AI-Based Antibody Design Targeting Recent H5N1 Avian Influenza Strains

Nicholas Santolla[1] **and Colby T. Ford**[1,2,3,4,✉]

[1]University of North Carolina at Charlotte, School of Data Science, Charlotte, NC, USA
[2]University of North Carolina at Charlotte, Center for Computational Intelligence to Predict Health and Environmental Risks (CIPHER), Charlotte, NC, USA
[3]University of North Carolina at Charlotte, Department of Bioinformatics and Genomics, Charlotte, NC, USA
[4]Tuple LLC, Charlotte, NC, USA

In 2025 alone, H5N1 avian influenza is responsible for thousands of infections across various animal species, including avian and mammalian livestock such as chickens and cows, and poses a threat to human health due to avian-to-mammalian transmission. There have been 70 human cases of H5N1 influenza in the United States since April 2024 and, as shown in recent studies, our current antibody defenses are waning. Thus, it is imperative to discover new therapeutics in the fight against more recent strains of the virus.

In this study, we present the *Frankies* framework for automated antibody diffusion and assessment. This pipeline was used to automate the generation of 30 novel anti-HA1 Fv antibody fragment sequences, fold them into 3-dimensional structures, and then dock against a recent H5N1 HA1 antigen structure for binding evaluation. Here we show the utility of artificial intelligence in the discovery of novel antibodies against specific H5N1 strains of interest, which bind similarly to known therapeutic and elicited antibodies.

Avian Influenza | Antibodies | Protein Modeling | Docking | AI
Correspondence: ***colby.ford@charlotte.edu***

Introduction

Highly pathogenic avian influenza A (H5N1) remains a persistent global health threat due to its zoonotic potential and historically high mortality rate in humans. Despite sporadic outbreaks and ongoing surveillance, the continual antigenic evolution of the hemagglutinin (HA) glycoprotein, particularly within the HA1 subunit, presents significant challenges to the development of broadly neutralizing therapeutics. The HA1 domain is critical for receptor binding and immune recognition, making it a prime target for antibody-based intervention (1).

To address the urgent need for rapid, scalable design of high-affinity antibodies against emerging H5N1 variants, we introduce *Frankies*, a computational pipeline that integrates state-of-the-art generative artificial intelligence (AI) modeling, protein structure prediction, and docking simulations. This pipeline begins by generating novel Fv antibody fragment sequences via conditional protein diffusion using EvoDiff, leveraging a curated set of reference sequences from the Therapeutic Structural Antibody Database (Thera-SAbDab). The resulting sequences were folded into 3-dimensional structures using structure prediction models such as AlphaFold3 or ESMFold v3, which have demonstrated near-experimental accuracy in modeling antibody-antigen interfaces. Finally, we employ HADDOCK3 to perform rigid and flexible docking of the modeled Fv regions to the HA1 target, enabling the assessment of binding orientation, stability, and epitope accessibility.

By systematically combining data-guided AI generation with physics-based modeling, *Frankies* represents a promising step toward automated antibody discovery and optimization. In this study, we apply the *Frankies* pipeline to a recent H5N1 HA1 antigen and characterize a set of *de novo* antibody candidates with favorable structural and biophysical properties. This work lays the foundation for future *in vitro* empirical validation and paves the way for accelerated response to future influenza outbreaks.

Results

Pipeline Performance. The protein generation pipeline successfully produced 30 novel anti-HA1 Fv antibody fragment sequences, with a 100% success rate in AbNumber sequence validation. In our systematic tests, the reliability of Fv antibody sequence generation was consistently greater than 90%. The complete set of 30 experimental runs took 11 hours and 34 minutes to complete in serial, averaging 21.13 minutes per run. In a distributed computing context, this walltime would scale linearly with $\sim$20 minutes per pipeline run. Individual component processing times were distributed as follows: approximately 5 minutes for EvoDiff operations, less than 1 minute for structure predictions with ESM3, 16 minutes for HADDOCK3 computations, and less than 1 minute for the output report generation.

Generated Fv Sequences. The 30 diffused Fv heavy and light chain sequences are listed in Table 1. Note that all diffused sequences are a defined length, with heavy and light chains having 132 and 125 amino acids, respectively. This is controlled by the diffusion step, which defines a maximum sequence length to be conditionally generated. The top 10 candidates, by lowest HADDOCK score, are shown in Figure 1. Note the epitope diversity given the random surface residue selection for unbiased binding site selection.

