(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
04 November 2021 (04.11.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/222546 A1

(51) International Patent Classification:
G01N 33/68 (2006.01) *C07K 16/00* (2006.01)

(21) International Application Number:
PCT/US2021/029869

(22) International Filing Date:
29 April 2021 (29.04.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/018,559 01 May 2020 (01.05.2020) US

(71) Applicant: **UNIVERSITY OF PITTSBURGH-OF THE COMMONWEALTH SYSTEM OF HIGHER EDUCATION** [US/US]; 1st Floor Gardner Steel Conference Center, 130 Thackeray Avenue, Pittsburgh, Pennsylvania 15260 (US).

(72) Inventors: **SHI, Yi**; 600 Fairgate Dr., Wexford, Pennsylvania 15090 (US). **XIANG, Yufei**; 4628 Bayard St., Apt. 317, Pittsburgh, Pennsylvania 15213 (US). **SANG, Zhe**; 5522 Baum Blvd., Apt. 518, Pittsburgh, Pennsylvania 15232 (US).

(74) Agent: **PAVENTO, Lisa C.** et al.; Meunier Carlin & Curfman LLC, 999 Peachtree Street NE, Suite 1300, Atlanta, Georgia 30309 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: COMPOSITIONS AND METHODS FOR IDENTIFYING NANOBODIES AND NANOBODY AFFINITIES

(57) Abstract: Provided herein are methods of identifying a group of complementarity determining region (CDR)3, 2 and/or 1 nanobody amino acid sequences (CDR3, CDR2 and/or CDR1 sequences) wherein a reduced number of the CDR3, CDR2 and/or CDR1 sequences are false positives as compared to a control, methods for determining antigen affinity of nanobody peptide sequences, and related methods for training a deep learning model.

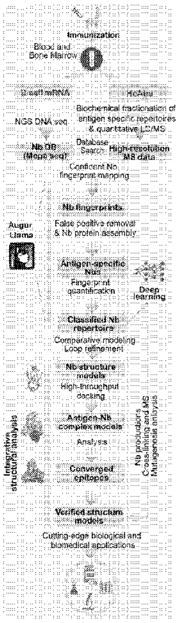

Immunization
Blood and Bone Marrow
B-cell mRNA
HcAbs
NGS DNA seq
Biochemical fractionation of antigen specific repertoires & quantitative LC/MS
Nb DB (Mega seq)
Database Search
High-resolution MS data
Confident Nb fingerprint mapping
Nb fingerprints
Augur Llama
False positive removal & Nb protein assembly
Antigen-specific Nbs
Fingerprint quantification
Deep learning
Classified Nb repertoire
Comparative modeling Loop refinement
Nb structure models
High-throughput docking
Antigen-Nb complex models
Analysis
Converged epitopes
Integrative structural analysis
Nb productions Cross-linking and MS Mutagenesis analysis
Verified structure models
Cutting-edge biological and biomedical applications

FIG. 2A

WO 2021/222546 A1

[Continued on next page]

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

COMPOSITIONS AND METHODS FOR IDENTIFYING NANOBODIES AND NANOBODY AFFINITIES

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 63/018,559, filed May 1, 2020, which is expressly incorporated herein by reference in its entirety.

BACKGROUND

Nanobodies (Nbs) are natural antigen-binding fragments derived from the V_HH domain of camelid heavy-chain only antibodies (HcAbs). They are characterized by their small size and outstanding structural robustness, excellent solubility and stability, ease of bioengineering and manufacturing, low immunogenicity in humans and fast tissue penetration. For these reasons, Nbs have emerged as promising agents for cutting-edge biomedical, diagnostic and therapeutic applications (Muyldermans, 2013; Beghein, 2017; Rasmussen, 2011; Jovcevska, I. & Muyldermans, S, 2020).

Display-based technologies have been developed for Nb discovery (Lauwereys, 1998; Pardon, 2014; McMahon, 2018; Egloff, 2019). These methods usually yield a small handful of target synthetic Nbs that bind specific targets with moderate affinities and do not directly analyze naturally circulating, antigen-specific HcAb/Nb repertoires. Recently, mass spectrometry-based proteomics has emerged as a promising technique for Nb discovery (Fridy, 2014). However, significant challenges remain towards a large-scale, sensitive, and reliable analysis of antigen-specific Nb proteomes for at least several reasons: (a) the diversity and dynamic range of circulating antibodies are orders of magnitude higher than any cellular proteome. (b) A Nb sequence database, obtained from an immunized camelid, usually contains millions of unique sequences posing a challenge for accurate database search (Savitski, 2015). (c) This massive database is overrepresented by conserved Nb framework sequences, which provide little specificity for identification. The specificity is largely determined by complementarity-determining regions (CDRs), among which CDR3 loops can be long, rendering it difficult for confident MS analysis. (d) Current methods are limited by the availability of efficient protocols and informatics that enable accurate quantification and classification of large Nb repertoires.

SUMMARY

Provided herein is a method of identifying a group of complementarity determining region (CDR)3, 2, and/or 1 nanobody amino acid sequences (CDR3, CDR2 and/or CDR1 sequences) wherein a reduced number of the CDR3, CDR2 and/or CDR1 sequences are false positives as compared to a control, the method comprising: (a) obtaining a blood sample from a camelid immunized with an antigen; (b) using the blood sample to obtain a nanobody cDNA library; (c) identifying the sequence of each cDNA in the library; (d) isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen; (e) digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products; (f) performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data; (g) selecting sequences identified in step c. that correlate with the mass spectrometry data; (h) identifying sequences of CDR3, CDR2 and/or CDR1 regions in the sequences from step g.; and (i) selecting from the CDR3, CDR2 and/or CDR1 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the selected sequences of step (i) comprise a group having the reduced number of false positive CDR3, CDR2 and/or CDR1 sequences. In some embodiments, step (d) comprises obtaining plasma from the blood sample and isolating nanobodies using one or more affinity isolation methods. In some aspects, the one or more affinity isolation methods of step (d) comprise one or more of protein G sepharose affinity chromatography and protein A sepharose affinity chromatography. In some aspects, step (d) further comprises a functional selection step comprising selecting antigen-specific nanobodies using an antigen-specific affinity chromatography and eluting the antigen-specific nanobodies under varying degrees of stringency thereby creating different nanobody fractions, and performing steps (e) through (i) on each fraction individually and estimating an affinity of each different step (i) CDR3, CDR2 and/or CDR1 region sequence for the antigen based on a relative abundance of the CDR3, CDR2 and/or CDR1 region sequence, respectively, in each of the nanobody fractions.

In some embodiments, a group of complementarity determining region (CDR)3 nanobody amino acid sequences (CDR2 sequences) wherein a reduced number of the CDR3 sequences are false positives as compared to a control, the method comprising: (a) obtaining a blood sample from a camelid immunized with an antigen; (b) using the blood sample to obtain a nanobody cDNA library; (c) identifying the sequence of each cDNA in the library; (d) isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen; (e) digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products; (f) performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data; (g) selecting sequences identified in step c. that correlate with the mass spectrometry data; (h) identifying sequences of CDR3 regions in the sequences from step g.; and (i) selecting from the CDR3 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the selected sequences of step (i) comprise a group having the reduced number of false positive CDR3 sequences. In some embodiments, step (d) comprises obtaining plasma from the blood sample and isolating nanobodies using one or more affinity isolation methods. In some aspects, the one or more affinity isolation methods of step (d) comprise one or more of protein G sepharose affinity chromatography and protein A sepharose affinity chromatography. In some aspects, step (d) further comprises a functional selection step comprising selecting antigen-specific nanobodies using an antigen-specific affinity chromatography and eluting the antigen-specific nanobodies under varying degrees of stringency thereby creating different nanobody fractions, and performing steps (e) through (i) on each fraction individually and estimating an affinity of each different step (i) CDR3 region sequence for the antigen based on a relative abundance of the CDR3 region sequence in each of the nanobody fractions.

In some embodiments, a group of complementarity determining region (CDR)2 nanobody amino acid sequences (CDR2 sequences) wherein a reduced number of the CDR2 sequences are false positives as compared to a control, the method comprising: (a) obtaining a blood sample from a camelid immunized with an antigen; (b) using the blood sample to obtain a nanobody cDNA library; (c) identifying the sequence of each cDNA in the library; (d) isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen; (e) digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products; (f) performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data; (g) selecting sequences identified in step c. that correlate with the mass spectrometry data; (h) identifying sequences of CDR2 regions in the sequences from step g.; and (i) selecting from the CDR2 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the selected sequences of step (i) comprise a group having the reduced number of false positive CDR2 sequences. In some embodiments, step (d) comprises obtaining plasma from the blood sample and isolating nanobodies using one or more affinity isolation methods. In some aspects, the one or more affinity isolation methods of step (d) comprise one or more of protein G sepharose affinity chromatography and protein A sepharose affinity chromatography. In some aspects, step (d) further comprises a functional selection step comprising selecting antigen-specific nanobodies using an antigen-specific affinity chromatography and eluting the antigen-specific nanobodies under varying degrees of stringency thereby creating different nanobody fractions, and performing steps (e) through (i) on each fraction individually and estimating an affinity of each different step (i) CDR2 region sequence for the antigen based on a relative abundance of the CDR2 region sequence in each of the nanobody fractions.

In some embodiments, a group of complementarity determining region (CDR)1 nanobody amino acid sequences (CDR1 sequences) wherein a reduced number of the CDR1 sequences are false positives as compared to a control, the method comprising: (a) obtaining a blood sample from a camelid immunized with an antigen; (b) using the blood sample to obtain a nanobody cDNA library; (c) identifying the sequence of each cDNA in the library; (d) isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen; (e) digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products; (f) performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data; (g) selecting sequences identified in step c. that correlate with the mass spectrometry data; (h) identifying sequences of CDR1 regions in the sequences from step g.; and (i) selecting from the CDR1 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the selected sequences of step (i) comprise a group having the reduced number of false positive CDR1 sequences. In some embodiments, step (d) comprises obtaining plasma from the blood sample and isolating nanobodies using one or more affinity isolation methods. In some aspects, the one or more affinity isolation methods of step (d) comprise one or more of protein G sepharose affinity chromatography and protein A sepharose affinity chromatography. In some aspects, step (d) further comprises a functional selection step comprising selecting antigen-specific nanobodies using an antigen-specific affinity chromatography and eluting the antigen-specific nanobodies under varying degrees of stringency thereby creating different nanobody fractions, and performing steps (e) through (i) on each fraction individually and estimating an affinity of each different step (i) CDR1 region sequence for the antigen based on a relative abundance of the CDR1 region sequence in each of the nanobody fractions.

In some embodiments, the antigen-specific affinity chromatography is a resin conjugated to the antigen. In some embodiments, the antigen-specific affinity chromatography is a resin coupled to a protein tag and the antigen. In some embodiments, the antigen-specific affinity chromatography is a resin coupled to a maltose binding protein and the antigen.

Some aspects further comprise creating a CDR3, CDR2, or CDR1 peptide having a sequence identified in step (i). Some aspects further comprise creating a nanobody comprising a CDR3, CDR2, and/or CDR1 region having a sequence identified in step (i).

Also included herein is a nanobody comprising an amino acid sequence selected from SEQ ID NOs: 1-2536 and SEQ ID NOs: 2665-2667.

Further provided herein is a computer-implemented method, comprising: (a) receiving a nanobody peptide sequence; (b) identifying a plurality of complementarity-determining region (CDR) regions of the nanobody peptide sequence, the CDR regions including CDR3, CDR2 and/or CDR1 regions; (c) applying a fragmentation filter to discard one or more false positive CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence; (d) quantifying an abundance of one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence; and (e) inferring an antigen affinity based on the quantified abundance of the one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence.

In some embodiments, the computer-implemented method further comprises classifying the one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence as having a low antigen affinity, mediocre antigen affinity, or high antigen affinity.

In some embodiments, the computer-implemented method further comprises assembling the one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence classified as having the high antigen affinity into a nanobody protein.

In some aspects of the computer-implemented method, the fragmentation filter is configured to require a minimum calculated fragmentation coverage percentage. In other or further aspects, the minimum calculated fragmentation coverage percentage is about 30. In some aspects, the minimum calculated fragmentation coverage percentage is about 50 for trypsin-treated samples and about 40 for chymotrypsin-treated samples.

In some embodiments, the computer-implemented method further comprises receiving a plurality of nanobody peptide sequences; and comparing each of the nanobody peptide sequences to a database to separate the nanobody peptide sequences into an excluded subgroup and a non-excluded subgroup, wherein the nanobody peptide sequences of the excluded subgroup are not found in the database, and wherein the CDR regions are only identified in the nanobody peptide sequences of the non-excluded subgroup.

In some embodiments of the computer-implemented method, the abundance of one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence is quantified based on relative MS1 ion signal intensities. In some embodiments, the antigen affinity is inferred using k-means clustering based on epitope similarity.

Also provided herein is a method for training a deep learning model, comprising: creating a dataset using the computer-implemented method described above; and training, using the dataset, a deep learning model to classify nanobody peptide sequences having low antigen affinity and nanobody peptide sequences having high antigen affinity, wherein the dataset comprises a plurality of nanobody peptide sequences and corresponding antigen-affinity labels. In some embodiments, the deep learning model is a convolutional neural network.

