2043F1-RD) was included to ensure binding specificity. Antigens were diluted to a concentration of 1 µg/mL in bicarbonate ELISA coating buffer and 30 µL (30 ng/well) was added to each well of a 384 well MaxiSorp plate (Nunclon). Plates were then incubated overnight at 4°C. Plates were washed 5 times with phosphate buffered saline (PBS) + 0.1% Tween using a Zoom HT plate washer (Berthold Technologies) and blocked with PBS containing 4% non-fat dry milk for 2 hours at room temperature, with shaking. All antibodies were diluted to a concentration of 1 µg/mL in PBS and 30 µL was added to each well (30 ng/well). Diluted antibodies were added to the plates with shaking for 2 hours at room temperature. Additional wells were coated with antigen and PBS was added in place of the adalimumab/pembrolizumab antibodies. Following sample incubation, plates were washed 5 times in PBS + 0.1% Tween using the Zoom plate washer before 1 hour incubation in Goat anti-human IgG – Fc specific HRP secondary antibody (1:70,000 dilution, Sigma A0170). Plates were washed a final time as before and 40 µL TMB II solution (Biopanda reagents) was added for 10 minutes at 37°C. The reaction was terminated by adding 20 µL of 1M Hydrochloric acid per well. Absorbance at 450 nm was then assessed using a Clariostar plate reader (BMG Labtech).

Variants that produced a positive signal (>0.2 nm at 450 nm) with the relevant antigen during ELISA analysis were selected for further binding analysis by Biolayer interferometry (BLI). A single-point analysis assay was performed by capturing the biotinylated TNF-α at 0.53 µg/mL and PD-1 at 0.28 µg/mL using SAX2 biosensors (Cat #18-5136 from Sartorius). The antigen-captured biosensors were submerged in wells containing each antibody, prepared at a single concentration of 33 nM (association stage), followed by a dissociation step in running buffer. To allow for double reference correction, antigen-captured sensors were dipped into wells containing only buffer. Blank sensors (no antigen present) were also dipped into wells containing each antibody. This referencing provided a means of compensating for any non-specific binding of the antibody to the sensor surface and also any baseline drift in the running buffer. Steps were performed at 25°C at a constant flow-rate of 1000 rpm. New sensors were used for each sample. All consumables used were those recommended by Sartorius. All samples were diluted in freshly prepared running buffer. Antigens were immobilized onto the surface of biosensors using the capture methods described previously. Antibody variants were passed over the surface to generate a binding response. Binding data for the Antibody:Antigen interactions were collected at 25°C on the biosensors.

Statistics and Reproducibility. The statistical linear correlation in Figure 2 was determined by the Pearson correlation coefficient implementation of SciPy[36] on 500 samples. No replicates were performed for the experimental analysis.

Data availability

The filtered and cleaned data for training and testing have been deposited on Zenodo (https://doi.org/10.5281/zenodo.15917096[37]) . Numeric source data are provided in the Supplementary Data.

Code availability

LICHEN is available open source as both a python package (https://github.com/oxpig/LICHEN and https://zenodo.org/records/18459226[38]) and web application (https://opig.stats.ox.ac.uk/webapps/lichen).

ACKNOWLEDGEMENTS
This work was supported by the Engineering and Physical Sciences Research Council (grant number EP/S024093/1) and research funding by Fusion Antibodies plc awarded to HC and IE, and research funding from Exscientia awarded to AGW. The authors thank Benjamin H. Williams for his guidance on designing the web tool, and Lorna M. Stewart for helpful conversations on antibody design.

AUTHOR CONTRIBUTIONS
CMD and IE conceptualised the study. CMD, AGW, RJB, and CJM supervised the project. IE initialised model development. HLC continued model development, performed the research, and analysed the data. CJM prepared sequences for experimental validation. PD performed minipreps. GM performed the transient gene expression. MB and BC performed purification, SEC, and Quality Control. CB performed the ELISA and ST the BLI analysis. AGW developed the web tool. The manuscript was written by HLC and reviewed by RJB and CMD.