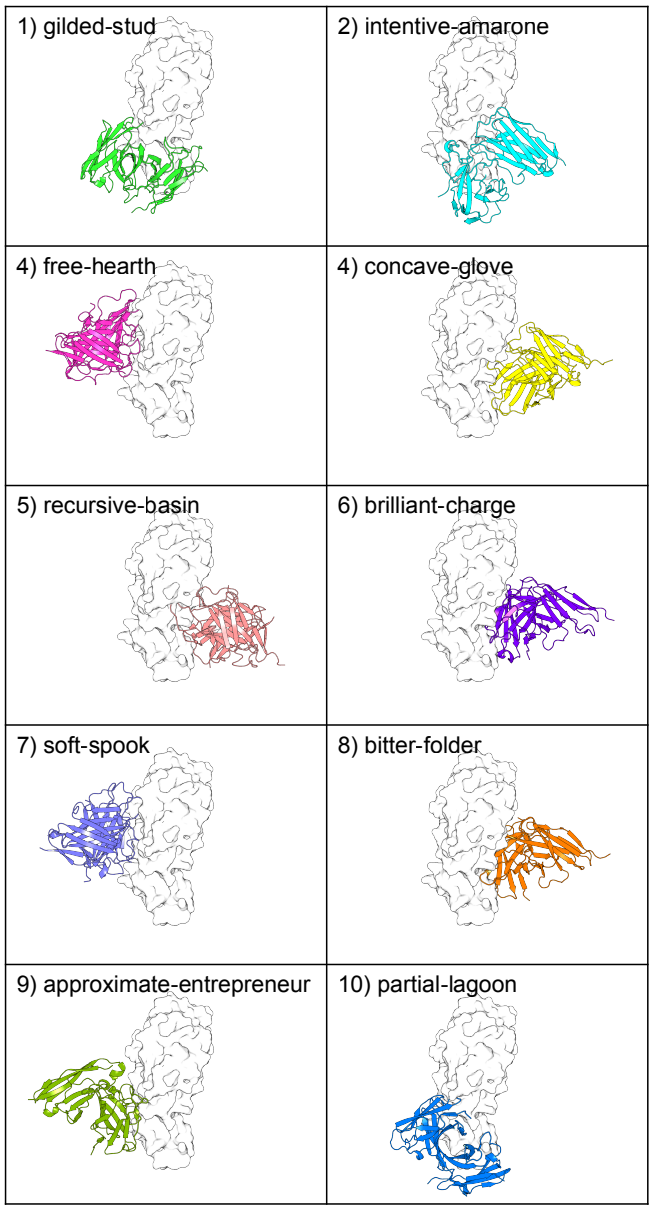
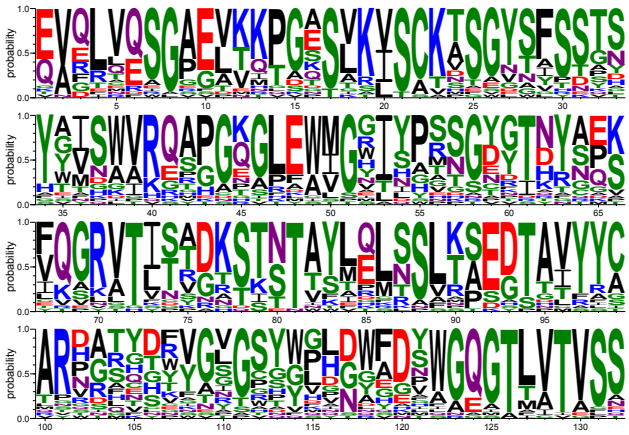
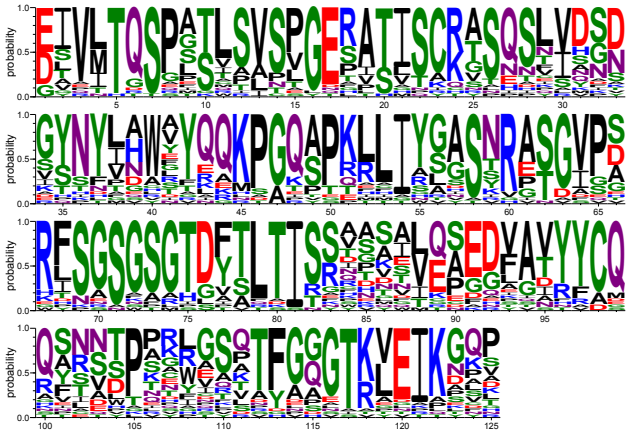


Fig. 1. Top 10 diffused Fv candidates by HADDOCK score. HA1 is shown as a white surface model and the diffused Fv structure is shown in a colorized cartoon style.

Also, note the sequence diversity depicted in the logos in Figures 2a and 2b. While natural heavy chain sequences often start with EVQ or QVQ, there is additional sequence diversity in position 3 with the introduction of arginine (R), phenylalanine (F), glycine (G), and aspartate (D) amino acids, which are atypical here. However, the conditional sampling shows a consistent selection of glycine at the 6th position and VTSS at the end of the VH sequence, which are very common in natural antibodies. For the light chains, the common starting sequences of EIV and DIV were seen in the diffused sequences along with expected sampling of methionine (M) or leucine (L) at position 4, and EIK toward the end of the VL sequence (positions 120-122). Note that all of these sequences were successfully numbered using the Chothia numbering scheme.



(a) Heavy chain probabilities



(b) Light chain probabilities

Fig. 2. Position-level amino acid probabilities of the diffused Fv chain sequences. Created with WebLogo (2).