Further provided herein is a method for determining antigen affinity of nanobody peptide sequences, comprising: receiving a nanobody peptide sequence; inputting the nanobody peptide sequence into a trained deep learning model; and classifying, using the trained deep learning model, the nanobody peptide sequence as having low antigen affinity or high antigen affinity. In some embodiments, the deep learning model is a convolutional neural network. In some embodiments, the trained deep learning model is trained according to method for training a deep learning model described above

DESCRIPTION OF DRAWINGS

FIG. 1(A-K). *In-silico* analysis of a NGS Nb database reveals the superiority of chymotrypsin for Nb proteomics. (A) A Nb crystal structure (PDB: 4QGY). CDR loops are color coded. (B) Sequence length distributions of CDRs of the database. (C) *In-silico* digestion of the Nb database by two proteases and a cumulative plot of corresponding peptide masses. (D) The length distributions for both trypsin and chymotrypsin digested CDR3 peptides. (E) Complementarity of trypsin and chymotrypsin for Nb mapping based on simulation. 10,000 Nbs with unique CDR3 sequences were randomly selected and *in silico* digested to produce CDR3 peptides. Peptides with molecular weights of 0.8- 3 kDa and with sufficient CDR3 coverage (≥ 30%) were used for Nb mapping. (F-G) Evaluations of unique CDR3 peptide identifications (1F: trypsin; 1G: chymotrypsin) based on the percentage of CDR3 fragment ions that were matched in the MS/MS spectra. CDR3 peptides were identified by database search using either the "target" database (in salmon) or the "decoy" database (in grey). (H-K) 3D plots of the normalized CDR3 peptide identifications from the target database search, the percentages of CDR3 fragmentations, and CDR3 length. FDR: false discovery rate. FDRs of CDR3 identifications are colored on the 3D plots. The color bar shows the scale of FDR. FDR below 5% are presented in gradient red. (1H: analysis by trypsin; 1I: analysis by chymotrypsin.) (J-L). Representative high-quality MS/MS spectra of trypsin and chymotrypsin-digested CDR3 peptides. The sequence in FIG. 1K is NTVYLEMNSLKPEDTAVYSCAAGVSDYGCYR (SEQ ID NO: 2656). The sequence in FIG. 1L is YCAAAEGLASGSY (SEQ ID NO: 2657).

FIG. 2(A-G). Schematics of the hybrid proteomic pipeline for reliable and in-depth analysis of antigen-engaged Nb proteomes. (A) Schematic of the pipeline for Nb proteomics. The pipeline consists of three main components: camelid immunization and purification of antigen-specific Nbs, proteomic analysis of Nbs (facilitated by a dedicated software Augur Llama and deep-learning), and high-throughput integrative structural analysis of antigen-Nb complexes. (B) ELISA measurements of the camelid immune responses of three antigens of GST, HSA and the PDZ. (C) Identifications of unique CDR combinations and unique CDR3 sequences for different antigens. (D) A comparison between trypsin and chymotrypsin for CDR3 mapping of high-quality Nb$_{GST}$. (E)

Comparisons of Nb_{GST} CDR3 identifications by three different proteases (gluC, trypsin and chymotrypsin). The results were based on three independent experiments. (F) The solubility of the randomly selected antigen-specific Nbs. (G) Verifications of the selected Nbs for antigen binding.

FIG. 3(A-L). Classification of Nb repertoires for GST, HSA and PDZ binding. (A) Label-free MS quantification and heat map analysis of $CDR3_{GST}$ fingerprints by chymotrypsin. (B) Reproducibility and precision of label-free $CDR3_{GST}$ peptide quantifications by chymotrypsin. (C) Percentages of different Nb affinity clusters that were classified by quantitative proteomics. (D) Linear Correlation ($R^2 = 0.85$) of Nb ELISA affinities (LogIC50 of O.D. 450nm) with SPR K_D measurements. (E) Boxplots of ELISA affinities of different Nb clusters. The p values were calculated based on the student's t test. * indicates a p value of < 0.05, ** indicates < 0.01, *** indicates < 0.001, **** indicates < 0.0001, ns indicates not significant. (F) A plot summarizing ELISA affinities of 25 Nb_{HSA} (circles), O.D. at 450 nm. K_D affinities of the top 14 ranked Nbs by ELISA were measured by SPR (triangles). (G) A plot summarizing the ELISA affinities of 11 soluble Nb_{PDZ}. (H) SPR kinetics analysis of representative Nb_{GST} from three different affinity clusters. For G60(C1), Ka(1/Ms)=4.9e3, Kd(1/s)=5.9e-3, K_D=1.3μM; for G95(C2), Ka(1/Ms)=1.4e4, Kd(1/s)=1.1e-3, K_D=77nM; For G13(C3), Ka(1/Ms)=4.74e5, Kd(1/s)=1.7e-4, K_D=360pM. (I) A representative SPR kinetics measurements of high-affinity Nb_{HSA}. For H14, Ka(1/Ms)=2.5e5, Kd(1/s)=5.75e-6, K_D=22.3pM. (J) The SPR kinetics measurement of Nb_{PDZ} P10. For P10, Ka(1/Ms)=2.06e6, Kd(1/s)=9.03e-6, K_D=4.4pM. (K) Immunoprecipitations of GST (1nM) by different Nbs-coupled dynabeads and GSH resin. (L) Schematic of the PDZ domain of the mammalian mitochondrial outer membrane protein 25. Fluorescence microscopic analysis of Nb_{PDZ} P10. The Nb was conjugated by Alexa Fluor 647 for native mitochondrial immunostaining of the COS-7 cell line. Mitotracker was used for positive control.

FIG. 4(A-K). The structural landscapes of HSA-specific Nb proteomes revealed by the integrative structural methods. (A) The sequence variations of pI and hydropathy between human and camelid serum albumins (upper panel,). The heatmap of the major epitopes mapped by structural docking (lower panel). (B) Cartoon representations of the four dominant HSA epitopes. HSA are presented in gray. E1, E2 and E3 are in salmon, orange and cyan, respectively. (C) Surface representations showing co-localizations of electrostatic potential surfaces with three major epitopes. (D) The HSA epitopes and their fractions (%) based on converged cross-link models (E1: residues 57-62, 135-169; E2: 322-331, 335, 356-365, 395-410; E3: 29-37, 86-91, 117-123, 252- 290; E4: 566-585, 595, 598-606 and E5:188-208, 300-306, 463-468). (E-G) Representative cross-link models of HSA-Nb complexes. The best scoring models were presented. Satisfied DSS or EDC cross-links are shown as blue sticks. (H) A putative salt bridge between glutamic acid 400 (HSA) and arginine 108 of a Nb CDR3 is presented. The local sequence alignment between HSA and camelid albumin is shown. (I) ELISA affinity screening (heatmap) of 19 different Nbs for binding to wild type HSA and the point mutant (E400R). * indicates decreased affinity. (J) A plot of the RMSDs (room-median-square-deviations) of HSA-Nb cross-link models. (K) Bar plots showing the percentage of all the DSS and EDC cross-links of HSA-Nbs that satisfied the models.

FIG. 5(A-K). Mechanisms of Nb affinity maturation. (A) Distributions of CDR3 lengths of high-affinity (dark) and low-affinity (light) Nb_{GST} and $Nb_{HSA.}$ (B) Comparisons of the pI of different Nbs. (C-D) Comparisons of pI and hydropathy of CDRs between different Nbs. (E) A plot of CDR3 sequences. The alignment is based on a random selection of 1,000 unique CDR3 sequences with the identical length of 15 residues. Schematic of CDR3 architecture: the hypervariable "head" is in dark grey and the semi-variable "torso" is in pale grey. (F) Pie charts of the amino acid compositions of the CDR3 heads (Nb_{GST} and Nb_{HSA}) and the CDR2s (Nb_{GST}). Only the top 6 abundant residues are shown. (G) The relative changes of abundant amino acids on CDR3 heads of both Nb_{GST} and Nb_{HSA}. Positive charged residues of K(lysine)/R(arginine)/H(histidine), negative charged residues of D(aspartic acid)/E(glutamic acid), aromatic residue of Y(tyrosine) and small flexible amino acids of G(glycine)/S(serine) are shown. (H) Comparisons of the relative abundance of Y, G and S on the CDR3 heads between high-affinity and low-affinity Nb_{HSA}. Their relative abundances are plotted as a function of the relative position of the respective residues. A representative structure (PDB: 5F1O) of antigen-Nb complex showing two tyrosines on the CDR3 head are inserted into the deep pockets of the antigen. (I) Correlation plots of the ELISA affinities and the number of specific amino acids on the CDR3 heads of Nb_{HSA}. Pearson correlation coefficients and the statistical values are shown. (J) The correlation plot of ELISA affinities and the number of positively charged residues on the CDR2s of Nb_{GST}. (K) Sequence logo of two representative convolutional CDR3 filters (Filter 14 for high-affinity Nb_{HSA}; filter 3 for low-affinity Nb_{HSA}) learned by a deep learning model. The sequence of the top panel of Figure 5K is SEQ ID NO: 2661 (YXXXXXX, residue 2 can be Y, L, D, R, or I; residue 3 can be K or G; residue 4 can be R, Y, T, or D; residue 5 can be P, D, or R, residue 6 can be E, Y, V, P, W or D; residue 7 can be G, W, D, or P). The sequence of the bottom panel of Figure 5K is SEQ ID NO: 2662 (YXXXLXX, residue 2 can be D, P, K, or A; residue 3 can be F, P, D, or A; residue 4 can be H, T, or G, residue 6 can be G, N; residue 7 can be R, P, D, or Y.

FIG. 6(A-H): The outstanding versatility of Nbs for antigen binding. (A) The electrostatic potential surface and the dominant E2 epitope of PDZ domain (PDB: 2JIK; E1: 7-8, 35-36, 43, 99-100, and E2: 25-26, 45-46, 48, 78-79, 82-83, 85-86). (B) A docking model by a long CDR3 (in deep salmon) of a high-affinity Nb$_{PDZ}$P10. (C) Comparison between a crystal structure of PDZ- peptide ligand complex (PDB:1EB9) and a docking model of PDZ-Nb complex. The conserved ligand binding sites are shown in cyan. Side chains of both CDR3 and the peptide ligand are shown. (D) A heatmap showing the ELISA affinities of 11 different Nbs for binding to wild type or a mutant (R46E: K48D) PDZ. * indicates a decrease of 10-100,000 fold ELISA affinity. (E) Plot comparisons of both the CDR3 lengths (upper panel) and pIs (lower panel) of different Nbs (high-affinity Nb$_{HSA}$, Nb$_{GST,}$ Nb$_{PDZ}$ and Nbs from the sequence database). The data was smoothed with a gaussian function. (F) Comparisons of pI and hydropathy among different Nbs. (G) Pie charts of the top 6 most abundant amino acids on the Nb CDR3 heads. (H) A schematic model for antigen binding by Nbs.

FIG. 7(A-F). Analysis of NGS Nb databases and representative false positive CDR3 peptide identifications. (A) The normalized variability of Nb sequences. Approximately 0.5 million unique Nb sequences were aligned based on IMGT numbering scheme to generate the plot. Amino acids were grouped based on their properties (i.e., positive, negative, polar, and nonpolar) and were color-coded. (B) The mass distribution of ~1.5 million peptide identifications of human proteins from PeptideAtlas. (C) In silico digestion of Nb NGS database by different proteases (AspN, GluC, LysC, Trypsin and Chymotrypsin) and plot of peptide masses. (D) The overlaps between the target Nb sequence database of the immunized Llama and a decoy database from another native Llama. ~ 0.5 million sequences were included in each database. (E) A representative low quality/false positive MS/MS spectrum (HCD) of a tryptic CDR3 peptide. (F) That of a chymotryptic CDR3 peptide. Few high-resolution fragment ions were matched in the spectra. The sequences in FIG. 7E are NTVYLQMNSLKPE (SEQ ID NO: 2658) and DTSIYYCAATPVFQSMSTMATESVYDYWGQGTQVTVSSEPK (SEQ ID NO: 2659). The sequence in FIG. 7F is CAAGSGVGLY (SEQ ID NO: 2660).

FIG. 8(A-J). The informatics pipeline of “Augur Llama” for Nb proteomics and validation of Nb binders. (A) Schematics of the informatic pipeline. Three modules including 1) peptide identifications, 2) Nb peptide and protein quality control, and 3) quantification and classifications were presented. Nb proteomics data is first searched against the search engine. The initial identifications that pass the search engine can be automatically annotated, and evaluated based on different quality filters at peptide and protein levels. High-quality fingerprint peptides that pass the quality filters can be quantified and clustered. (B) Illustrations of the Nb CDR3 spectrum and coverage quality filters. (C) Illustrations of peptide classification method. (D) Phylogenetic tree and Web logo analyses of 230 unique CDR3s of the identified Nb$_{PDZ}$. (E) Schematic of PCR amplifications of HcAb variable domain (V$_{H}$H) from B lymphocytes of the camelid. (F) DNA gel electrophoresis of the V_HH PCR amplicons from the cDNA libraries prepared from the immunized bone marrow/blood. (G) SDS-PAGE analysis of fractionated Nb_{GST} based on different fractionation protocols. (H) SDS-PAGE analysis of Nb_{PDZ}. Maltose-binding protein (MBP) tag was fused to PDZ domain and the fusion protein was used as affinity handle for isolation. MBP was used as a negative control for quantification. (I) Unique Nb identifications for different antigens. (J) Comparison of antigen-specific Nbs identified by either chymotrypsin or trypsin-based method. Y axis stands for the % of the positive hits that were randomly selected for verifications.

FIG. 9(A-D). Proteomic quantifications, biochemical verifications and affinity measurements of Nb_{GST}. (A) Proteomic quantifications and heatmap analysis of Nb_{GST} based on different fractionation methods. (B) Pearson correlations of LC retention times of different fractionated Nb peptide samples. (C) Representative GST beads-binding assay. GST coupled resin was used to specifically isolate recombinant Nb from the E.coli lysis. Red arrows indicate enriched Nbs. Inactivated resin was used for negative control. (D) SPR kinetic measurements of 10 representative Nb_{GST}.

FIG. 10(A-B). Characterizations of High-quality HSA and PDZ Nbs. (A) SPR kinetic measurements of representative high-affinity Nb_{HSA}. (B) Beads-binding assays of selected high-quality Nb_{PDZ}. Recombinant MBP fusion PDZ was used as an affinity handle for isolation of Nbs from E.coli lysates. MBP coupled resin was used for negative control. I: E.coli lysate input, B: beads control, P: affinity pullout by PDZ.