COMPETING INTERESTS
RJB is a director and shareholder in Fusion Antibodies plc. CJM, GM, MB, BC, CB, ST, and PD were employed by Fusion Antibodies plc. CMD discloses membership of the Scientific Advisory Board of Fusion Antibodies plc and AI proteins, and is a founder of DaltonTx. All other authors declare no conflict of interest.

Figure Captions

Caption Figure 1 Number of mutations from the germline V-gene sequence per light sequence complementarity determining regions (CDRs) and framework regions (FR). The native (grey) and LICHEN's generated (red) light sequences of the 500 samples of the standard_ds were compared against the closest germline V-gene according to germline annotations by ANARCI[19]. Generated sequences by LICHEN contain in general more mutations away from the germline sequence.

Caption Figure 2 The average germline identity scores of the V-gene and J-gene derived from ANARCI[19] calculated for both chains of the native antibodies (A), random paired native antibodies (B), and antibodies with generated light sequences by LICHEN (C) on the 500 samples of the standard_ds. A linear regression line, determined by SciPy[36], is shown as well as the corresponding Pearson correlation coefficient (R) and p-value (P). LICHEN captured the co-evolutionary relationship between the heavy and light chain during affinity maturation.

Caption Figure 3 The normalised difference between the probability of native over germline reverted pairing. The conditional probability of

generating the native light sequence and germline-reverted light sequence were normalised by the unconditional probability. The difference was normalised by the sum of the probabilities. Results are shown for the complete conditioning_ds (1000 samples) and on subset containing either five to ten (607 samples) or ten to fifteenth (393 samples) mutations. Values above zero (dotted line) indicates the native pairing is preferred over the germline-reverted pairing. LICHEN in general prefers native pairing over germline-reverted pairings.

Caption Figure 4 Percentage change in potential yield from therapeutic and binding data of generated light sequences by LICHEN for the VH of therapeutics adalimumab (A and B) and pembrolizumab (C and D). The therapeutics were used as positive controls. The germline V-gene sequences IGKV1-39*01, IGKV1D-33*01, and IGKV7-3*01 for adalimumab and IGKV3-20*01, IGKV3-11*01, and IGKV7-3*01 for pembrolizumab, labeld "Germline-1", "Germline-2", and "Germline-3" respectively, combined with IGKJ1*01 were used as baselines. All three light CDRs (Kabat definition) were grafted in the germlines labeled with the suffix "+" (dark grey). Three use cases where tested: pairing_CDRs ("LICHEN-1" to "LICHEN-5", navy), pairing_eCDR ("LICHEN-6" to "LICHEN-13", blue), and pairing ("LICHEN-14" to "LICHEN-23", light blue). For pairing_CDRs LICHEN was conditioned to all light therapeutic CDRs, for pairing_eCDR LICHEN was conditioned to essential light CDRs only (CDRL3 for adalimumab, CDRL1 and CDRL3 for pembrolizumab). LICHEN was restricted to make IGKV1, IGKV2, and IGKV3 light sequences and diverse sequences were tested. Binding was tested to the targets of adalimumab (TNF-α) and pembrolizumab (PD-1) ("Target", pink and orange) and the unrelated non-human protein RBD was uses as a control ("Control", light pink and light orange). Phosphate buffered saline ("PBS", black) was used as secondary control. LICHEN generates well expressing antibodies which maintain binding ability to the target when additionally conditioned on the CDRs.

References

1. Raybould, M. I., Turnbull, O. M., Suter, A., Guloglu, B. & Deane, C. M. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *Communications Biology* **7**, 62 (2024).

2. Chiu, M. L., Goulet, D. R., Teplyakov, A. & Gilliland, G. L. Antibody structure and function: the basis for engineering therapeutics. *Antibodies* **8**, 55 (2019).

3. Janeway, C., Travers, P., Walport, M., Shlomchik, M. J. *et al. Immunobiology: the immune system in health and disease*, vol. 2 (Garland Pub. New York, NY, USA, 2001).

4. Lu, R.-M. *et al.* Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science* **27**, 1–30 (2020).

5. Shuai, R. W., Ruffolo, J. A. & Gray, J. J. Iglm: Infilling language modeling for antibody sequence design. *Cell Systems* **14**, 979–989 (2023).