Sequence ID	Heavy Chain	Light Chain
kinetic-template	EVQLARSGAEPTMPGETVKLSCKTSGYNASDTSFYIGAARQTGGKGLEWMGHISPTNGNPIYYSEK IQARLTLTADTTTETTYIQLLAFKSEDSAMFYAARHRTGHYYGYGSYWPLNGGDIWGGGTLVTVSA	ESVLTQSPGSLIISVGERATISCKASQDLVNDTGHSFPHWRYLGKPGTAPKLLGYGASNRAS GAPGRFNGSGSGTDFSLTISRTASELKPKDVATYYCQQYNATPPTYRSQTYGGGTRAEIKSQP
partial-lagoon	QVGLVQSGPEVKTPGESVKISCTAGGYSFSSGLYWIDWVRERHGQGLEWMGMIHPSTSENTKYNPS FQSRVTISVNNSTNTAKEELSSLKAEDTATYLAARSAAVDVYTRGAYGKADRFEAWGQGALATVSS	ELVLTQSPLSTSVSPGERATLSCRASQSLVYGNGYNNLAAVTRQMPAEATRLLISGGSTRAT GVGSRLSGSGSGTDYTLTINSQTKQLQSEDFAAYYAMRYNNWPTRLIKQTFGGGTKLEIKDFP
glowing-avocet	QARLVESGAEVTKTGQSTKVSAKMSANTYSSADYTVSAVHQYPGKALFWIGHIYYPNGYYRDNAGS FQGRVTISTDASKSTTYLTLSSLKSEATAIYRCQYRRRYYKVGVSSYVPLDWYDNLGQGTLVTVSS	DVVMTQSPESLAVNPGERISISCKGSQSLIASDGQNYVHWRYKAKPGQSPKILAYSASNRAT GVPARISGSGPGTGFTLAISSISNAIQAEGLTTNRCQRELETAIVRSFRPFAGGTVLEIKALK
approximate-entrepreneur	EVQLLQSGAGAKKPGATVTISCVVSGYSYSSYYSGIDKVRQRPGHGKDCVGGIYPRSGYYTHYTEK FQGRVTYPSGKQTNNAYLQLNSVTTEDTAVYYCARPGLESFYGVGVNWSHNNSMSIGQGTLVTVSE	GIRITQSRASLSVSAGEPASIQARVSQSLIKGDISNYLEYYYQRKPGQPPKLFIYSASTRAT GVPARLPGSGSGTDFTLTISSRDGFVGSEDFGIYYCQRVENTPPQLGQPTFDQGTKLEIKGYK
recursive-basin	EVQLVEIGPEDKKPGTTLKLSCVPSGVSFHSTNYAVSGVKQSPGQGPEAMGNTYPSSGCDTDYCQK FQVRATITTRRSTSTVYLELNSLKPDDPAVYYCARVSHNTEKTPGSYFHPGDEELWGQGTLVTALS	DIVLTDSPSGLSASTGERPTISKRVSNSLSDFDRSTTLAWFFQARPGRSEKALIRAASSRAS GVPVRFSGSGSMTSFAFTISSAAAALEAEHPATYYCQQSRDDPAARASAAFGGGTRVEIAPLA
free-hearth	EVQEVDEGGEVVKPGKSARISCKNSGYGFSSQSYAVSWIREAEGKGLAAMTVISPTGAAGIHRNEK VQGRVTISKDHSSNTVYLEMNSLKSEGTAIYAAARNGQYNFVVVGSYWGYGLFDSWGEGALCTVSS	ETVLTQSPFTLSVSPGEPITVSCRSPQNLVDSNVYNNINWEYQKKPGQAPPRLIYVNSNGAS GAPDELKGSGSGTDFVLTIRRCPKEIKDEEVVVYYCQKSIVSPVEWVVATFGQNTRVEIKKKV
smoky-latitude	EVRLVQSGREVVKVGESLKISCKASEYNFSSTNYYTGWVRKPPGQVRMWISYIYHTTTEGTNYSAV VKPYAGVCYGKSINTVTLHMNSVKASDTAVYYCANTIRYDVYSVSPTWDHDWVDSWGQGTLVTVPQ	DSQLTQSQSSWSLSVGDFVHGTCKTSQSLSQPDVYSYTHWSGRRHPAQTPKLLIYLASTRGS GVGTHISGSGSGTDYTLTISSFGSTTRAGGFATRFCQQSRADTTRAGRQTDGSGTKVGTKGSR
internal-rundown	QAGLVQSGAELKKTGSSLRVSCKSSWYTYSSSYYAIHIIREAPSKGLEWVSRINSRHGYYTTYAPS IQGRVTFSTDKSTSTIYMPLSSLSSEDTAVYFCAPHASEHYVGSGTSWLHDWAESSGQGTTVTASN	DIVITQSPLTCSVSLGETTSVSRRTSQNLIDNNKYHYFAWLYQQKPGQAPTLLIYAGSYRAS AVSDRPSRSGSGIDYTLTISSKSTIVECDGVAVRYCQRSNSTPTRLAALDFAAGTKPEIKNPS
critical-tin	QVFLVQSGAELTNPGASVKVSCKTSGYSFSTTSYGMSWIRRAGGQGLEAIGWISHRSGYRTNASPK FQGRVTINTGASTSTVYTQLRSLKPEDTTVYYSARDGAHSFVAAGSAWGLDWGGYAGEGTIVTVSA	DIVITTSPMTRSVSVGEAASISCCRSQSCIDGNGYNYMNWIYRQKSGQAPRELIYGASKVAT GFPARFSASGSGTDFTLTISYTAVNVEPGGVGSYYCQRARSTPSKRAYQTFGAGTRVEIKLAN
gilded-stud	VAQTLQSGPELMQPGASVKISCTDSGYTISSTSYAFSWARGSGGKSLEWVGWIHWHTGVGTQYADS FQGRATSDRDKSKNTASAQFNSARSEHSGVAYGASDRTTTTYGLGRPVVVGWADSWNQGTLATASS	DLVLTQSPASVSVTPGTSASISCQSSQSQVDSNDYNYANRAYQQMPGQAPTLMIRSASYRPS GVPSRISSSGSGTSASLTISRVNANLQEENEANYRCQTSSVSGNRICSPVFNSGTKLEIKGPP
concave-glove	EVQLQESGAGLVKTSESGSISCTTSGGGGPSGGYWMSWGRQGPGGGLEWSGRIYGVSGDGTNGRGS LKERGTLSPDTSTNTASLGMSSVTASDTALYYGARGAMGGPVGGGSYGGLNGGDGQGQGTLVTVSS	DYVLTGSPASASVSPGESPTISCRATQTFVDGDGTKYVAWAAQAKPGQAPKLLISLDSNRPT GVPSRFSGSGSGTDYSLTITGAKNTLQNEDVADYYCQQVRSSPPARGSPSYAALTKLDAKNPS
boolean-burbot	EVELCESGAEVEKPGSSVKVTCKVTGYAFSSTSYAISKVVQAGNPSLASIGELSPSSGDYTRYNEK VTAKVTLTADKSTNTTYLELTPLTSEGTAIYICTRRARYDRSGVGSDYVGDWQDPAGQGTLATVSS	ETVLTQSPGTVTVSPGERATMSAKVSISTRMSVSTNYLNWAYEQKPGQAPRLLIHGASNRAS GVSARLSGSGSGSLFSRTISSNEDEVEAFQLAIYYCDQNTSDPECLPRDTYGGGTKLEIKEVP
magenta-food	EAQLQESGAELNKTGASAKVSCTHSGYSLSDTSYYINGAKQAPDKGPFALGGLYASSRYGDDTAQS TKSLVHVTRDRTKNTTSLELSSLKAEGTGIYYALGWGSYGKAGLGCSGLDGYFAYWAQSTLVTASS	EIVLTRSPATVTVSPLQRATVSCRNSNSNVDSDGYSYLHWYYQQKPGQAPKLAIGSASNRVS GVPSRFSGSGSLTDYALTISSDAAALQAGDAADYACGAAANDTPGRGSATFGPGTRVTIKGQL
exponential-cymbal	EVELVQSGPETVKPDKSVKVSCKTEAYSFSTPSHYVSAARSSTGQGLEWMPGIYASSGYKTDYAEP VQSRVTKTVDKTTTTAYTELSSLTAKDTAVYYIARDGTYDRYAGGHYGHHNWEDYWGQATMATVSH	EGVMTQSPATLRLSEGERVTISCTYSQNNISINSYNDIGWTYIQKPGQPPETLIYLSSVRAT GIQDHFSGSGARTDYALTITRATAAMQPEDLAVYYCQQSNEDPPNGGPTTFGGGSRVEIKGQP
soft-spook	QAERVRSGEELKQPADSVKISCKTSGNTFSSSHGEMNWVKHAPCQGREWLGYTLARSGYGTHYSPK FVGRTTITAGKTSSTTKMQLSSLMSEGSAVYRCARVSTTSCNGLPSYYPHGGADVWGQGTTVTVSS	ESVITQSPSSQPASPGELLTISCQASPINIVNKSYNHIACEYQQMPGQVSKLLTAGASIRPS VVPSRHSGSSSGTLYTLTISFIASILCSEDFAVYVCQNFCSLKACWGSVAGGGETKVEIKGQL