FIG. 11(A-G). Hybrid structural analysis of GST-Nb complexes. **(A)** Heatmap analysis of structural docking of 64,670 GST-Nb complexes showing three converged epitopes (E1: 75-88, 143-148; E2: 33-43, 107-127; E3: 158-200, 213-220). **(B)** Cartoon representations of the three dominant GST epitopes. GST dimers were presented in gray. E1, E2 and E3 were in pale yellow, orange, and deep teal respectively. **(C)** Surface representations showing colocalizations of electrostatic surfaces with three major epitopes. **(D)** GST epitopes and their abundances (%) based on converged cross-link models were shown with different colors.

FIG. 12(A-H). The analysis of the CDR sequences of different Nbs and the sequence conservation of camelid and human albumin. (A-B) Comparison of the abundance of amino acids on the CDR3 heads between high-affinity and low-affinity Nbs. (C-F) Comparison of CDR1 and CDR2 for different Nbs. (G) Comparison of the relative position of tyrosine (Y), glycine(G) and serine(S) on the CDR3 heads of GST Nbs. (H) Sequence alignment of human serum albumin and llama serum albumin. Conserved amino acids were highlighted.

FIG. 13(A-F). Comparison among different antigen epitopes. (A) Comparison of the geometries of a major epitope of three different antigens (i.e., E2 for PDZ, E3 for GST dimer and E3 for HSA). Different epitopes were color coded on the antigen structures. (B) The surface electrostatic potentials and the E1 epitope of the PDZ domain. (C) A plot of the solvent accessible areas of different epitopes. The y axis stands for the areas of different epitopes in square angstrom. (D) Net formal charges of the epitopes. (E) Relative abundance of different amino acids on the CDR3 heads. DB: NGS Nb sequence database. (F) Comparison of the pI of CDR1 and CDR2 among different antigen-specific Nbs.

FIG. 14 depicts an example of a computing system that executes methods and procedures described in certain embodiments of the present disclosure.

FIG. 15(A-B) shows the results of amino acid sequence filters that are derived from the deep learning approach. The sequence filters can be used to accurately separate high-affinity from low-affinity binding HSA Nbs. The sequence of FIG. 15A is SEQ ID NO: 2663 (LXYRXXX, residue 2 can be N, Y, V, or G; residue 5 can be L or W; residue 6 can be E, G, N, T, or S; residue 7 can be D or E). The sequence of FIG. 15B is SEQ ID NO: 2664 (XXXXXXX, residue 1 can be C, F, Q, S, H, K, L, Y, or R; residue 2 can be G, P, A, or N; residue 3 can be E, S, G, T, P, V, Y, H, or A; residue 4 can be C, A, S, P, or D; residue 5 can be I, W, V, T, or A; residue 6 can be M, Q, or H; residue 7 can be K, Y, Q, V, or W).

FIG. 16(A-C) shows the results of amino acid sequence filters that are derived from the deep learning approach. The sequence filters can be used to accurately separate high-affinity from low-affinity binding HSA Nbs. The sequence of FIG. 16A is SEQ ID NO: 2665 (TXXXLXX; residue 2 can be D, P, K,or A; residue 3 can be F, P, L, D, or A; residue 4 can be H, T, or G; residue 6 can be G, E, N, or R; residue 7 can be R, P, G, D, or Y). The sequence of FIG. 16B is SEQ ID NO: 2666 (XXRXXXX; residue 1 can be E, G, W, D, or I; residue 2 can be N, G, or C; residue 4 can be A, H, or D; residue 5 can be E, R, Y, A, or T; residue 6 can be G, A, or P; residue 7 can be L, S, or Y). The sequence of FIG. 16C is SEQ ID NO: 2667 (XXGAQXW; residue 1 can be R or A; residue 2 can be K or L; residue 6 can be L, G, Y, or W).

DETAILED DESCRIPTION

Here reported is an integrative proteomic platform for in-depth discovery, classification, and high-throughput structural characterization of antigen-engaged Nb repertoires. The sensitivity and robustness of the technologies were validated using antigens spanning three orders of magnitude in immune response including a small, weakly immunogenic antigen derived from mitochondrial membrane. Tens of thousands of highly diverse, specific Nb families were confidently identified and quantified according to their physicochemical properties; a significant fraction had sub-nM affinity. Using high-throughput structural modeling, structural proteomics, and deep learning, the structural landscapes of >100,000 antigen-Nb complexes were systematically surveyed to significantly advance the understanding of immunogenicity and Nb affinity maturation. The study has revealed a surprising efficiency, specificity, diversity, and versatility of the mammalian humoral immune system.

Terminology

As used in the specification and claims, the singular form "a," "an," and "the" include plural references unless the context clearly dictates otherwise. For example, the term "a cell" includes a plurality of cells, including mixtures thereof.

The term “about” as used herein when referring to a measurable value such as an amount, a percentage, and the like, is meant to encompass variations of ±20%, ±10%, ±5%, or ±1% from the measurable value.

“Administration” to a subject or “administering” includes any route of introducing or delivering to a subject an agent. Administration can be carried out by any suitable route, including oral, intravenous, intraperitoneal, intranasal, inhalation and the like. Administration includes self-administration and the administration by another.

The terms “antibody” and “antibodies” are used herein in a broad sense and include polyclonal antibodies, monoclonal antibodies, and bi-specific antibodies. In addition to intact immunoglobulin molecules, also included in the term “antibodies” are fragments or polymers of those immunoglobulin molecules, and human or humanized versions of immunoglobulin molecules or fragments thereof. Antibodies are usually heterotetrameric glycoproteins of about 150,000 daltons, composed of two identical light (L) chains and two identical heavy (H) chains. Each heavy chain has at one end a variable domain (V_H) followed by a number of constant domains. Each light chain has a variable domain at one end (V_L) and a constant domain at its other end.

As used herein, the terms “antigen” or “immunogen” are used interchangeably to refer to a substance, typically a protein, a nucleic acid, a polysaccharide, a toxin, or a lipid, which is capable of inducing an immune response in a subject. The term also refers to proteins that are immunologically active in the sense that once administered to a subject (either directly or by administering to the subject a nucleotide sequence or vector that encodes the protein) is able to evoke an immune response of the humoral and/or cellular type directed against that protein.

The terms “antigenic determinant” and “epitope” may also be used interchangeably herein, referring to the location on the antigen or target recognized by the antigen-binding molecule (such as the nanobodies of the invention). Epitopes can be formed both from contiguous amino acids (a “linear epitope”) or noncontiguous amino acids juxtaposed by tertiary folding of a protein. The latter epitope, one created by at least some noncontiguous amino acids, is described herein as a “conformational epitope.” An epitope typically includes at least 3, and more usually, at least 5 or 8-10 amino acids in a unique spatial conformation. Methods of determining spatial conformation of epitopes include, for example, x-ray crystallography and 2-dimensional nuclear magnetic resonance. See, e.g., Epitope Mapping Protocols in Methods in Molecular Biology, Vol. 66, Glenn E. Morris, Ed (1996).

The terms “antigen binding site”, “binding site” and “binding domain” refer to the specific elements, parts or amino acid residues of a polypeptide, such as a nanobody, that bind the antigenic determinant or epitope.

The term “biological sample” as used herein means a sample of biological tissue or fluid. Such samples include, but are not limited to, tissue isolated from animals. Biological samples can also include sections of tissues such as biopsy and autopsy samples, frozen sections taken for histologic purposes, blood, plasma, serum, sputum, stool, tears, mucus, hair, and skin. Biological samples also include explants and primary and/or transformed cell cultures derived from patient tissues. A biological sample can be provided by removing a sample of cells from an animal, but can also be accomplished by using previously isolated cells (e.g., isolated by another person, at another time, and/or for another purpose), or by performing the methods as disclosed herein in vivo. Archival tissues, such as those having treatment or outcome history can also be used.

The term “cDNA library” refers herein to a combination of different cDNA fragments, which constitute some portion of the transcriptome of a given organism.

The terms "CDR" and “complementarity determining region” are used interchangeably and refer to a part of the variable chain of an antibody that participates in binding to an antigen. Accordingly, a CDR is a part of, or is, an “antigen binding site.” In some embodiments, the nanobody comprises three CDR that collectively form an antigen binding site.

The term “comprising” and variations thereof as used herein is used synonymously with the term “including” and variations thereof and are open, non-limiting terms. Although the terms “comprising” and “including” have been used herein to describe various embodiments, the terms “consisting essentially of” and “consisting of” can be used in place of “comprising” and “including” to provide for more specific embodiments and are also disclosed.

“Composition” refers to any agent that has a beneficial biological effect. Beneficial biological effects include both therapeutic effects, e.g., treatment of a disorder or other undesirable physiological condition, and prophylactic effects, e.g., prevention of a disorder or other undesirable physiological condition. The terms also encompass pharmaceutically acceptable, pharmacologically active derivatives of beneficial agents specifically mentioned herein, including, but not limited to, a bacterium, a vector, polynucleotide, cells, salts, esters, amides, proagents, active metabolites, isomers, fragments, analogs, and the like. When the terms "composition" is used, then, or when a particular composition is specifically identified, it is to be understood that the term includes the composition per se as well as pharmaceutically acceptable, pharmacologically active vector, polynucleotide, salts, esters, amides, proagents, conjugates, active metabolites, isomers, fragments, analogs, etc.

A "control" is an alternative subject or sample used in an experiment for comparison purposes. A control can be "positive" or "negative."

"Effective amount" encompasses, without limitation, an amount that can ameliorate, reverse, mitigate, prevent, or diagnose a symptom or sign of a medical condition or disorder (e.g., cancer). Unless dictated otherwise, explicitly or by context, an "effective amount" is not limited to a minimal amount sufficient to ameliorate a condition. The severity of a disease or disorder, as well as the ability of a treatment to prevent, treat, or mitigate, the disease or disorder can be measured, without implying any limitation, by a biomarker or by a clinical parameter. In some embodiments, the term "effective amount of a recombinant nanobody" refers to an amount of a recombinant nanobody sufficient to prevent, treat, or mitigate a cancer. .

The "fragments" or "functional fragments," whether attached to other sequences or not, can include insertions, deletions, substitutions, or other selected modifications of particular regions or specific amino acids residues, provided the activity of the fragment is not significantly altered or impaired compared to the nonmodified peptide or protein. These modifications can provide for some additional property, such as to remove or add amino acids capable of disulfide bonding, to increase its bio-longevity, to alter its secretory characteristics, etc. In any case, the functional fragment must possess a bioactive property, such as binding to HSA and/or ameliorating cancer.

The term "fragmentation coverage percentage" refers to a percentage obtained using the following formula:

f(x,Enzyme) is the function to calculate fragmentation coverage (%) of peptides digested by Enzyme

x is the length of CDR3 that the peptide mapped

$f(x,\text{chymotrypsin}) = 0.0023x^2-0.0497x+0.7723,x[5,30]$

$f(x,\text{trypsin})=0.00006x^2 - 0.00444x+0.9194, x[5,30]$.

In some embodiments, a minimum calculated fragmentation coverage percentage is required. In other or further aspects, the required minimum calculated fragmentation coverage percentage is about 30. In some aspects, the required minimum calculated fragmentation coverage percentage is about 50 when trypsin is the enzyme and about 40 when chymotrypsin is the enzyme.

As used herein, a "functional selection step" is a method by which nanobodies are divided into different fractions or groups based upon a functional characteristic. In some embodiments, the functional characteristic is nanobody or CD3, CD2, or CD1 region antigen affinity. In other embodiments, the functional characteristic is nanobody thermostability. In other embodiments, the functional characteristic is nanobody intracellular penetration. Accordingly, the present invention includes a method of identifying a group of complementarity determining region (CDR)3, 2 or 1 region nanobody amino acid sequences (CDR3, CDR2 or CDR1 sequences) wherein a reduced number of the CDR3, CDR2 or CDR1 sequences are false positives as compared to a control, the method comprising: obtaining a blood sample from a camelid immunized with the antigen; using the blood sample to obtain a nanobody cDNA library; identifying the sequence of each cDNA in the library; isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen; performing a functional selection step; digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products; performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data; selecting sequences identified in step c. that correlate with the mass spectrometry data; identifying sequences of CDR3, CDR2 or CDR1 regions in the sequences from step g.; and excluding from the CDR3, CDR2 or CDR1 region sequences from step h. those sequences having less than a calculated fragmentation coverage percentage; wherein the non-excluded sequences comprise a group having the reduced number of false positive CDR3, CDR2 or CDR1 sequences. It should be understood that the method steps following the functional selection step can be performed separately on each different fraction or group created by the functional selection.

The "half-life" of an amino acid sequence, compound or polypeptide of the invention can generally be defined as the time taken for the serum concentration of the amino acid sequence, compound or polypeptide to be reduced by 50%, *in vivo*, for example due to degradation of the sequence or compound and/or clearance or sequestration of the sequence or compound by natural mechanisms. The *in vivo* half-life of a nanobody, amino acid sequence, compound or polypeptide of the invention can be determined in any manner known, such as by pharmacokinetic analysis. these, for example, Kenneth, A et al., Chemical Stability of Pharmaceuticals: A Handbook for Pharmacists; Peters et al., Pharmacokinete analysis: A Practical Approach (1996); "Pharmacokinetics", M Gibaldi & D Perron, published by Marcel Dekker, 2nd Rev. edition (1982).