6. Turnbull, O. M., Oglic, D., Croasdale-Wood, R. & Deane, C. M. p-iggen: a paired antibody generative language model. *Bioinformatics* **40**, btae659 (2024).

7. Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

8. Rajagopal, N. *et al.* Deep learning-based design and experimental validation of a medicine-like human antibody library. *Briefings in Bioinformatics* **26**, bbaf023 (2025).

9. Greenshields-Watson, A. *et al.* Anarcii: A generalised language model for antigen receptor numbering. *bioRxiv* 2025–04 (2025).

10. Tsuchiya, Y. & Mizuguchi, K. The diversity of h 3 loops determines the antigen-binding tendencies of antibody cdr loops. *Protein Science* **25**, 815–825 (2016).

11. 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 Science* **31**, 141–146 (2022).

12. Jaffe, D. B. *et al.* Functional antibodies exhibit light chain coherence. *Nature* **611**, 352–357 (2022).

13. Warszawski, S. *et al.* Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLoS computational biology* **15**, e1007207 (2019).

14. Fernández-Quintero, M. L. *et al.* Germline-dependent antibody paratope states and pairing specific vh-vl interface dynamics. *Frontiers in Immunology* **12**, 675655 (2021).

15. Seidler, C. A. *et al.* Data-driven analyses of human antibody variable domain germlines: pairings, sequences and structural features. In *mAbs*, vol. 17, 2507950 (Taylor & Francis, 2025).

16. Lefranc, M.-P. *et al.* Imgt, the international immunogenetics information system. *Nucleic Acids Research* **37** (2009).

17. Abhinandan, K. & Martin, A. C. Analysis and improvements to kabat and structurally correct numbering of antibody variable domains. *Molecular immunology* **45**, 3832–3839 (2008).

18. Olsen, T. H., Moal, I. H. & Deane, C. M. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances* **2**, vbac046 (2022).

19. Dunbar, J. & Deane, C. M. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2016).

20. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

21. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

22. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256 (JMLR Workshop and Conference Proceedings, 2010).

23. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915–10919 (1992).

24. Chinery, L., Jeliazkov, J. R. & Deane, C. M. Humatch-fast, gene-specific joint humanisation of antibody heavy and light chains. *MAbs* **16**, 2434121 (2024).

25. Olsen, T. H., Moal, I. H. & Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics* **40**, btae618 (2024).

26. Marks, C., Hummer, A. M., Chin, M. & Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics* **37**, 4041–4047 (2021).

27. Wong, W. K. *et al.* Scalop: sequence-based antibody canonical loop structure annotation. *Bioinformatics* **35**, 1774–1776 (2019).

28. Abanades, B. *et al.* Immunebuilder: Deep-learning models for predicting the structures of immune proteins. *Communications Biology* **6**, 575 (2023).

29. Olsen, T. H., Abanades, B., Moal, I. H. & Deane, C. M. Ka-search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports* **13**, 11612 (2023).

30. Dunbar, J., Fuchs, A., Shi, J. & Deane, C. M. Abangle: characterising the vh–vl orientation in antibodies. *Protein Engineering, Design & Selection* **26**, 611–620 (2013).

31. Mease, P. J. Adalimumab in the treatment of arthritis. *Therapeutics and clinical risk management* **3**, 133–148 (2007).

32. Kwok, G., Yau, T. C., Chiu, J. W., Tse, E. & Kwong, Y.-L. Pembrolizumab (keytruda). *Human vaccines & immunotherapeutics* **12**, 2777–2789 (2016).