creative-halftone	EAQLVFSGAELTQPGNSLAISAKSSEDSIYSVNYVVSWVREAPGQGHLIMGGIHPVPNTGTKYGQV FQGRVTITADNSTNTAYVKSTSFPSDDTAVYYCTRHTFCGGVNLGSGYLQTAFDYWGQGTAVIVSS	GIVMTQSPATLSASPGETATISCKGSNSISDNAGPNYLAWVYQQKPAPPPKLLIYSASNRAN GDPEGFSGSGSAPGVSLTSSSVPKIVEEGDAAARYCQQTNVVPAKWETKTFVPGIKLEIAGQG
brilliant-charge	TVQLRQSGSEAKRPVESLKVSAKASSVSFSSGAYYASDIRQAPGNTLEWMGAANAANSNDTAYNQS FQGRVTINRDKTITTAYLQLNNLTAEDTDCFYCATDASCDFITNGPYYFNDWADTWGQGTMVVVFS	EIVTTQSGSTMSVPLGEHATISCRGSESPVSSYESNLGAWSYQQKPAKAPKRLIYRYSNRPS GVPSRFSGSFSGTDVTLDISGWGSSLQSEDVAIRYAQQFSNLPSTFGLTTFGQGTKVVIKDCS
strong-bear	EVQRQQSGAEVTKPGGSLKVSCKTSGYTFSSTTAAVSWVKQPHGTGLEWTGWLYHESGDGTNYAES VRGRVTVSYGKSTSTASLQMSSLRSEGTHVYYSARPGTGDWWGVGWGWGGNWFDSWAQGTTVTGSS	EAPLTQSGLSLSASSGNRATHTCRTSQAKVQNSIYIYIHWGYQQKPAKSPQLLIYGASSRGT GVPSRFSGSGSGTDYTLTISSNSLTLQPEDYATYFCQQSNVSPNNYESQTHEQGTKEEIQDQT
minty-cylinder	QVGFVEWAGGVKIPSASKKLSCKASVGSVSSTNYGISAVRQAGAEGLKAVGWISGMGGTYTDYSES LKGVVTISAAKSTTTTFIELSSLRPSSTTVRYCAPPASQDRVGSGSPGGPGWFKPWGEGTLITVSS	DLEMTQSPLSLAVSLGESISIPCRTSQSLVDSDKYNFPDLLYAQKPGISPRLLIYTGSSRAT GSPDRISASGSGTDFTLTITKQGDGVEAERIATYYCQQPRNTPRRINSQAFAQSTKLEKKAKA
bitter-folder	EVRLMESGAVVKQPGQSLKVSAKDSGYAFRNTSYSISWPRGAPGQGLEWMGYIYPNSGDGTNRSQS VQGLVTISTNKSISTASLQLSSLKAEDTPVYATARHDGYHWFAYTCHWMHGAADHWGQITLVTESS	ETVLTQSAATLSVTPGEGASLSCKALQSLVHNNGYNFIAAFYNQKPGQSPKRLIRGGANVGS GIPSRYNASGSGTDTSQTITSDHSALQSEGVQVYYCEQYTTTPKSPTSKTFPGGTKVEILPQV
contemporary-wine	KVERTQRGAEVKKPDKSLKISCAASGYSASDTSHYINWVQQAPGKALEWIGIIYPSSGDRTKYAEA FQGRVTITRDGSKNTAYARCNSVTPEGTAVRYCARHGSQTRFAIGSYWPVDQEGFWRQGTFVTVCS	GIVLTQSPPSLSAPVGESATASARGSQSEVDADGYNYLQADYQQKPGQAGQLLIYGVSNRES DVPARLSNAGAGTGYTTTISSAAVWIQSADFGVYFCQQANNTPSGRVSTRFAGGAATLPKGKT
quadratic-format	EVDTTQSLASAKMLGESVRISCKASGYTFTKPYYTYQWVKQTKAEILYWVGVTDPANSDVINYQPK EQGRVTLGVRKSTSTNWMRRRSLRSEDTNVYYCRRVRTYHYVNNGGGWVDNWFHNFGEGTMVTVSS	EIVLTQSPASIALSTGERATISCRANHPFIHSDGSNYLDWVRQQKPGQSPTRHIYGASYHET DIPDWFSGSGTGTDFTLTIRRSTSVVEAEDTGVYYCQQFSVSPPDWEASNYGDGTRVEIPGVH
avocado-bumper	EAERVESGAEVKKPGASTKISCKAAGYSFSSTSYWMHWVRQMPGEGLEWMGRIYPSKSCGSNRSMK CQGRVTLSTDTSTNTASLQLRSLTPSDTATYRAARQAFHGWVGIGSTWPDDWADVWAQMTLVTVSS	EIVNTQSPGTLSVSPGERATITCKASESTIAGNSYPYIGVNYLKKPGQAPKFLIYSASNRIS GIPSKFNESWSGHDFALTISNPPQIIQSFDFADYYCQHINSSPPRYQSLTFGAETKVEIKTQP
cold-electricity	EVFLLQAGPFLTHTGSSLKVTCKNSGNSFTTGSYTIKAVRQSGGTATFWIGSIIPSNGYGTNTAKT IKGRATISADTSTNTAYMELSSLASEGSALYSCARDAQNSWVGRGWYYGLNGFGMAGQGTTVTVSS	ESVMTGSEASLSVCPGESATISCRTTQSLIYSDGTNYLHWTTQQKTGQSPKLCIYSHSKRAS GVSGRTSGSGFRTDATLTISSHSYSTTAEDVSTYYDQQALNPPAHHGSSTYGQGTRLEIKNAP
flat-gutter	EVDLNQSGAETKITGQSIKVSCKTSGVSFPEADYATPLTRQHHGKALEWMGNTNYGTGYTTNYGPK IQVRVTLNSGKSTSTAYLPKKSLKAEYTTIYYCVRDGHQTNVESTGQGQIGYFNYWGEGTLVTVSN	GIAMTGSGSTISVSPGERATISCRASQSTVDKSVSNYVHWVFQQLPATSTKRIINGSSNRES DVPSRTSGSKSGHDPTLTISRRSSDLEPEDVAVYYCQSYTSTPSELVSQTYGQATKAEITGQD
symmetric-pad	QAQLVQSGAGVTKGAASVKLSCKTSGYSISSYSYGVSAVRQAPGQGPEWVGGISPMSGPYTHYAQS VQARLTLTVDKSTSTAYLELTASNPEDTATYYAARNARGTRVGVGPHYLLDWHDYWGAGTLVTVST	DIVMHQSPTTLSVNVYEPATISCKTSNTLANGDGSNYVVWYYQQKAGQSPKRLIAGISTRAT GVEHKFSGNASGTDITLTISSTHTAVEPEDFAVHYDQQYRNWLKKLISPTFGGGTKGNRPSKV
antique-structure	QAQLEQSGVEVVKPGSSVKVSSKTSGYWASTTSHWISWVRRSPAKGFEWMGGIQPGSGNYTNFNEK YQGRATITAGKSSNTAYTQLTSLTAEGTTTYYCARNNTHDTYGSGSSYPLDYFDVWGQATTITVSS	EVVLTQSPGTTSLPPGERATLSIHASHHLVDSDGSTYVSWVYQEKSGQATRRTIYGASNRAS GIVGRFSGSGSATGYTLTIRRADVSVESEQSAVFFAQQFSSTPQKWGSVTFGHVTRLEIKGSP
cream-callback	GNFLVESGAGATKPAPSLSVSCKVSGESFSSGSYGISWARQAGGPGLEWMGGIIPSSGEFINRGPS FQGKATITAGRSTTTAFFELSSLTSEDTAVYYCMRPRRFDFYGLTSYQPLGWHGYWGQGTLATVSS	DAVMTQSPPTLPVSVGESASISKKAAESVVSSDAYNYLNWAYEENPGQSPEMLIWAGTNRES GIPDRFSGSGSGTGFTLSISRVRSATEAGAVAVTYAMGSIAHPKPWGTKTFGQGTKVEIKGQD
inventive-amarone	QAQIVQSGPGLVKTGTSVKVSAKTTGYNFSNKNYIVSWVREVPGRGLEAMGRIYGRDGDYTDRAEK VVGKVTISTDKDKNTWYLQMSSLKAEDTAVSYAARNDLVCYGGGGRYGLHNAYDNAGQGTLVTVSS	DIVTTQTSGKLSISLGERVTINYKTSQSYVDGSGYNYTHHAYEQKDGKYPKLLIYGGSNRES GVPDRDSGSNAGTDVTLTISEVVMVVQSDDKINRYCSQSTDYTLYLDAVTFLQGTTYEIKYNP
messy-discriminator	EVRLVQAGPEVKQQKESAKLSCKTFGLSVSSTHYGNNWAHGAPGNGPEAIGHILPMNGYGIHYCPK VQGNSTISTDKTTSTAYMDLSSATSEDTAIYYCTVPATKLTYGTACGWGLSYFDPWAQGTLATVSS	EIVITQSPITLPVSPGEPASITCRASQSVLHSDGYNYLDWGVKQKPGQAPQHLIALASRRAS GVGARFSGSGSGHDFTLKIRAYNAIVQSEGVGVYYCQAANQTPQGFGQQTFGGGTKLEIKNDP