The term "identity" or "homology" shall be construed to mean the percentage of nucleotide bases or amino acid residues in the candidate sequence that are identical with the bases or residues of a corresponding sequence to which it is compared, after aligning the sequences and introducing gaps, if necessary to achieve the maximum percent identity for the entire sequence, and not considering any conservative substitutions as part of the sequence identity. A polynucleotide or polynucleotide region (or a polypeptide or polypeptide region) that has a certain percentage (for example, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%,94%, 95%, 96%, 97%, 98%, 99% or higher) of "sequence identity" to another sequence means that, when aligned, that percentage of bases (or amino acids) are the same in comparing the two sequences. This alignment and the percent homology or sequence identity can be determined using software programs known in the art. Such alignment can be provided using, for instance, the method of Needleman et al. (1970) J. Mol. Biol. 48: 443-453, implemented conveniently by computer programs such as the Align program (DNAstar, Inc.). In some embodiments, percent identity is determined along the entire length of the compared sequences.

The term “increased” or “increase” as used herein generally means an increase by a statically significant amount; for the avoidance of any doubt, “increased” means an increase of at least 10% as compared to a reference level, for example an increase of at least about 20%, or at least about 30%, or at least about 40%, or at least about 50%, or at least about 60%, or at least about 70%, or at least about 80%, or at least about 90% or up to and including a 100% increase or any increase between 10-100% as compared to a reference level, or at least about a 2-fold, or at least about a 3-fold, or at least about a 4-fold, or at least about a 5-fold or at least about a 10-fold increase, or any increase between 2-fold and 10-fold or greater as compared to a reference level.

The term “isolating” as used herein refers to isolation from a biological sample, i.e., blood, plasma, tissues, exosomes, or cells. As used herein the term “isolated,” when used in the context of, e.g., a nucleic acid, refers to a nucleic acid of interest that is at least 60% free, at least 75% free, at least 90% free, at least 95% free, at least 98% free, and even at least 99% free from other components with which the nucleic acid is associated with prior to isolation.

The term “mass spectrometry” refers to a measurement of the mass-to-charge ratio (m/z) of one or more molecules present in a sample. “Mass spectrometry data” refers to mass, charge, mass-to-charge ratio, molecular weight and/or amino acid identity or sequence of the one or more molecules present in a sample. In some embodiments, the mass spectrometry data is the amino acid sequence of a molecule present in the sample. Sequences, including cDNA sequences, that "correlate" with mass spectrometry data have an expected same or highly similar amino acid sequence determined in the mass spectrometry step of the method. In some embodiments, a sequence correlates with mass spectrometry data when there is about 80%, about 85%, about 90%, about 91%, about 92%, about 93%, about 94%, about 95%, about 96%, about 97%, about 98%, or about 99% similarity or identity. In some embodiments, a sequence correlates with mass spectrometry data when there is about 90-100% similarity or identity.

As used herein, the terms "nanobody", "V_HH", "V_HH antibody fragment" are used indifferently and designate a variable domain of a single heavy chain of an antibody of the type found in *Camelidae*, which are without any light chains, such as those derived from Camelids as described in PCT Publication No. WO 94/04678, which is incorporated by reference in its entirety. As used herein, "single domain antibody" refers to a nanobody and an Fc domain.

The term "nucleic acid" as used herein means a polymer composed of nucleotides, e.g. deoxyribonucleotides (DNA) or ribonucleotides (RNA). The terms "ribonucleic acid" and "RNA" as used herein mean a polymer composed of ribonucleotides. The terms "deoxyribonucleic acid" and "DNA" as used herein mean a polymer composed of deoxyribonucleotides.

As used herein, "operatively linked" refers to the arrangement of polypeptide segments within a single polypeptide chain, where the individual polypeptide segments can be, without limitation, a protein, fragments thereof, linking peptides, and/or signal peptides. The term operatively linked can refer to direct fusion of different individual polypeptides within the single polypeptides or fragments thereof where there are no intervening amino acids between the different segments as well as when the individual polypeptides are connected to one another via a "linker" that comprises one or more intervening amino acids.

The term "reduced", "reduce", "reduction", or "decrease" as used herein generally means a decrease by a statistically significant amount. However, for avoidance of doubt, "reduced" means a decrease by at least 5% as compared to a reference level, for example a decrease by at least about 10%, or at least about 20%, or at least about 30%, or at least about 40%, or at least about 50%, or at least about 60%, or at least about 70%, or at least about 80%, or at least about 90% or up to and including a 100% decrease (i.e., absent level as compared to a reference sample), or any decrease between 10-100% as compared to a reference level.

The terms "polynucleotide" and "oligonucleotide" are used interchangeably, and refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof. Polynucleotides may have any three-dimensional structure, and may perform any function, known or unknown. The following are non-limiting examples of polynucleotides: a gene or gene fragment, exons, introns, messenger RNA (mRNA), transfer RNA, ribosomal RNA, ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers. A polynucleotide may comprise modified nucleotides, such as methylated nucleotides and nucleotide analogs. If present, modifications to the nucleotide structure may be imparted before or after assembly of the polymer. The sequence of nucleotides may be interrupted by non-nucleotide components. A polynucleotide may be further modified after polymerization, such as by conjugation with a labeling component. The term also refers to both double- and single-stranded molecules. Unless otherwise specified or required, any embodiment of this invention that is a polynucleotide encompasses both the double-stranded form and each of two complementary single-stranded forms known or predicted to make up the double-stranded form.

The term "polypeptide" is used in its broadest sense to refer to a compound of two or more subunit amino acids, amino acid analogs, or peptidomimetics. The subunits may be linked by peptide bonds. In another embodiment, the subunit may be linked by other bonds, e.g. ester, ether, etc. As used herein the term "amino acid" refers to either natural and/or unnatural or synthetic amino acids, including glycine and both the D or L optical isomers, and amino acid analogs and peptidomimetics. A peptide of three or more amino acids is commonly called an oligopeptide if the peptide chain is short. If the peptide chain is long, the peptide is commonly called a polypeptide or a protein. The terms “peptide,” “protein,” and “polypeptide” are used interchangeably herein.

“Recombinant” used in reference to a polypeptide refers herein to a combination of two or more polypeptides, which combination is not naturally occurring.

The term “specificity” refers to the number of different types of antigens or antigenic determinants to which a particular antigen-binding molecule (such as the nanobody of the invention) can bind. A nanobody with low specificity binds to multiple different epitopes (or polypeptide regions) via a single antigen binding site or binding domain, whereas a nanobody with high specificity binds to one or a few epitopes (or polypeptide regions) via a single antigen binding site or binding domain. In some embodiments, the few epitopes (or polypeptide regions) are similar or highly similar, such as, for example, cross-species epitopes. As used herein, the term "specifically binds," as used herein with respect to a nanobody refers to the nanobody’s preferential binding to an epitope (or polypeptide region) as compared with other epitopes (or polypeptide regions). Specific binding can depend upon binding affinity and the stringency of the conditions under which the binding is conducted. In one example, a nanobody specifically binds an epitope when there is high affinity binding under stringent conditions. In some embodiments, the HSA binding polypeptide or nanobody described herein specifically binds to human serum albumin.

It should be understood that the specificity of an antigen-binding molecule (e.g., the HSA binding polypeptides, the nanoantibodies of the present invention) can be determined based on affinity and/or avidity. The affinity, represented by the equilibrium constant for the dissociation of an antigen with an antigen-binding molecule (K_D), is a measure for the binding strength between an antigenic determinant and an antigen-binding site on the antigen-binding molecule: the lesser the value of the K_D, the stronger the binding strength between an antigenic determinant and the antigen-binding molecule (alternatively, the affinity can also be expressed as the affinity constant (K_A), which is 1/ K_D). Methods for determining affinity are well known to those of ordinary skill in the art. Avidity is the measure of the strength of binding between an antigen-binding molecule (such as the HSA binding polypeptides and the nanobodies of the present invention) and the pertinent antigen. Avidity is related to both the affinity between an antigenic determinant and its antigen binding site on the antigen-binding molecule and the number of pertinent binding sites present on the antigen-binding molecule. Typically, antigen-binding proteins (such as the HSA binding polypeptides and the nanobodies of the invention) will bind to their antigen with a dissociation constant (K_D) of 10^{-5} to 10^{-12} moles/liter or less, and preferably 10^{-7} to 10^{-12} moles/liter or less and more preferably 10^{-8} to 10^{-12} moles/liter (i.e., with an association constant (K_A) of 10^5 to 10^{12} liter/moles or more, and preferably 10^7 to 10^{12} liter/moles or more and more preferably 10^8 to 10^{12} liter/moles). In some embodiments, the Ka (on rate, 1Ms) is about 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, or 10^{11}. In some embodiments, the Ka is about 10^7. In some embodiments, the Kd (off rate, s) is about 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}, or 10^{-11}. In some embodiments, the K_D is about 10^{-7}. In some embodiments, the antigen-binding protein disclosed herein binds to its antigen with a K_D of less than about 10^{-9} moles/liter. Any K_D value greater than 10 μM is generally considered to indicate non-specific binding. The dissociation constant may be the actual or apparent dissociation constant, as will be clear to the person of ordinary skill in the art.

The term “subject” is defined herein to include animals such as mammals, including, but not limited to, primates (e.g., humans), cows, sheep, goats, horses, dogs, cats, rabbits, rats, mice and the like. In some embodiments, the subject is a human.

Compositions and Methods

In some aspects, disclosed herein is a method of identifying a group of complementarity determining region (CDR)3, 2 or 1 region nanobody amino acid sequences (CDR3, CDR2 or CDR1 sequences) wherein a reduced number of the CDR3, CDR2 or CDR1 sequences are false positives as compared to a control. The term “false positive” herein refers to a result that indicates something is present when it is not. Herein the phrase “sequences are false positive” refers to the CDR3, CDR2 and/or CDR1 sequences that do not specifically bind to the tested antigens, or to the CDR3, CDR2 and/or CDR1 sequences contained within a nanobody, which nanobody cannot specifically bind to the tested antigens. It should be understood that the number or amount of false positive CDR3, CDR2 and/or CDR1 sequences can be reduced using the methods disclosed herein with a fragmentation filter set at about at least 30% (for example, at least about 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 99%) for trypsin-treated samples and/or about at least 30% (for examples, at least about 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 99%) for chymotrypsin-treated samples. In some examples, the false positive CDR3, CDR2 and/or CDR1 sequences can be mostly removed using the methods disclosed herein with a fragmentation filter set at about 50% for trypsin-treated samples and/or about 40% for chymotrypsin-treated samples.

Accordingly, the disclosed method of identifying CDR3, CDR2 and/or CDR1 sequences can reduce the number of the CDR3, CDR2 and/or CDR1 sequences that are false positives as compared to a control. The reduction can be, for example, at least about a 2-fold, at least about a 3-fold, at least about a 4-fold, at least about a 5-fold, at least about a 10-fold, at least about a 20-fold, at least about a 50-fold, or at least about a 100-fold compared to the number of false positive CDR3, CDR2 and/or CDR1 sequences that are identified without using the method described herein.

In some embodiments, the method comprises:

a. obtaining a blood sample from a camelid immunized with an antigen;
b. using the blood sample to obtain a nanobody cDNA library;
c. identifying the sequence of each cDNA in the cDNA library;
d. isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen;
e. digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products;
f. performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data;
g. selecting sequences identified in step c. that correlate with the mass spectrometry data;
h. identifying sequences of CDR3, CDR2 and/or CDR1 regions in the sequences from step g.; and

i. selecting from the CDR3, CDR2 and/or CDR1 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the selected sequences comprise a group having the reduced number of false positive CDR3, CDR2 and/or CDR1 sequences.

In some embodiments, the method comprises:

a. obtaining a blood sample from a camelid immunized with an antigen;

b. using the blood sample to obtain a nanobody cDNA library;

c. identifying the sequence of each cDNA in the library;

d. isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen;

e. digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products;

f. performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data;

g. selecting sequences identified in step c. that correlate with the mass spectrometry data;

h. identifying sequences of CDR3, CDR2 and/or CDR1 regions in the sequences from step g.; and

i. selecting from the CDR3, CDR2 and/or CDR1 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the fragmentation coverage percentage is determined by a formula f(x,chymotrypsin) = 0.0023x2-0.0497x+0.7723,x[5,30] when chymotrypsin is used in step e. or a formula f(x,trypsin)=0.00006x2 - 0.00444x+0.9194, x[5,30] when trypsin is used in step e., and wherein x is the length of the CDR3, CDR2 and/or CDR1 region sequence; and

j. wherein the selected sequences of step i. comprise a group having the reduced number of false positive CDR3, CDR2 and/or CDR1 sequences.

In some aspects, the selected CDR3, CDR2 and/or CDR1 region sequences in step i. have a minimum required fragmentation coverage percentage of about 30. In some aspects, the selected CDR3, CDR2 and/or CDR1 region sequences in step i. have a minimum required fragmentation coverage percentage of about 50 and trypsin is used in step e. In some embodiments, the selected CDR3, CDR2 and/or CDR1 region sequences in step i. have a minimum required fragmentation coverage percentage about 40 and chymotrypsin is used in step e.

It should be understood that the nanobody cDNA library in step b. is obtained from a biological sample (e.g., a blood sample or bone marrow) of the immunized subject. In some embodiments, the cDNA library is obtained from the B cells. A cDNA (cloned cDNA or complementary DNA) library is a combination of cDNAs that are produced from mRNAs in a biological sample (e.g., a blood sample or bone marrow sample) using reverse transcription technology. The method of producing cDNA library is well-known in the art. Accordingly, in some embodiments, step b. further comprises a step of isolating mRNAs from a biological sample (e.g., a blood sample or a bone marrow sample) and/or a step of reverse transcribing the isolated mRNA to cDNAs.