33. Akbar, R. *et al.* A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports* **34** (2021).

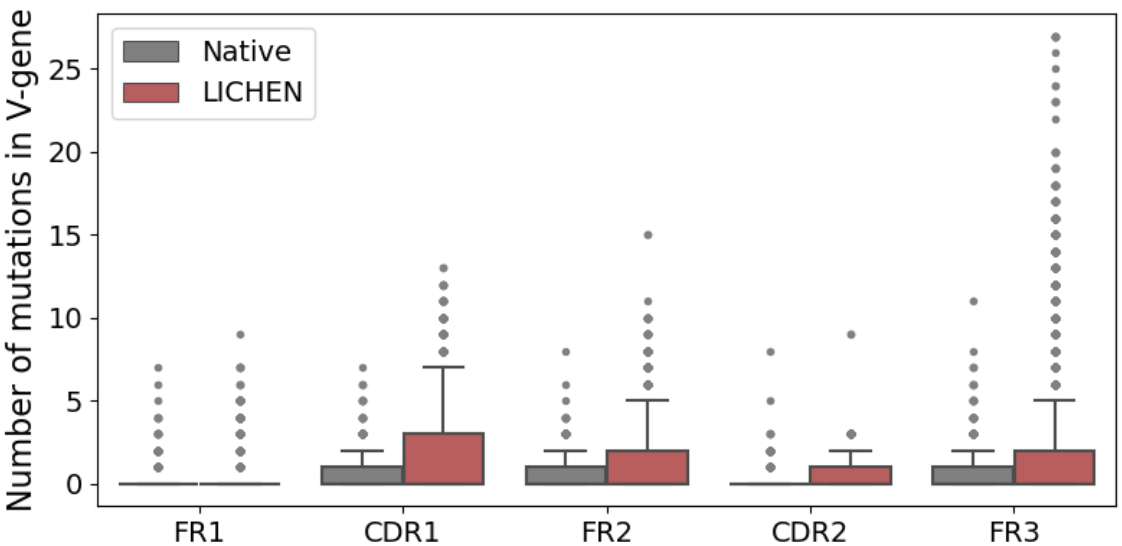
34. Raybould, M. I. *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 4025–4030 (2019).

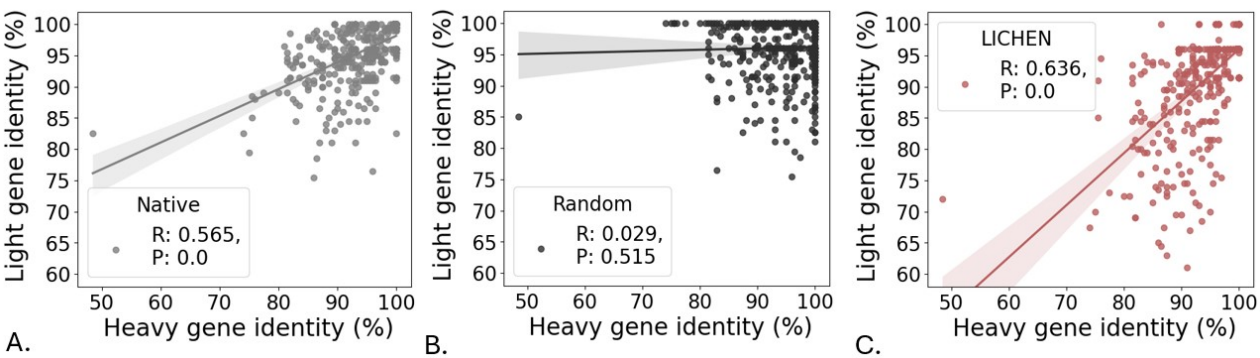
35. Jayaram, N., Bhowmick, P. & Martin, A. C. Germline vh/vl pairing in antibodies. *Protein Engineering, Design & Selection* **25**, 523–530 (2012).