Table 1. The 30 Fv heavy and light chain sequences diffused and evaluated in this study.

Comparison to Existing Antibodies. The *Frankies* pipeline consistently generated Fv sequences and structures that bound well to the HA1 epitope. For comparison purposes, as shown in Figure 3, the binding performance of the 11 antibodies used in Ford et al. 2025 that were bound to H5N1 isolate EPI3009174 were selected as a reference. Then, these were compared to the 30 Fv structures generated in this study, which were docked against the same isolate antigen structure.

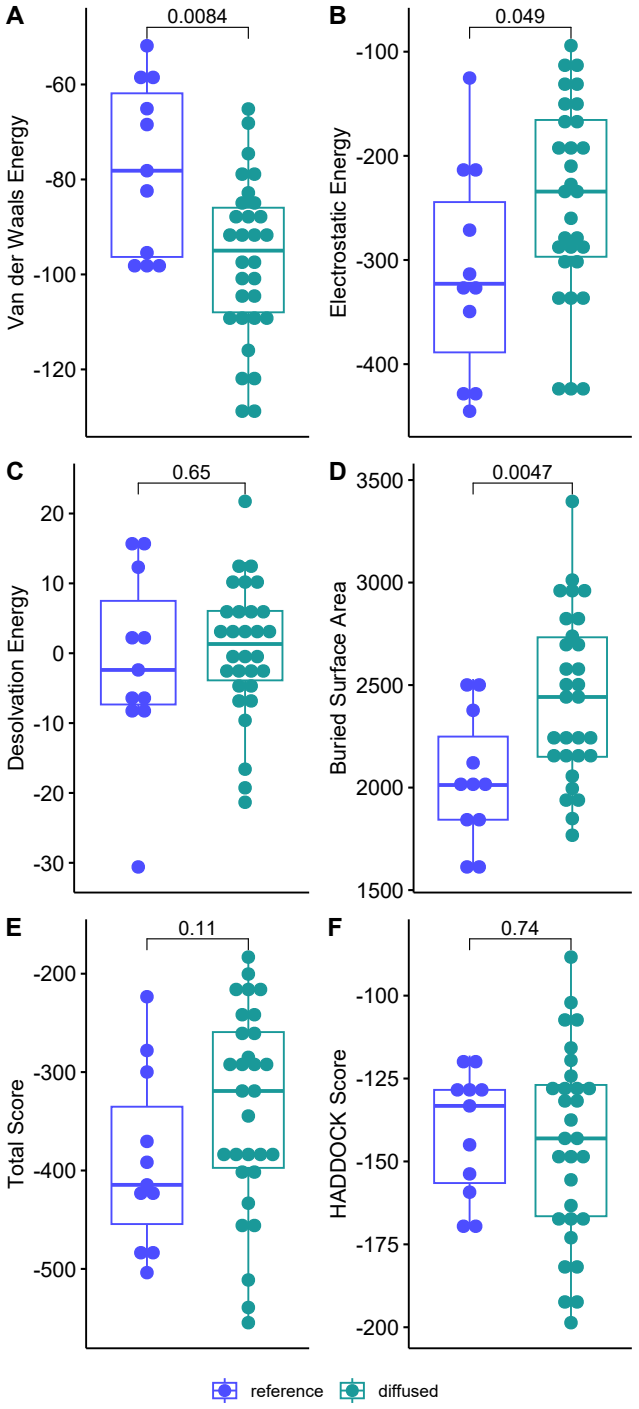


Fig. 3. Comparison of the docking metrics between reference and diffused antibodies. Pairwise comparisons are shown as p-values from the Wilcoxon signed-rank test. For all metrics except buried surface area and desolvation energy, lower is likely indicative of better binding.