The produced cDNAs are then sequenced as described in step c. In some embodiments, step c. further comprises a step of amplifying camelid IgG heavy chain cDNA sequences from the variable domain to the CH2 domain using specific primers (e.g., SEQ ID NO: 2646 and SEQ ID NO: 2647), a step of separating the V_HH genes that lack CH1 domain from conventional IgG (having CH1 domain) using DNA gel electrophoresis, a step of re-amplifying from framework 1 to framework 4 using a 2nd-Forward primer (e.g., SEQ ID NO: 2648) and a 2nd-Reverse primer (e.g., SEQ ID NO: 2649), a step of purifying the amplicon of this second PCR (e.g., using a PCR clean up kit or isolation kit), a step of another PCR with primers to add adapter for sequencing analysis (e.g., using forward primer SEQ ID NO: 2650 and reverse primer SEQ ID NO: 2651) for sequencing analysis (e.g., MiSeq sequencing analysis). The methods for sequencing analysis can be, for example, single molecule real time (SMRT) sequencing, nanopore DNA sequencing, massively parallel signature sequencing (MPSS), polony sequencing, 454 pyrosequencing, Illumina (Solexa) sequencing, combinatorial probe anchor synthesis (cPAS), SOLiD sequencing, or MiSeq sequencing.

Step d. above can be performed concurrently, prior, or following steps a, b, and/or c. In some examples, step d. further comprises obtaining plasma from the blood sample and isolating nanobodies using one or more affinity isolation methods. The affinity isolation methods can be any affinity isolation methods known in the art, including, for example, protein G sepharose affinity chromatography, protein A sepharose affinity chromatography, hydroxylapatite chromatography, gel electrophoresis, or dialysis. Protein G sepharose affinity chromatography and protein A sepharose affinity chromatography are two well-known affinity chromatography methods (Grodzki A.C., Berenstein E. (2010) Antibody Purification: Affinity Chromatography – Protein A and Protein G Sepharose. In: Oliver C., Jamur M. (eds) *Immunocytochemical Methods and Protocols. Methods in Molecular Biology (Methods and Protocols)*, vol 588. Humana Press.) The methods rely on the reversible interaction between a protein and a specific ligand immobilized in a chromatographic matrix. The sample is applied under conditions that favor specific binding to the ligand as the result of electrostatic and hydrophobic interactions, van der Waals' forces, and/or hydrogen bonding. After washing away the unbound material, the bound protein is recovered by changing the buffer conditions to those that favor desorption. Protein A sepharose affinity chromatography and G sepharose affinity chromatography are commonly used in antibody purification due to the high binding affinity and specificity of Protein A or G with the Fc region of the antibody. In some embodiments, the one or more affinity isolation methods of step d. comprise one or more of protein G sepharose affinity chromatography and protein A sepharose affinity chromatography.

In some examples, step d. also further comprises a functional selection step comprising selecting antigen-specific nanobodies using an antigen-specific affinity chromatography and eluting the antigen-specific nanobodies under varying degrees of stringency thereby creating different nanobody fractions, and performing steps e. through i. on each fraction individually and estimating an affinity of each different step i. CDR3, CDR2 and/or CDR1 region sequence for the antigen based on a relative abundance of the CDR3, CDR2 and/or CDR1 region sequence in each of the nanobody fractions, respectively. In some embodiments, the antigen-specific affinity chromatography is a resin conjugated to the antigen. In some embodiments, the antigen-specific affinity chromatography is a resin coupled to maltose binding protein and the antigen.

It should be understood and herein contemplated that the term “degrees of stringency” refers to different concentrations of salt buffer (e.g., from about 0.1M to about 20 M $MgCl_2$ in neutral pH buffer, preferably from about 1M to about 10 M $MgCl_2$ in neutral pH buffer, or preferably from about 1M to about 4.5 M $MgCl_2$ in neutral pH buffer), alkaline solutions with different pH values (e.g., 1-100 mM NaOH, about pH 11, 12 and 13), acidic solutions with different pH values (e.g., 0.1 M glycine, about pH 3, 2 and 1), or a combination thereof. It should also be understood that the term “different nanobody fractions” or “different biochemistry fractions” refers to different fractions of nanobodies that are eluted from an antigen-coupled solid support (e.g., a resin) under the different degrees of stringency. The nanobodies that are most resistant to high salt, high acidity or high alkalinity conditions have the highest affinity to the antigen.

The term “digestion products” herein, such as in step e., refers to the mixture of peptides following the step of digestion with an enzyme (including, for example, trypsin, chymotrypsin, LysC, GluC, and AspN). In some examples, the nanobodies are digested with trypsin(such as Pierce™ Trypsin Protease, MS Grade, Catalog number: 90057), chymotrypsin (such as Pierce™ Chymotrypsin Protease (TLCK treated), MS Grade, Catalog number: 90056), LysC (or Lys-C protease, such as Pierce™ Lys-C Protease, MS Grade, Catalog number: 90051), GluC (or Glu-C

Protease, such as Pierce™ Glu-C Protease, MS Grade, Catalog number: 90054), and/or AspN (or Asp-N protease, such as Pierce™ Asp-N Protease, MS Grade, Catalog number: 90053) to create the corresponding digestion products. Trypsin, chymotrypsin, LysC, GluC, and AspN are enzymes that digest proteins. The cleavage rules for digestion of nanobodies by these enzymes are:

Trypsin:	C-terminal to K/R, not followed by P
Chymotrypsin:	C-terminal to W/F/L/Y, not followed by P
GluC:	C-terminal to D/E, not followed by P
AspN:	N-terminal to D
LysC:	C-terminal to K

The digestion step can be performed at a temperature from about 2 °C to about 60 °C (e.g., at about 2 °C, 4 °C, 6 °C, 8 °C, 10 °C, 12 °C, 14 °C, 16 °C, 18 °C, 20 °C, 22 °C, 24 °C, 26 °C, 28 °C, 30 °C, 32 °C, 34 °C, 36 °C, 38 °C, 40 °C, 42 °C, 44 °C, 46 °C, 48 °C, 50 °C, 52 °C, 54 °C, 56 °C, 58 °C, or 60 °C) for about 5 min, 10 min, 30 min, 45 min, 1 hour, 2 hours, hours, 4 hours, 6 hours, 8 hours, 10 hours, 12 hours, 14 hours, 16 hours, 18 hours, 20 hours, 22 hours, 24 hour, 36 hours, 48 hours, or 72 hours.

Amino Acid Abbreviations		
Amino Acid	**Abbreviations**	
Alanine	Ala	A
allosoleucine	AIle	
Arginine	Arg	R
asparagine	Asn	N
aspartic acid	Asp	D
Cysteine	Cys	C
glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isolelucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
phenylalanine	Phe	F
proline	Pro	P
pyroglutamic acid	pGlu	
Serine	Ser	S
Threonine	Thr	T
Tyrosine	Tyr	Y
Tryptophan	Trp	W
Valine	Val	V

Step f. comprises performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data. The methods of using mass spectrometry for peptide analysis are well-known in the art. In some embodiments, the mass spectrometry analysis herein is performed in combination with gas chromatography (GC-MS), liquid chromatography (LC-MS), capillary electrophoresis (CE-MS), ion mobility spectrometry-mass spectrometry (IMS/MS or IMMS), Matrix Assisted Laser Desorption Ionisation (MALDI-TOF), Surface Enhanced Laser Desorption Ionization (SELDI-TOF), or Tandem MS (MS-MS). This step can identify the sequence of the nanobody, or a portion of a nanobody in the sample, based on mass of the amino acids and sequence homology search in a database of polypeptides translated from the cDNA library of step b. In some examples, mass spectrometry is used to analyze and generate a spectrum of digestion products from each nanobody fraction separately. In some examples, the spectrum of the digestion productions refers to the electron ionization data that are present as intensity versus m/z (mass-to-charge ratio) plot.

It should be understood herein that the nanobody sequence determination is not only based on mass spectrometry. It is determined by matching/correlating the sequences identified by mass spectrometry with the sequences the cDNA library identified by sequencing. The matched sequences are then selected. Accordingly, step g. comprises selecting sequences identified in step c. that correlate with the mass spectrometry data and step h comprises identifying sequences of CDR3 regions in the sequences from step g.

Step i. comprises selecting from the CDR3, CDR2 and/or CDR1 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage. In some embodiments, the fragmentation coverage percentage is equal to or more than about 30% (for example, about 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 99%) for trypsin-treated samples. In some embodiments, the fragmentation coverage percentage is equal to or more than about 30% (for examples, at least about 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 99%) for chymotrypsin-treated samples. In some embodiments, the fragmentation coverage percentage is about 50% for trypsin-treated samples and about 40% for chymotrypsin-treated samples.

In some embodiments, the method described herein further comprises creating a nanobody comprising a CDR3, CDR2 and/or CDR1 region having a sequence identified in step i. The nanobody genes are cloned into a vector, which is then transformed into competent cells for nanobody protein expression, extraction and purification.

In some embodiments, the nanobody comprises an amino acid sequence at least 80% (for examples, at least about 80%, 85%, 90%, 95%, 98% or 99%) identical to a sequence selected from the group consisting of SEQ ID NOs: 1-157. In some embodiments, the nanobody has a sequence selected from the group consisting of SEQ ID NOs: 1-157. In some embodiments, the nanobody comprises an amino acid sequence at least 80% (for examples, at least about 80%, 85%, 90%, 95%, 98% or 99%) identical to a sequence selected from the group consisting of SEQ ID NOs: 158-2536. In some embodiments, the nanobody has a sequence selected from the group consisting of SEQ ID NOs: 158-2536. In some embodiments, the nanobody comprises an amino acid sequence at least 80% (for examples, at least about 80%, 85%, 90%, 95%, 98% or 99%) identical to a sequence selected from the group consisting of SEQ ID NOs: 2665-2667. In some embodiments, the nanobody has a sequence selected from the group consisting of SEQ ID NOs: 2665-2667.

Disclosed herein is a PDZ-specific nanobody, wherein the PDZ-specific nanobody comprises an amino acid sequence selected from the group consisting of SEQ ID NOs: 158-2536. Also disclosed herein is a PDZ-specific nanobody, wherein the PDZ-specific nanobody comprises an amino acid sequence selected from the group consisting of SEQ ID NOs: 143-157. As used herein, “PDZ” refers to an 80-100 amino acid domain found in signaling proteins that have also been referred to as DHR (Dlg homologous region) or GLGF (glycine-leucine-glycine-phenylalanine) domains. PDZ domains bind to a short region of the C-terminus of other specific proteins. PDZ domains are conventionally divided into three different classes, categorized by the chemical nature of their ligands. Different ligand classes are distinguished by differences in the penultimate binding residues found at the extreme COOH of target proteins. Type I domains recognize the sequence, X-S/T-X-Φ* (where X= any amino acid, Φ = hydrophobic amino acid, * COOH terminus). Type II domains bind to ligands with the sequence X-Φ-X-Φ*. Type III domains interact with sequences with X-X-C*. Binding specificity within each domain class can be conferred by the variant (X) residues as well as residues outside the canonical binding motif. Moreover, a few PDZ domains do not fall into any of these specific classes. Proteins that contain PDZ domains include, but are not limited to, Erbin, GRIP, Htra1, Htra2, Htra3, PSD-95, SAP97, CARD10, CARD11, CARD14, PTP-BL, and SYNJ2BP. In some embodiments, the PDZ domain is from SYNJ2BP.

Disclosed herein is a GST-specific nanobody, wherein the GST-specific nanobody comprises an amino acid sequence in Table 4. Also disclosed herein is a GST-specific nanobody, wherein the GST-specific nanobody comprises an amino acid sequence selected from the group consisting of SEQ ID NOs: 1-98. “Glutathione S-transferase” or “GST” refers herein to glutathione-S-transferases (GSTs) are a family of Phase II detoxification enzymes that catalyze the conjugation of glutathione (GSH) to a wide variety of endogenous and exogenous electrophilic compounds. In some embodiments, the GST polypeptide is that in the pGEX6p-1 vector.

Disclosed herein is a HSA-specific nanobody, wherein the HSA-specific nanobody comprises an amino acid sequence in Table 5. Also disclosed herein is a HSA-specific nanobody, wherein the HSA-specific nanobody comprises an amino acid sequence selected from the group consisting of

SEQ ID NOs: 99-142. "Human serum albumin" or "HSA" refers herein to a polypeptide encoded by the *ALB* gene. In some embodiments, the HSA polypeptide is that identified in one or more publicly available databases as follows: HGNC: 399, Entrez Gene: 213, Ensembl: ENSG00000163631, OMIM: 103600, UniProtKB: P02768. In some embodiments, the HSA polypeptide comprises the sequence of SEQ ID NO: 2668, or a polypeptide sequence having at or greater than about 80%, about 85%, about 90%, about 95%, or about 98% homology with SEQ ID NO: 2668, or a polypeptide comprising a portion of SEQ ID NO: 2668. The HSA polypeptide of SEQ ID NO: 2668 may represent an immature or pre-processed form of mature HSA, and accordingly, included herein are mature or processed portions of the HSA polypeptide in SEQ ID NO: 2668.

Here a robust proteomic pipeline was developed for large-scale quantitative analysis of antigen-engaged Nb proteomes and epitope mapping based on high-throughput structural characterization of antigen-Nb complexes.

EXAMPLES

Example 1. The superiority of chymotrypsin for large-scale Nb proteomics analysis.

The variable domains of HcAb (V_HH/Nb) cDNA libraries were amplified from the B lymphocytes of two *lama glamas*, recovering 13.6 million unique Nb sequences in the databases by the next-generation genomic sequencing (NGS) (DeKosky, 2013). Approximately half a million Nb sequences were aligned to generate the sequence logo (**FIG. 1A, 7A**). CDR3 loops have both the largest sequence diversity and length variation providing excellent specificity for Nb identifications (**FIG. 1B, 1C**). *In silico* analysis of Nb databases revealed that trypsin predominantly produced large CDR3 peptides due to the limited number of trypsin cleavage sites on Nbs (**FIG. 1A**). As a result, the majority of the CDR3 residues (77%) were covered by large tryptic peptides of more than 2.5 kDa (**FIG. 1D, 1E**), which are suboptimal for proteomic analysis (**FIG. 7B**). In comparison, chymotrypsin, which is infrequently used for proteomics cleaving specific aromatic and hydrophobic residues, appears to be more suitable (**Methods, FIG. 1A**, **7B**). 91% of CDR3 sequences can be covered by chymotryptic peptides less than 2.5 kDa (**FIG. 1D, 1E**). Random selection and simulation confirmed that significantly more CDR3 sequences can be covered by chymotrypsin than trypsin (**FIG. 1F**). Moreover, there was a small overlap (~9%) between the two enzymes, indicating their good complementarity for efficient Nb analysis.