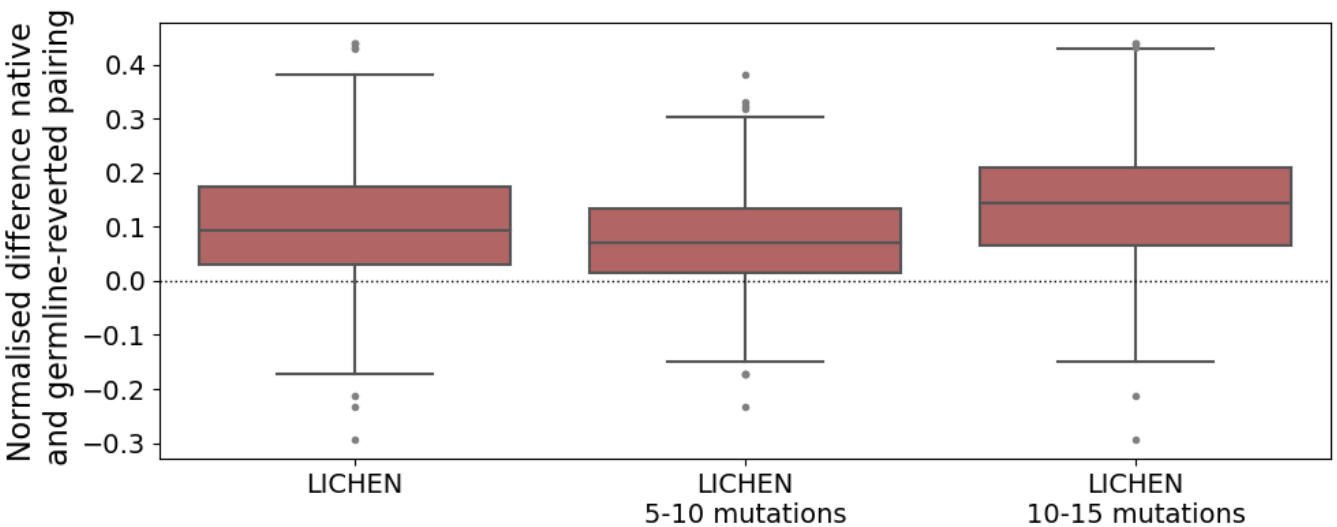
36. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).

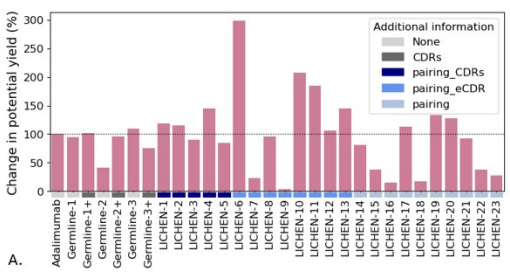
37. Capel, H., Greenshields-Watson, A. & Deane, C. Lichen: Light-chain immunoglobulin sequence generation conditioned on the heavy chain and experimental needs (dataset and model weights) (2025). URL https://doi.org/10.5281/zenodo.15917096.

38. Capel, H., Greenshields-Watson, A. & Deane, C. oxpig/lichen: Light-chain immunoglobulin sequence generation conditioned on the heavy chain and experimental needs (source code) (2026). URL https://doi.org/10.5281/zenodo.18459226.

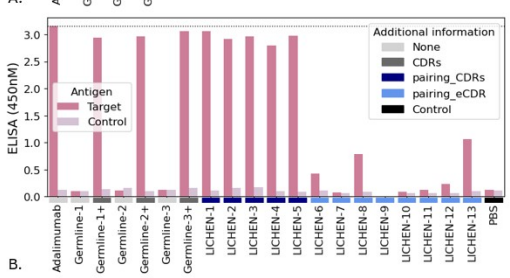

Native
LICHEN
Number of mutations in V-gene
0
5
10
15
20
25
FR1
CDR1
FR2
CDR2
FR3



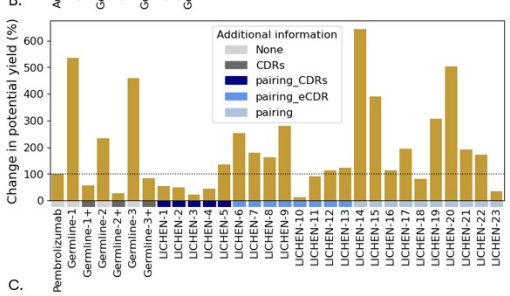

Normalised difference native and germline-reverted pairing
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
LICHEN
LICHEN 5-10 mutations
LICHEN 10-15 mutations



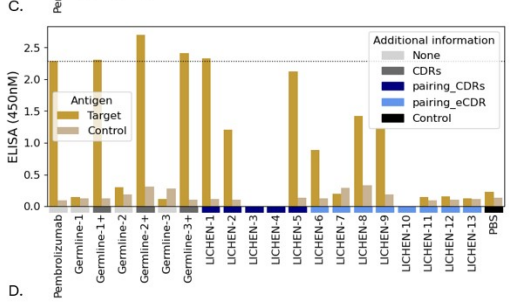
A.



B.



C.



D.