The Van der Waals energy is significantly better (more negative, statistically) in the diffused antibodies. In contrast, the reference antibody set displayed better electrostatic energy and buried surface area, though there are multiple examples of diffused antibodies that are within the same range. However, the desolvation energy, total score (electrostatic + Van der Waals energies), and HADDOCK score were quite similar overall and showed no statistical difference according to the Wilcoxon signed-rank test at the $\alpha = 0.05$ level.

Thus, these metrics show that this pipeline was able to consistently generate antibody sequences with similar performance to therapeutically-derived or elicited antibodies.

Methods

The *Frankies* pipelines was designed as an automated workflow for producing antibody candidates and predicting their binding affinity at-scale. This pipeline is written in Snakemake (4), which provides a reproducible framework for running all of the steps mentioned below.

Inside the Snakemake pipeline, various steps are run as Python or Shell scripts, while other complex tools are executed using Docker Containers. The modularity of the pipeline is designed such that each run produces a single antibody candidate, tagged with a randomly generated experiment name, along with an evaluation of its performance against a given target. Thus, the *Frankies* pipeline can be run n times, without modifying the initial configuration, to produce n unique Fv candidates. The overall workflow is shown in Figure 4.

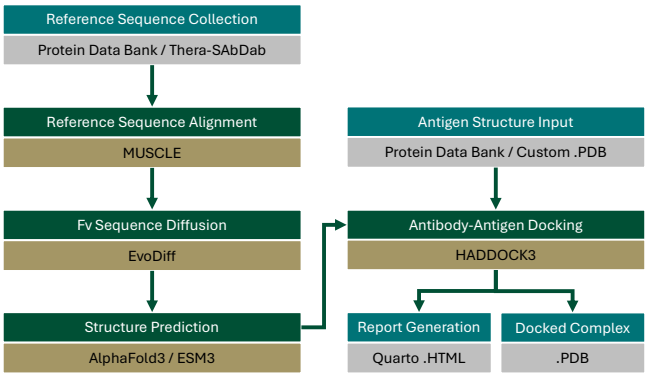


Fig. 4. *Frankies* pipeline workflow steps

Reference Dataset Preparation. To guide conditional sequence generation, we curated a reference set of HA1-targeting Fv sequences from the Therapeutic Structural Antibody Database (Thera-SAbDab) combined with additional antibodies used in Ford et al. 2025. Sequences were trimmed to retain only the Fv portion of the antibody and then were aligned using MUSCLE v3.8.425 (5). This produced separate input .a3m alignment files for heavy chain and light chain sequences.

Conditional Sequence Generation. We used EvoDiff's `MSA_OA_DM_MAXSUB` model for conditional sequence generation (6). The conditioned sequence generation sampled

from the curated reference sequence files using the 'Max-Hamming' distance to produce novel, HA1-targeted Fv chain sequences. EvoDiff operates as a protein diffusion model trained on general sequence databases and can be guided by user-provided templates or alignments of desired reference sequences in .a3m format.

Thirty Fv candidate sequences were generated and filtered for naturalness using AbNumber, a Python wrapper for ANARCI (7). This attempted to number the diffused sequences with the Chothia numbering system and, if it failed, the diffusion process would restart generate another sequence. This was performed to help ensure that the diffused sequences were as antibody-like as possible, lending to more confident subsequent folding and improve future protein synthesis capabilities.

Structure Prediction. Diffused heavy and light chain sequences were folded using either AlphaFold3 or ESM3.

AlphaFold3 may be desirable for groups wishing to perform purely local folding. While AlphaFold3 provides highly accurate structure predictions, it requires the user to download and store >600GB of reference databases and model weights to run. Plus, the multimer prediction of the heavy and light sequences together take approximately 15 minutes on a GPU.

Alternatively, Evolutionary Scale's ESM package and `esm3-medium-multimer-2024-09` model allow for API-based multimer structure prediction without any reference databases. This package returns the predicted structure in a few seconds.

Both AlphaFold3 and ESM3 consistently produce reliable antibody structures and thus we provide support for both models.

Complementarity-determining region (CDR) loops were detected using ANARCI and residues belonging to the CDR loop structures were selected as "active residues" in the subsequent docking process.

HA1 Antigen Preparation. A reference target antigen structure—HA1 subunit of H5N1 hemagglutinin—was obtained from the Protein Data Bank (PDB ID: 2VIR). This structure was used to test and validate the *Frankies* pipeline.

Then, the HA1 structure of a more recent isolate EPI3009174 was folded and used for subsequent novel antigen docking. Isolate EPI3009174 was collected in from a 9-year-old Cambodian male patient who passed away in 2024. This isolate was previously analyzed in Ford et al. 2025 against 11 reference antibodies.

For docking, the residue numbers in the antigen structure were incremented +1000, to avoid overlapping numbers with the antibody structure, and all residues were assigned to chain B.

Surface residues are automatically detected on the antigen structure and a 25% random subset of the surface residues was selected as "active residues".

Antibody-Antigen Docking. Fv structures were docked to the HA1 antigen using HADDOCK3 (8, 9) following the methodologies shown our recent previous studies (10–12). Docking configuration files were generated using an antibody-antigen docking template from the Bonvin Lab (13). These configuration files, along with the other required docking files, were generated in a docking preparatory step in the *Frankies* pipeline.

Docking was performed in a multi-stage protocol including rigid-body energy minimization, semi-flexible refinement, and explicit (water-based) solvent modeling. Docked complexes were scored using HADDOCK's built-in scoring function and further evaluated using the other biochemical/biophysical binding metrics listen below. The template for the HADDOCK protocol is available in the GitHub repository.