The estimated false discovery rate (FDR) of CDR3 identifications can be inflated due to the large database size and the unusual Nb sequence structure. To test this, antigen-specific HcAbs were proteolyzed with trypsin or chymotrypsin, and a state-of-the-art search engine was employed for identification using two different databases: a specific "target" database derived from the immunized llama, and a "decoy" database of similar size from an irrelevant llama with literally no identical sequences (**FIG. 7D**). Any CDR3 peptides identified from the decoy database search were thus considered as false positives (Elias, J.E. & Gygi, S.P, 2007). A large number of false positive CDR3 peptides were nonspecifically identified from the decoy database search. It was found that these spurious peptide-spectrum-matches generally contained poor MS/MS fragmentations on the CDR3 fingerprint sequences (**FIG. 7E, 7F**). The vast majority (95%) of these erroneous matches can be removed by using a simple fragmentation filter that we have implemented, requiring a minimum coverage of 50% (by trypsin, **FIG. 1G**) and 40% (by chymotrypsin, **FIG. 1H**) of the CDR3 high-resolution diagnostic ions in the MS2 spectra (**FIG. 1K, 1L**). The filter was further optimized based on the CDR3 length (**FIG. 1I, 1J**) before integrating into the new, open-source software "Augur Llama" (**FIG. 8A-8C**) for reliable Nb proteomic analysis.

Example 2. Development of an integrative proteomics pipeline for Nb discovery and characterization.

A robust platform is shown herein for comprehensive quantitative Nb proteomics and high-throughput structural characterizations of antigen-Nb complexes (**Methods, FIG. 2A**). A domestic camelid was immunized with the antigens of interest. The Nb cDNA library was then prepared from the blood and/or bone marrow of the immunized camelid (Fridy, 2014). NGS was performed to create a rich database of >10^7 unique Nb protein sequences **(FIG. 8E, 8F)**. Meanwhile, antigen-specific V_HHs were affinity isolated from the sera and eluted using step-wise gradients of salts or pH buffers. Fractionated HcAbs were efficiently digested with trypsin or chymotrypsin to release Nb CDR peptides for identification and quantification by nanoflow liquid chromatography coupled to high-resolution MS. Initial candidates that pass database searches were annotated for CDR identifications. CDR3 fingerprints were filtered to remove false positives, their abundances from different biochemical fractions were quantified to infer the Nb affinities, and assembled into Nb proteins – all of the above steps were automated by Augur Llama. The pipeline enables identification and characterization of an unprecedented scale of diverse, specific, and high-quality Nbs. In parallel, to enable structural analysis of tens of thousands of antigen-Nb interactions, a robust method have been developed to integrate high-throughput computational docking (Schneidman-Duhovny, 2005), cross-linking and mass spectrometry (CXMS) (Chait, 2016; Rout, 2019; Yu, 2018; Leitner, 2016), and mutagenesis. A deep-learning approach was further developed to learn the latent features associated with the Nb repertoires.

Example. 3. Robust, in-depth, and high-quality identifications of antigen-specific Nbs.

To validate this pipeline, three benchmark antigens were chosen: glutathione S-transferase (GST), human serum albumin (HSA)- an important drug target (Larsen, 2016), and a small PDZ domain derived from mitochondrial outer membrane protein 25. These antigens span three orders of magnitude of immune responses with PDZ only weakly immunogenic (**FIG. 2B**) and are ideal to assess the robustness of our technologies.

Here 64,670 unique Nb_{GST} sequences (9,915 unique CDR combinations from 3,453 CDR3 Nb families), 34,972 unique Nb_{HSA} (7,749 unique CDRs from 2,286 unique CDR3 Nb families) and a smaller cohort of 2,379 high-quality Nb_{PDZ} sequences (495 unique CDRs from 230 CDR3 families) were identified (**Methods, FIG. 2C, 8G**). It was confirmed that chymotrypsin provided the most useful fingerprint information for Nb identification from the various proteases tested (**FIG. 2D, 2E**). The Nb repertoires exhibited exceptional CDR3 diversity (**FIG. 8D**).

A random set of 146 Nbs was selected from among the three antigen-specific Nb groups and expressed in *E.coli.* A group of 130 Nbs (89%) exhibited excellent solubility and can be readily purified in large quantities (**FIG. 2F**). Complementary approaches were taken, including immunoprecipitation, ELISA, and SPR, to evaluate the antigen binding (**Methods, FIG. 2G, 9C, 9D, 10, Tables 1-3**). Nbs identified by trypsin and chymotrypsin were comparably high-quality (**FIG. 8H**). 86.2% ($CI_{95\%}$: 6.8%), 90.5% ($CI_{95\%}$: 11.5%), and 100% true Nb binders were confirmed for GST, HSA and PDZ, respectively. These results demonstrate the high sensitivity and specificity of this approach.

Example 4. Accurate large-scale quantification and clustering of Nb proteomes.

Different strategies were evaluated for accurate classification of Nbs based on affinities. Briefly, antigen-specific HcAbs were affinity isolated from the serum and eluted by the step-wise high-salt gradients, high pH buffers, or low pH buffers (**Methods, FIG. 8I, 8J**). Different HcAbs fractions were accurately quantified by label-free quantitative proteomics (Zhu, 2010; Cox, J. & Mann, M, 2008). The CDR3 peptides (and the corresponding Nbs) were then clustered into three groups based on their relative ion intensities (**FIG. 3A, 3B, 9A, and 9B**). This classification assigns 31% of Nb_{GST} and 47% of Nb_{HSA} into the C3 high affinity group by the high pH method (**FIG. 3C**). A number of Nb_{GST} with unique CDR3 sequences from each cluster were randomly expressed and their affinities were measured by ELISA and SPR ($R^2 = 0.85$, **FIG. 3D, Table 1**) to evaluate different fractionation methods. While the low pH method did not provide sufficient resolution to separate different affinity groups, the salt gradient and particularly the high pH method, enabled significant and reproducible separations of Nbs based on their affinities (**FIG. 3E**). Nbs from high pH clusters

1 and 2 (C1, C2) generally have low and mediocre affinities, respectively, from μM to dozens of nM, while over 50% of C3 were ultrahigh affinity, sub-nM binders (**FIG. 3H, 9D**). To further verify this result, a random set of 25 Nb_{HSA} (with divergent CDR3s) were purified from C3, and ranked their ELISA affinities (**FIG. 3F**, **Table 2**). The top 14 Nb_{HSA} were selected for SPR measurements, in which 11 have dozens to hundreds of pM affinities with diverse binding kinetics. The remaining 3 Nb_{HSA} demonstrated single-digit nM K_D's. (**FIG. 3I, 10A**). 13 soluble Nb_{PDZ} were purified and their high affinities were confirmed by ELISA and immunoprecipitation (**FIG. 3G**, **10B, and Table 3**). The K_D of a representative, highly soluble Nb_{PDZ} P10 was 4.4 pM (**FIG. 3J**).

The ultrahigh affinity Nbs for immunoprecipitation (Nb_{GST}) and fluorescence imaging (Nb_{PDZ}) of native mitochondria (**FIG. 3K, 3L**) were further positively evaluated. The quantitative approach enables large-scale and accurate classification of Nb proteomes based on desirable properties such as affinities.

Example 5. The landscapes of antigen-engaged Nb proteomes revealed by integrative structure determination methods.

Identification and classification of large repertoires of high-quality Nbs allow to the investigation on the global structure landscapes of antigen-engaged humoral immune response. Structural docking and clustering of 34,972 Nb_{HSA} revealed three dominant HSA epitopes (**FIG. 4A**). The presence of abundant native serum albumin (76% identical to HSA, **FIG. 12H**) allowed the investigation on the specificity of the camelid humoral immunity. The two albumin sequences were aligned and their variations were calculated based on pI and hydropathy (**Methods, FIG. 4A**). All three epitopes are co-localized with the major peaks of pI and hydropathy which correspond to the large sequence differences. This result illustrates the exceptional specificity of antigen recognition by Nbs. It appears that Nbs preferentially bind stable helical secondary structures (**FIG. 4B**). It was found that the epitopes were highly charged. E2 and E3 were predominantly negative (-4 and -5 net formal charges respectively, **FIG. 13D**), while E1 was more heterogeneous with mixed charges -2 net formal charges) (**FIG. 4C**).

19 HSA-Nb complexes (Shi, 2014; Kim, 2018) were cross-linked to verify the epitopes identified by docking. Overall, 92% of cross-links were satisfied by the models, which have a median RMSD of 5.6 Å (**FIG. 4J, 4K**). Cross-linking confirmed the docking results and identified two epitopes (E2, E3) that were heavily populated (65% and 20%, respectively) (**FIG. 4D, Table 2**). E1 was identified by cross-links with low abundance (5%). Cross-linking also identified additional two minor epitopes that were not revealed by docking (**FIG. 4D**). High shape complementarity was observed between HSA and Nbs involving convex Nb paratopes and concave HSA epitopes (**FIG.**

4E – 4G). To further confirm the dominant E2, we introduced a single point mutation on HSA, E400R with minimal impact on the overall structure (Pires, 2016). The resulting mutation reverses the surface charge to mimic the positive charge at the orthologous position in E2 of camelid albumin, potentially disrupting a salt bridge formed between it and an arginine in the Nb CDR3 (**FIG. 4H**). 19 high-affinity binders were then selected and this point mutation on HSA-Nb interactions was evaluated by ELISA (**FIG. 4I**, **Table 2**). E400R almost completely abolished the binding of 5 out of 19 Nbs (26%) that were tested, indicating that E2 is a *bona fide* major epitope.

This approach was further employed to map the epitopes of 64,670 GST-Nb complexes. Three major epitopes on GST were accurately identified (**FIG. 11A, 11B, 11F, 11G**) and were verified by cross-links with relative abundances of 18.75%, 31.25%, and 50% for E1, E2, and E3, respectively (**FIG. 11D, 11E**). E1 and E3 contain negatively charged surface patches. E2 overlapped with GST dimerization cavity (**FIG. 11C**); in the models shown herein E2 Nbs insert their CDR3s into this cavity. Similar to HSA, preference to charged surface residues and high shape complementarity of Nbs were confirmed. Together, these results indicate that Nbs can bind diverse protein surfaces and prefer highly charged cavities on the antigen.

Example 6. Exploring the mechanisms of Nb affinity maturation.

The physicochemical and structural features that distinguish high-affinity (matured) and low-affinity Nbs were investigated, based on the high pH dataset that was most reliably classified. Shorter CDR3s with distinct distributions for high-affinity binders for HSA and GST, respectively (**FIG. 5A**), lowering the entropy for antigen binding. A significant increase of pI was observed (**FIG. 5B**), from slightly acidic for low-affinity to relatively basic for high-affinity Nbs.

The contribution of CDRs to pI and hydropathy of the Nbs were compared, and it was determined that $CDR3_{HSA}$ was primarily responsible for polarity shifts in Nb_{HSA} while $CDR1_{GST}$ and $CDR2_{GST}$ were primarily responsible for polarity shifts in Nb_{GST} (**FIG. 5C**). It was observed that high-affinity Nbs are slightly more hydrophilic (**FIG. 5D**).

The structure of a CDR3 can be considered as having a “head” region consisting of the highest sequence variability, and a “torso” region of lower specificity (Finn, 2016) (**FIG. 5E**). Certain residues were enriched on CDR3 heads, including aspartic acid and arginine (forming strong electrostatic interactions) (Tiller, 2017), small and flexible residues of glycine and serine, hydrophobic residues such as alanine and leucine, and aromatic residue of tyrosine (**FIG. 5F, and FIG. 12**). Nbs of different affinity groups were compared and three major differences were found. First, high-affinity Nbs were more enriched with charged residues (Mitchell, L.S. & Colwell, L.J, 2018) (**Methods, FIG. 5G**). Second, intricate differences were identified for different antigens: high-affinity Nb_{HSA} tend to strengthen the electrostatics by increasing positively charged residues (39%) and decreasing (46%) negatively charged residues on the CDR3 heads. High-affinity Nb_{GST} predominantly altered their charges on other CDRs. Increases of 29.2% and 117.2% of positively charged residues and decreases of 44.2% and 21.5% of negatively charged residues were found on CDR1 and CDR2, respectively. The changes in charge may increase the physicochemical complementarity between the Nb and the epitope. Third, tyrosine (51%), glycine and serine (58%) were more enriched on CDR3 heads for high-affinity Nb_{HSA}. For high-affinity Nb_{GST}, there was an increase in tyrosines (73%) in CDR3 heads but the fractions of glycine and serine were hardly affected.

To further explore the putative roles of these residues for augmenting HSA binding affinity, their location frequency was calculated along the CDR3 heads (**FIG. 5H**). Tyrosine is more frequently found at the center of CDR3 heads for high-affinity Nb_{HSA} enabling its bulky, aromatic side chain to insert into specific epitope pocket(s) (Desmyter, 1996; Li, 2016). Glycine and serine tend to be placed away from the CDR3 center, providing additional flexibilities and facilitating the orientation of the tyrosine side chain in the antigen pocket. These results were confirmed by the correlation analysis between the number of these residue groups and ELISA affinities of our purified Nbs (**FIG. 5I, 5J**).

A deep learning model was developed to learn the latent features that enable Nb affinity classification (**Methods**). The most informative Nb_{HSA} CDR3 filter for high-affinity binder classification revealed a pattern of consecutive lysine and arginine, tyrosines and glycines (**FIG. 5K, Table 4**). For low-affinity binders, the most informative filter has preference for phenylalanine, histidine, and two consecutive aspartic acids. Moreover, this analysis revealed a tendency for consecutive pairs of negative and positive charges for high- and low- affinity binders, respectively.

Example 7. The outstanding versatility and resilience of Nbs for antigen recognition.

Identification of hundreds of divergent, high-affinity Nb_{CDR3} families for the weakly immunogenic PDZ domain prompted the investigation of the structural basis of such interactions. Two putative epitopes were identified based on docking (**FIG. 6A, 13B**). E2 can be the major epitope because it has a large positively charged surface (**FIG. 6A, 6B**) and it is more structured with an α helix and two β-strands. E2 overlapped with the conserved ligand binding sites that are shared among numerous PDZ interacting proteins (Sheng, 2001; Doyle, 1996) (**FIG. 6C**). Remarkably, Nb_{PDZ} have obtained >100,000-fold higher affinity than natural PDZ ligands (in μM affinity) (Niethammer, 1998) (**FIG. 3J**). Such high affinity likely was achieved by a long CDR3 loop wrapping around the small and shallow epitope, forming extensive electrostatic and hydrophobic interactions (**FIG. 6C, 13A**).

Modeling results indicated that R46 and K48 of the second β strand in the PDZ epitope formed salt bridges with the corresponding residues in Nb_{PDZ}. A double mutant PDZ (R46E:K48D) was produced and its affinity was evaluated to Nb_{PDZ} by ELISA. The majority (8/11) of Nb_{PDZ} exhibited significantly decreased or no affinity for the mutant, confirming that E2 is indeed the major epitope (**FIG. 6D**).

There are several other observations on Nb_{PDZ}. First, the distribution of CDR3 loop length formed one major peak with a median of ~20 aa that pushed the upper limit of its natural distribution (**FIG. 6E**). Second, Nb_{PDZ} are rather acidic with a median pI of 4.9 (**FIG. 6F**), which is largely contributed by CDR3 (**FIG. 6E**, **13F**). Third, despite their acidic nature, Nb_{PDZ} did not seem to appreciably alter hydropathy, due to the compensation of hydrophobic residues (**FIG. 6G**, **13E**). Finally, there were significant increases of negatively charged aspartic acid and small glycines and serines, accounting for half of the CDR3 head residues; decrease of bulky tyrosine was also evident compared with high-affinity Nb_{GST} and Nb_{HSA} reflecting the rather shallow pocket of E2 for binding (**FIG. 7C, 7E**). Collectively, these results demonstrated a remarkable versatility of Nbs for antigen binding.

This study reports the development of a robust platform integrating proteomics, informatics, and structural modeling technologies for analysis of antigen-engaged Nb proteomes. The pipeline enables sensitive and reliable identification of a large repertoire of high-quality Nbs against different challenging antigens. It also enables accurate classification of circulating Nbs based on their physicochemical properties. Thousands of ultrahigh-affinity Nbs were identified by our technologies. Combining computational docking and structural proteomics, the present study have structurally characterized 102,673 antigen-Nb complexes, mapped, and validated the dominant epitopes. This “big data” analysis permits for the first time, global-scale proteomic and structural dissections of the humoral immune response.

These results revealed, at unprecedented depth, the efficiency, specificity, diversity, and versatility of antigen-engaged Nbs that together shape the epic landscapes of camelid antibody immunity (**FIG. 6H**).

Efficiency: Nbs efficiently utilize both shape and electrostatic complementarity for binding. Specific residues such as charged aspartic acids and arginines, aromatic tyrosines, and small, flexible glycines and serines permit loop flexibility that result in high-affinity Nbs. Intricate and fine-tuned interactions specific for different CDRs were revealed. Moreover, the presence of multiple dominant epitope for Nb binding was confirmed, which can act as a general mechanism for efficiently recognizing pathogens (Akram, A. & Inman, R.D, 2012).

Specificity and Diversity: Thousands of highly divergent Nbs were discovered that evolved to recognize specific HSA surface pockets with some of the most pronounced sequence variations (**FIG. 4A**) to ensure a specific, effective, and safe immune response.

Versatility: for antigens that tend to evade immune response such as the PDZ, Nbs can drastically alter the size and the physicochemical properties of paratopes to mimic natural ligand binding with outstanding affinity and specificity. The study shows the fascinating rapid evolution of protein-protein interactions.

Nbs are highly potent in viral neutralization and inhibition of enzymatic activities (Lauwereys, 1998; Desmyter, 1996; Acharya, 2013; Arabi, 2017). These findings indicate that these highly robust and efficient camelid HcAbs are evolutionarily advantageous for their survival in both arid natural habitats and aggressive pathogenic challenges, while the driving force(s) behind such an incredible selection and adaptation remains enigmatic (Flajnik, 2011).

These technologies can find broad utility in challenging biomedical applications such as cancer biology, brain research, and virology. These informatics tools for Nb proteomics can be freely available to the research community. The high-quality Nb datasets can serve as a blueprint to study antibody-antigen and can facilitate computational antibody design (Sircar, 2011; Baran, 2017; Chevalier, 2017).

Example 8. Methods

Animal immunization. Two Llamas were respectively immunized with HSA, and a combination of GST and GST fusion PDZ domain of Mitochondrial outer membrane protein 25 (OMP25) at the primary dose of 1 mg, followed by three consecutive boosts of 0.5 mg every 3 weeks. The bleed and bone marrow aspirates were extracted from the animals 10 days after the last immuno-boost. All the above procedures were performed by Capralogics, Inc. following the IACUC protocol.

mRNA isolation and cDNA preparation. Approximately 1 - 3 x10^9 peripheral mononuclear cells were isolated from 350 ml immunized blood and 5 - 9 x10^7 plasma cells were isolated from 30 ml bone marrow aspirates using Ficoll gradient (Sigma). The mRNA was isolated from the respective cells using RNeasy kit (NEB) and was reverse-transcribed into cDNA using Maxima™ H Minus cDNA Synthesis Master Mix (Thermo). Camelid IgG heavy chain cDNA sequences from the variable domain to the CH2 domain were specifically amplified using primers CALL001 (GTCCTGGCTGCTCTTCTACAAGG, SEQ ID NO: 2646) and CH2FORTA4 (CGCCATCAAGGTACCAGTTGA, SEQ ID NO: 2647) (Abrabi, 1997). The V_HH genes that lack CH1 domain were separated from conventional IgG and purified (Qiagen) by DNA gel electrophoresis, and were subsequently re-amplified from framework 1 to framework 4 using the

2nd-Forward (ATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNATGGCT[C/G]A[G/T]GTGCAGCTGGTGGAGTCTGG, SEQ ID NO: 2648, wherein N represents A, T, C or G) and 2nd-Reverse (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNGGAGACGGTGACCTGGGT, SEQ ID NO: 2649, wherein N represents A, T, C or G). The random 8-mers replacing adaptor sequences were added to aid in cluster identification for Illumina MiSeq. The amplicon of the second PCR (approximately 450-500 bp) was purified using Monarch PCR clean up kit (NEB). The final round of PCR with primer MiSeq-F (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTA, SEQ ID NO: 2650) and MiSeq-R (CAAGCAGAAGACGGCATACGAGATTTCTGAATGTGACTGGAGTTCA, SEQ ID NO: 2651) was performed to add P5/P7 adapters with the index before MiSeq sequencing.

Next generation sequencing by Illumina Miseq. Sequencing was performed based on the Illumina MiSeq platform with the 300 bp paired-end model. More than 30 million reads were generated for each database. Read QC tool in FastQC v0.11.8 (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality check and control of the FASTQ data. Raw Illumina reads were processed by the software tools from the BBMap project (github.com/BioInfoTools/BBMap/). Duplicated reads and DNA barcode sequences were removed successively before converting the nucleotide sequences into amino acid sequences.

Isolation and biochemical fractionation of V_HH antibodies from immunized sera. Approximately 175 ml of plasma was isolated from 350 ml of immunized blood by Ficoll gradient (Sigma). Camelid single-chain V_HH antibodies were isolated from the plasma supernatant by a two-step purification procedure using protein G and protein A sepharose beads (Marvelgent), acid-eluted, before neutralized and diluted in 1xPBS buffer to a final concentration of 0.1- 0.3 mg/ml. To purify antigen-specific V_HH antibodies, the GST or HSA-conjugated CNBr resin was incubated with the V_HH mixture for 1 hr at 4°C and extensively washed with high salt buffer (1xPBS and 350 mM NaCl) to remove non-specific binders. Specific V_HH antibodies were then released from the resin by using one of the following elution conditions: alkaline (1-100 mM NaOH, pH 11, 12 and 13), acidic (0.1 M glycine, pH 3, 2 and 1) or salt elution (1M – 4.5 M $MgCl_2$ in neutral pH buffer). For purification of PDZ-specific V_HH, a fusion protein of MBP-PDZ (where the maltose binding protein/MBP was fused to the N terminus of PDZ domain to avoid steric hindrance of the small PDZ after coupling) was produced and was used as the affinity handle. MBP coupled resin was used for control **(FIG.**

6J). All the eluted V_HHs were neutralized and dialyzed into 1x DPBS separately prior to proteomics analysis.

Proteolysis of Antigen Specific Nbs and Nanoflow Liquid Chromatography coupled to Mass spectrometry (nLC/MS) Analysis. For GST and HSA V_HHs, each elution was processed separately according to the following protocol. For PDZ specific V_HHs, only the most stringent biochemical elutes (i.e., pH 13, pH 1, $MgCl_2$ 3M and 4.5M) and the respective nonspecific MBP binders (negative controls) from different fractions were pooled for proteolysis. For instance, For PDZ-specificV_HHs that were eluted by pH13 buffer, non-specific MBP binding Nbs were pooled from pH 11, pH12 and pH13 fractions for negative control to improve the stringency of our downstream LC/MS quantification. V_HHs were reduced in 8M urea buffer (with 50 mM Ammonium bicarbonate, 5 mM TCEP and DTT) at 57°C for 1hr, and alkylated in the dark with 30 mM Iodoacetamide for 30 mins at room temperature. The alkylated sample was then split into two and in-solution digested using either trypsin or chymotrypsin. For trypsin digestion samples, 1:100 (w/w) trypsin and Lys-C were added and digested at 37°C overnight, with additional 1:100 trypsin the other morning for 4 hrs at 37°C water bath. For chymotrypsin digestion samples, 1:50 (w/w) chymotrypsin was added and digested at 37 °C for 4 hrs. After proteolysis, the peptide mixtures were desalted by self-packed stage-tips or Sep-pak C18 columns (Waters) and analyzed with a nano-LC 1200 that is coupled online with a Q Exactive™ HF-X Hybrid Quadrupole Orbitrap™ mass spectrometer (Thermo Fisher). Briefly, desalted Nb peptides were loaded onto an analytical column (C18, 1.6 μm particle size, 100 Å pore size, 75 μm × 25 cm; IonOpticks) and eluted using a 90-min liquid chromatography gradient (5% B–7% B, 0–10 min; 7% B–30% B, 10–69 min; 30% B–100% B, 69 – 77 min; 100% B, 77 - 82 min; 100% B - 5% B, 82 min - 82 min 10 sec; 5% B, 82 min 10 sec - 90 min; mobile phase A consisted of 0.1% formic acid (FA), and mobile phase B consisted of 0.1% FA in 80% acetonitrile (ACN)). The flow rate was 300 nl/min. The QE HF-X instrument was operated in the data-dependent mode, where the top 12 most abundant ions (mass range 350 – 2,000, charge state 2 - 8) were fragmented by high-energy collisional dissociation (HCD). The target resolution was 120,000 for MS and 7,500 for tandem MS (MS/MS) analyses. The quadrupole isolation window was 1.6 Th and the maximum injection time for MS/MS was set at 80 ms.

Nb DNA synthesis and cloning. Nb genes were codon-optimized for expression in *Escherichia coli* and the nucleotides were *in vitro* synthesized (Synbiotech). After verification by Sanger sequencing, the Nb genes were cloned into a pET-21b (+) vector at BamHI and XhoI (for GST Nbs), or EcoRI and NotI restriction sites (for HSA and PDZ Nbs).

Purification of recombinant Proteins. DNA constructs were transformed into BL21 (DE3) competent cells according to manufacturer's instructions and plated on Agar with 50 µg/ml ampicillin at 37 °C overnight. A single colony was inoculated in LB medium with ampicillin for overnight culture at 37 °C. The culture was then inoculated at 1:100 (v/v) in fresh LB medium and shaked at 37 °C until the O.D.600 nm reached 0.4-0.6. GST, GST-PDZ and Nbs were induced with 0.5 mM of IPTG while MBP and MBP-PDZ were induced with 0.1 mM of IPTG. The inductions were performed at 16°C overnight. Cells were then harvested, briefly sonicated and lysed on ice with a lysis buffer (1xPBS, 150 mM NaCl, 0.2% TX-100 with protease inhibitor). After lysis, soluble protein extract was collected at 15,000 x g for 10 mins. GST and GST-PDZ were purified using GSH resin and eluted by glutathione. MBP (maltose binding protein) and MBP-PDZ fusion protein were purified by using Amylose resin and were eluted by maltose according to the manufacturer's instructions. Nbs were purified by His-Cobalt resin and were eluted using imidazole. The eluted proteins were subsequently dialyzed in the dialysis buffer (e.g., 1x DPBS, pH 7.4) and stored at -80 °C before use.

Nb immunoprecipitation assay. After Nb induction and cell lysis, the cell lysates were run on SDS-PAGE to estimate Nb expression levels. Recombinant Nbs in the cell lysis were diluted in 1x DPBS (pH 7.4) to a final concentration of ~ 5 µM (for GST Nbs) and ~ 50 nM (for PDZ Nbs). To test the specific interactions of Nbs with antigens, different antigens were coupled to the CNBr resin. Inactivated or MBP-conjugated CNBr resin was used for control. Antigen coupled resins or control resins were incubated with Nb lysates at 4°C for 30 mins. The resins were then washed three times with a washing buffer (1x DPBS with 150 mM NaCl and 0.05% Tween 20) to remove nonspecific bindings. Specific antigen bound Nbs were then eluted from the resins by the hot LDS buffer containing 20 mM DTT and ran on SDS-PAGE. The intensities of Nbs on the gel were compared between antigen specific signals and control signals to derive the false positive binding.