- Van der Waals intermolecular energy (*vdw*) in kcal/mol
- Electrostatic intermolecular energy (*elec*) in kcal/mol
- Desolvation energy (*desolv*) in kcal/mol
- Restraints violation energy (*air*) in arbitrary units
- Buried surface area (*bsa*) in $Å^2$
- Total energy (*total*): $1.0vdw + 1.0elec$ in kcal/mol
- HADDOCK score: $1.0vdw + 0.2elec + 1.0desolv + 0.1air$

Candidate Ranking, Reporting, and Visualization. The "best" complex is selected from the output of complexes based on the docking conformation with the best (lowest) HADDOCK score of the best cluster of complexes.

This best cluster and the best scoring complexes are reported in a rendered Quarto dashboard as the final step of the *Frankies* pipeline (14). This report shows the distribution of the various binding metrics of all complexes in the best cluster. This also renders the 3D structure of the best model as an interactive object using Py3Dmol, a Python wrapper for 3Dmol.js (15). A screenshot of the dashboard is shown in Figure 5.

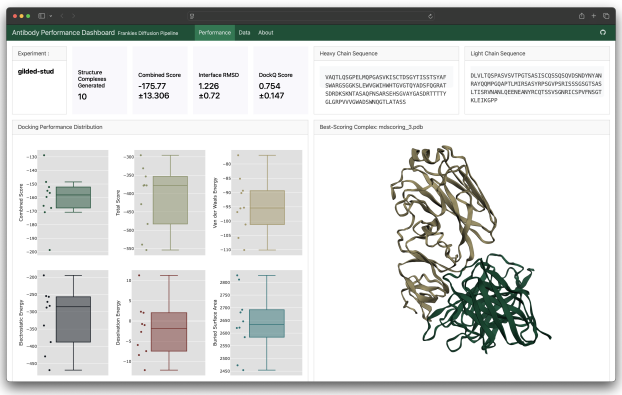


Fig. 5. *Frankies* output dashboard report. Generated with Quarto.

Discussion

As H5N1 continues to evolve, it is imperative that therapeutic advances continue to better target modern influenza clades of interest (e.g., 2.3.3.4b (16)) and therefore reduce the risk of mortality and morbidity in humans. As of late 2024, there are over 50 licensed H5 vaccine candidates (17), though many of these are based on significantly older strains. For example, in the United States, the 3 licensed vaccines are from 2007, 2013 and 2020 (18).

Some vaccine candidates are currently being developed using mRNA technologies on newer strains in the U.S. by Moderna

(mRNA-1018) (19) and Arcturus Therapeutics (ARCT-2304) (20). These were shown to be effective in animal trials and are currently in Phase I/II clinical trials.

Also, therapeutic antibodies are being developed that target H5 HA1 from clade 2.3.3.4b. One promising candidate was reported in a recent preprint by Alzua et al. 2025.

The *Frankies* pipeline presented in this study offers a streamlined and modular approach for the design of *de novo* antibody candidates against rapidly evolving viral targets such as H5N1 influenza hemagglutinin. By integrating generative protein diffusion, structure prediction, and flexible molecular docking, *Frankies* enables end-to-end discovery of Fv candidates with structural and biophysical characteristics that support high-affinity binding to the desired antigen.

A key innovation in *Frankies* lies in its conditional generative architecture. By seeding EvoDiff with therapeutically validated antibody sequences from Thera-SAbDab, we impose domain-specific constraints that preserve critical structural motifs while enabling conditional exploration of novel sequence space. This balances diversity with developability, reducing the likelihood of generating non-functional or unstable designs. Furthermore, our use of ANARCI and other quality control steps help to ensure that only the most promising candidates advance through the pipeline.

The application of *Frankies* to the H5N1 HA1 domain highlights the feasibility of using AI-driven approaches for pandemic preparedness. HA1 remains a challenging target due to its rapid antigenic drift and glycan shielding. Nonetheless, several *Frankies*-designed Fv candidates exhibited strong binding affinity scores and favorable interface properties, suggesting potential for neutralization. Importantly, these designs were generated *in silico* within minutes, underscoring the value of generative pipelines for rapid therapeutic prototyping.

Despite its strengths, *Frankies* has some limitations. The current conditional sequence generation can be improved in the future by implementing structure-aware diffusion (i.e., diffusing the CDR loops while a reference antibody is bound at the epitope site on the antigen). Also, better flexibility in CDR loop lengths is necessary.

While our predictions showcase the consistency of the Fv sequence generation, experimental validation will be crucial to confirm the binding specificity, affinity, and neutralization potency of the proposed candidates. Plus, the Fv sequences generated will need to be expanded to include the other parts of the antibody (the rest of the Fab region, the hinge, and the Fc region).

Today, the pipeline utilizes Docker to run the diffusion, folding, and docking steps, enabling users to run these processes without complex dependency installation in their local environment. While this is useful for cloud-based scalability, we will also implement the ability to choose the containerization engine. For example, Apptainer/Singularity are more common in on-premises high-performance computing (HPC) environments.

The current docking approach does not account for glycosylation, which plays a significant role in HA1 surface shielding and antibody accessibility. While HADDOCK3 provides valuable binding predictions, it lacks the full thermodynamic and kinetic accuracy of more computationally intensive free energy calculations. Thus, future improvements to the pipeline will include the integration of a molecular dynamics step, such as OpenMM (22), to model the trajectory of the antibody-antigen complex and to predict the stability of the interaction.

Future directions include fine-tuning EvoDiff on antibody–antigen co-evolution datasets, and benchmarking against existing broadly neutralizing antibodies targeting H5N1. Additionally, the modular nature of *Frankies* allows for straightforward extension to other antigens, including other influenza subtypes, entirely different viral families, or even targets in other diseases (such as in oncology, as shown in Ford 2024).

In conclusion, *Frankies* demonstrates how recent advances in protein generative modeling and structure prediction can be combined into an automated and scalable pipeline for therapeutic antibody design. As AI tools continue to mature, we anticipate that pipelines like *Frankies* will become central components of the next-generation biosecurity and drug discovery infrastructure for pandemic preparedness and beyond.

Contributors

Authors NS and CTF developed the *Frankies* pipeline. Author CTF performed data curation of the anti-HA1 antibody Fv sequences and their multiple sequence alignment. Author NS performed the docking experiments and statistical analyses. CTF and NS generated all visualization figures. NS and CTF performed the formal analysis of the structure and docking predictions of the antibodies. All authors wrote the original draft of the manuscript. All authors read and approved the final version of the manuscript.

Declaration of Interests

Author CTF is the owner of Tuple, LLC, a biotechnology consulting firm. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

We gratefully acknowledge all GISAID data contributors (i.e., the authors and their originating laboratories) responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

We acknowledge the following entities at the University of North Carolina at Charlotte: Academic Affairs, The Office of Research, The Center for Computational Intelligence to Predict Health and Environmental Risks (CIPHER), The Department of Bioinformatics and Genomics, The College of Computing and Informatics, and the University Research Computing group. We gratefully acknowledge the support of the Belk Family.

Data Sharing Statement

All code, data, results, and additional analyses are openly available on GitHub at: https://github.com/Santollan/Frankies. These data include all sequences and folded structures for the reference antibodies and H5N1 isolates used in this study, analysis scripts, and docking metrics. This also includes the experimental outputs for the 30 diffused Fv candidates.

Funding Statement

No external funding was used for this study.

Bibliography

1. Michael G. Ison and Jeanne Marrazzo. The emerging threat of h5n1 to human health. *New England Journal of Medicine*, 392(9):916–918, 2025. doi: 10.1056/NEJMe2416323.
2. Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, June 2004.
3. Colby T. Ford, Shirish Yasa, Khaled Obeid, Rafael Jaimes III, Phillip J. Tomezsko, Sayal Guirales-Medrano, Richard Allen White III, and Daniel Janies. Large-scale computational modelling of H5 influenza variants against HA1-neutralising antibodies. *eBioMedicine*, 114, Apr 2025. ISSN 2352-3964. doi: 10.1016/j.ebiom.2025.105632.
4. Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 34(20):3600–3600, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty350.
5. Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 03 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh340.
6. Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Robert Strome, Alan M. Moses, Alex X. Lu, Nicolò Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2024. doi: 10.1101/2023.09.11.556673.
7. James Dunbar and Charlotte M. Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 09 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv552.
8. G.C.P. van Zundert, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E. Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries, and A.M.J.J. Bonvin. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, 428(4):720–725, 2016. doi: 10.1016/j.jmb.2015.09.014.
9. João M.C. Teixeira, Rodrigo Vargas Honorato, Marco Giulini, Alexandre Bonvin, SarahAlidoost, Victor Reys, Brian Jimenez, Douwe Schulte, Charlotte van Noort, Stefan Verhoeven, Barbara Vreede, SSchott, and Regen Tsai. haddocking/haddock3: v3.0.0-beta.5. January 2024. doi: 10.5281/zenodo.10527751.
10. Colby T. Ford. PD-1 Targeted Antibody Discovery Using AI Protein Diffusion. *Technology in Cancer Research & Treatment*, 23:15330338241275947, 2024. doi: 10.1177/15330338241275947. PMID: 39228166.
11. Phillip J. Tomezsko, Colby T. Ford, Avery E. Meyer, Adam M. Michaleas, and Rafael Jaimes. Human cytokine and coronavirus nucleocapsid protein interactivity using large-scale virtual screens. *Frontiers in Bioinformatics*, Volume 4 - 2024, 2024. ISSN 2673-7647. doi: 10.3389/fbinf.2024.1397968.
12. Colby T. Ford, Shirish Yasa, Denis Jacob Machado, Richard Allen White, and Daniel A. Janies. Predicting changes in neutralizing antibody activity for sars-cov-2 xbb.1.5 using in silico protein modeling. *Frontiers in Virology*, Volume 3 - 2023, 2023. ISSN 2673-818X. doi: 10.3389/fviro.2023.1172027.
13. Bonvin Lab. HADDOCK3 Antibody-Antigen Tutorial. https://www.bonvinlab.org/education/HADDOCK3/HADDOCK3-antibody-antigen/, 2024. Accessed: 2024-06-21.
14. J.J. Allaire, Charles Teague, Carlos Scheidegger, Yihui Xie, Christophe Dervieux, and Gordon Woodhull. Quarto. November 2024. doi: 10.5281/zenodo.5960048.
15. Nicholas Rego and David Koes. 3dmol.js: molecular visualization with webgl. *Bioinformatics*, 31(8):1322–1324, 12 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu829.
16. Daniel Janies, Kary Ocaña, Sayal Guirales-Medrano, Khaled Obeid, Rachel Alexander, and Colby T. Ford. Analyses of phylogenetics, natural selection, and protein structure of clade 2.3.4.4b h5n1 influenza a reveal that recent viral lineages have evolved promiscuity in host range and improved replication in mammals in north america. *bioRxiv*, 2025. doi: 10.1101/2025.03.15.641219.
17. Jessica Taaffe, Shuyi Zhong, Shoshanna Goldin, Kate S. Rawlings, Benjamin J. Cowling, and Wenqing Zhang. An overview of influenza H5 vaccines. *The Lancet Respiratory Medicine*, 13(4):e20–e21, Apr 2025. ISSN 2213-2600. doi: 10.1016/S2213-2600(25)00052-9.
18. U.S. Food and Drug Administration. FDA Briefing Document: Vaccines and Related Biological Products Advisory Committee Meeting October 10, 2024 – Highly Pathogenic Avian Influenza (H5) Virus Vaccines, 2024. Accessed: 2025-04-24.
19. Inc. Moderna. Moderna announces updates on pandemic influenza program, 2025. Accessed: 2025-04-24.
20. Arcturus Therapeutics Holdings Inc. Arcturus therapeutics receives u.s. fda fast track designation for arct-810, mrna therapeutic candidate for ornithine transcarbamylase deficiency, 2023. Accessed: 2025-04-24.
21. Garazi Peña Alzua, André Nicolás León, Temima Yellin, Disha Bhavsar, Madhumathi Loganathan, Kaitlyn Bushfield, Philip J.M. Brouwer, Alesandra J. Rodriguez, Trushar Jeevan, Richard Webby, Christine Marizzi, Julianna Han, Andrew B. Ward, J. Andrew Duty, and Florian Krammer. Human monoclonal antibodies that target clade 2.3.4.4b H5N1 hemagglutinin. *bioRxiv*, 2025. doi: 10.1101/2025.02.21.639446.
22. Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charlles R. A. Abreu, Stephen E. Farr, Emilio Gallicchio, Anton Gorenko, Michael M. Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João PGLM Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. Openmm 8: Molecular dynamics simulation with machine learning potentials. *The Journal of Physical Chemistry B*, 128(1):109–116, Jan 2024. ISSN 1520-6106. doi: 10.1021/acs.jpcb.3c06662.