ELISA (enzyme-linked immunosorbent assay). Indirect ELISA was carried out to evaluate the camelid immune response of an antigen and to quantify the relative affinities of antigen-specific Nbs. An antigen was coated onto a 96-well ELISA plate (R&D system) at an amount of approximately 1-10 ng per well in a coating buffer (15 mM sodium carbonate, 35 mM sodium bicarbonate, pH 9.6) overnight at 4°C. The well surface was then blocked with a blocking buffer (DPBS, 0.05% Tween 20, 5% milk) at room temperature for 2 hours. To test an immune response, the immunized serum was serially 5-fold diluted in the blocking buffer. The diluted sera were incubated with the antigen coated wells at room temperature for 2 hours. HRP-conjugated secondary antibodies against llama Fc (Bethyl) were diluted 1:10,000 in the blocking buffer and incubated with each well for 1 hour at room temperature. For Nb affinity tests, scramble Nbs that do not bind the antigen of interest were used for negative controls. Nbs of both specific binders for test and scramble negative controls were serially 10-fold diluted from 10 μM to 1 pM in the blocking buffer. HRP-conjugated secondary antibodies against His-tag (Genscript) or T7-tag (Thermo) were diluted 1:5,000 or 1:10,000 in the blocking buffer and incubated for 1 hour at room temperature. Three washes with 1x PBST (DPBS, 0.05% Tween 20) were carried out to remove nonspecific absorbance between incubations. After the final wash, the samples were further incubated under dark with freshly prepared w3,3′,5,5′-Tetramethylbenzidine (TMB) substrate for 10 mins at room temperature to develop the signals. After the STOP solution (R&D system), the plates were read at multiple wavelengths (450 nm and 550 nm) on a plate reader (Multiskan GO, Thermo Fisher). A false positive Nb binder was defined if any of the following two criteria was met: i) the ELISA signal can only be detected at a concentration of 10 μM and was under detected at 1 μM concentration. ii) At 1 μM concentration, a pronounced signal decrease (by more than 10-fold) was detected compared to the signal at 10 μM, while there were no signals can be detected at lower concentrations. The raw data was processed by Prism 7 (GraphPad) to fit into a 4PL curve and to calculate logIC50.

Nb affinity measurement by SPR. Surface plasmon resonance (SPR, Biacore 3000 system, GE Healthcare) was used to measure Nb affinities. Antigen proteins immobilized on the activated CM5 sensor-chip by the following steps. Protein analytes were diluted to 10-30 μg/ml in 10 mM sodium acetate, pH 4.5, and were injected into the SPR system at 5 μl/min for 420 s. The surface of the sensor was then blocked by 1 M ethanolamine-HCl (pH 8.5). For each Nb analyte, a series of dilution (spanning three orders of magnitude) was injected in HBS-EP+ running buffer (GE-Healthcare) containing 2 mM DTT, at a flow rate of 20- 30 μl/min for 120- 180 s, followed by a dissociation time of 5 – 20 mins based on dissociation rate. Between each injection, the sensor chip surface was regenerated with the low pH buffer containing 10 mM glycine-HCl (pH 1.5- 2.5), or high pH buffer of 20-40 mM NaOH (pH 12- 13). The regeneration was performed with a flow rate of 40-50 μl/min for 30 s. The measurements were duplicated and only highly reproducible data was used for analysis. Binding sensorgrams for each Nb were processed and analyzed using BIAevaluation by fitting with 1:1 Langmuir model or 1:1 Langmuir model with mass transfer.

Cross-linking and mass spectrometric analysis of antigen-nanobody complex. Different Nbs were incubated with the antigen of interest with equal molarity in an amine-free buffer (such as 1x DPBS with 2 mM DTT) at 4°C for 1 - 2 hours before cross-linking. The amine-specific disuccinimidyl suberate (DSS) or heterobifunctional linker 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) was added to the antigen-Nb complex at 1 mM or 2 mM final concentration, respectively. For DSS cross-linking, the reaction was performed at 23°C for 25 mins with constant agitation. For EDC cross-linking, the reaction was performed at 23°C for 60 mins. The reactions were quenched by 50 mM Tris-HCl (pH 8.0) for 10 mins at room temperature. After protein reduction and alkylation, the cross-linked samples were separated by a 4–12% SDS-PAGE gel (NuPAGE, Thermo Fisher). The regions corresponding to the cross-linked species were cut and in-gel digested with trypsin and Lys-C as previously described (Shi, 2014; Shi, 2015). After proteolysis, the peptide mixtures were desalted and analyzed with a nano-LC 1200 (Thermo Fisher) coupled to a Q Exactive™ HF-X Hybrid Quadrupole-Orbitrap™ mass spectrometer (Thermo Fisher). The cross-linked peptides were loaded onto a picochip column (C18, 3 μm particle size, 300 Å pore size, 50 μm x 10.5 cm; New Objective) and eluted using a 60 min LC gradient : 5% B–8% B, 0 – 5 min; 8% B – 32% B, 5 – 45 min; 32% B–100% B, 45 – 49 min; 100% B, 49 - 54 min; 100% B - 5 % B, 54 min - 54 min 10 sec; 5% B, 54 min 10 sec - 60 min 10 sec; mobile phase A consisted of 0.1% formic acid (FA), and mobile phase B consisted of 0.1% FA in 80% acetonitrile. The QE HF-X instrument was operated in the data-dependent mode, where the top 8 most abundant ions (mass range 380–2,000, charge state 3 - 7) were fragmented by high-energy collisional dissociation (normalized collision energy 27). The target resolution was 120,000 for MS and 15,000 for MS/MS analyses. The quadrupole isolation window was 1.8 Th and the maximum injection time for MS/MS was set at 120 ms. After MS analysis, the data was searched by pLink2 for the identification of cross-linked peptides (Chen, 2019). The mass accuracy was specified as 10 and 20 p.p.m. for MS and MS/MS, respectively. Other search parameters included cysteine carbamidomethylation as a fixed modification and methionine oxidation as a variable modification. A maximum of three trypsin missed-cleavage sites was allowed. The initial search results were obtained using the default 5% false discovery rate, estimated using a target-decoy search strategy. The crosslink spectra were then manually checked to remove false-positive identifications essentially as previously described (Shi, 2014; Kim, 2018; Shi, 2015).

Site-directed mutagenesis. Mammalian expression plasmid of HSA was obtained from Addgene. E400R point mutation was introduced to the HSA sequence by the Q5 site-directed mutagenesis kit (NEB) using the primer HSA-F (GGTGTTCGACCGGTTCAAGCCTCTGG, SEQ ID NO: 2652) and HSA-R (TTGGCGTAGCACTCGTGA, SEQ ID NO: 2653). After sequence verification by Sanger Sequencing, plasmids bearing wild type HSA and the mutant were transfected to HeLa cells using Lipofectamine 3000 transfection kit (Thermo) and Opti-MEM (Gibco) according to the manufacturer's protocol. The cells were cultured overnight before change of medium to DMEM without FBS supplements to remove BSA. After a 48 h culture at 37°C, 5% CO_2, the media expressing HSA were collected and stored at -20°C. The media were analyzed by SDS-PAGE and Western Blotting to confirm protein expression.

The PDZ domain (in the pGEX6p-1 vector) was obtained from the General Biosystems. A double point mutant of PDZ (i.e., R46E: K48D) was introduced by the Q5 Site-directed mutagenesis kit using specific primers of PDZ-F (TGATGAAAATGGCGCAGCCGCC, SEQ ID NO: 2654) and PDZ-R (ATTTCACTCACATAGATACCACTATCATTACTAACATAC, SEQ ID NO: 2655). After verification by Sanger Sequencing, the mutant vector was transformed into BL21(DE3) cells for expression. The GST fusion PDZ mutant protein was purified by GSH resin as previously described.

Fluorescence Microscopy. COS-7 cells were plated onto the glass bottom dish at an initial confluence of 60-70% and cultured overnight to let the cells attach to the dish. Cells were with MitoTracker Orange CMTMRos (1:4000) at 37 °C for 30 minutes, washed once with PBS and fixed with pre-cold methanol/ethanol (1:1) for 10 minutes. After being washed with PBS, the cells were blocked with 5% BSA for 1 hour. Alexa Fluor™ 647-conjugated Nb (1:100) was then added to the cells, incubated for 15 minutes at room temperature. Two-color wide-field fluorescence images were acquired using our custom-built system on an Olympus IX71 inverted microscope frame with 561 nm and 642 nm excitation lasers (MPB Communications, Pointe-Claire, Quebec, Canada) and a 100X oil immersion objective (NA=1.4, UPLSAPO 100XO; Olympus).

Text-based CDR (complementarity-determining region) Annotation. The CDR annotation method was modified from (Fridy, 2014). [*] denotes any residue.

CDR1 annotation: The short sequence motif "SC" was first searched, which is localized between the residue 20- residue 26 of a Nb sequence. The start of a CDR1 sequence is defined as the 5th residue followed by the "SC" motif. Once the first residue is identified, we then look for another sequence motif "W[*]R" which is localized between Nb residue 32- residue 40, and define the end of the CDR1 sequence as the first residue preceding the "W[*]R" motif.

CDR2 annotation: The start of a CDR2 sequence is defined as the 14th residue followed by the "W[*]R" motif. Once the first residue is identified, motif "RF" which is localized between Nb residue 63- residue 72 was then identified, and the end of the CDR2 sequence as the 8th residue preceding the "RF" motif was defined.

CDR3 annotation: The motif of "Y[*]C" or "YY[*]" was first searched, which is localized between Nb residue 90- residue 105. The start of a CDR3 sequence is defined as the 3rd residue followed by the "Y[*]C" or "YY[*]" motif. Once the first residue of a CDR3 was identified, either one of the following sequence motifs ("WG[*]G", "WGQ[*]","W[*]Q[*]", "[*]GQG","[*][*]GQ" and "WG[*][*]") was then used to locate the end of the CDR3. These motifs are located within the last 14 residues of the C terminal Nb sequence. CDR3 ends at 1 residue ahead of the sequence motif. More information can be found in the Augur Llama scripts.

The cleavage rules for in-silico digestion of Nbs by different proteases:

Trypsin:	C-terminal to K/R, not followed by P
Chymotrypsin:	C-terminal to W/F/L/Y, not followed by P
GluC:	C-terminal to D/E, not followed by P
AspN:	N-terminal to D
LysC:	C-terminal to K

Sequence alignment of Nb database: Nb sequences were aligned using the software ANARCI (Dunbar, J. & Deane, C.M, 2016). Three CDRs (CDR1-CDR3) and four Framework sequences (FR1-FR4) were annotated according to IMGT numbering scheme (Lefranc, 2003). Alignments below the threshold e-value of 100 were removed and the remaining sequences were plotted by WebLogo (Crooks, 2004).

In-silico digestion of Nb database by different proteases and analysis of Nb CDR3 mapping. A high-quality database containing approximately 0.5 million unique Nb sequences was *in-silico* digested using different enzymes including trypsin,chymotrypsin, LysC, GluC, and AspN according to the above cleavage rules. CDR3 containing peptides were obtained to calculate the sequence coverages. The CDR3 coverages were then summed to generate FIG. 1D & 7B. The CDR3 peptide length distributions (by trypsin and chymotrypsin) were plotted to generate FIG. 1E.

Simulation of trypsin and chymotrypsin-aided MS mapping of Nbs. 10,000 Nb sequences with unique CDR3 fingerprint sequences were randomly selected from the database. The selected Nbs were then *in-silico* digested by either trypsin or chymotrypsin (with no-miscleavage sites allowed) to generate CDR3 peptides. The following criteria were applied to these peptides to better simulate Nb identifications by MS: 1) peptides of favorable sizes for bottom-up proteomics (between 850- 3,000 Da) were first selected. 2) Peptides containing the highly conserved C-terminal FR4 motif of WGQGQVTS were further discarded. Based on our observations, such peptides are often dominated by C terminal y ion fragmentations, while having poorly fragmented ions on the CDR3 sequence which are essential for unambiguous CDR3 peptide identifications. 3) CDR3 peptides with limited Nb fingerprint information (containing less than 30% CDR3 sequence coverage) were removed. As a result, 2,111 unique tryptic peptides and 5,154 unique chymotryptic peptides were obtained. These peptides were then used to map Nb proteins. After protein assembly, only Nb identifications with sufficiently high CDR3 fingerprint sequence coverages (≥ 60%) were used to generate the venn diagram in FIG. 1F.

Phylogenetic analysis of Nb CDR3 sequences. Phylogenetic trees were generated by Clustal Omega (Sievers, 2014) with the input of unique Nb CDR3 sequences and the additional flanking sequences (i.e., YYCAA to the N-term and WGQG to the C-term of CDR3 sequences) to assist alignments. The data was plotted by ITol (Interactive Tree of Life) (Letunic, I. & Bork, P, 2007). Isoelectric points and hydrophobicities of Nb CDR3s were calculated using the BioPython library. Sequence alignments were visualized by Jalview (Waterhouse, 2009).

Evaluation of the reproducibility of Nb peptide quantification. Shared peptide identifications among different LC runs were used to evaluate the reproducibility of the label-free quantification method. For a typical 90 min LC gradient, the peptide peak width or full width at half maximum (FWHM) in general was less than 5s. The differences of peptide retention time among different LC runs were calculated to generate the kernel density estimation plots in FIG. 3B. Peptide retention times from different LC runs were used to calculate pearson correlation and were plotted in FIG. 9B.

Sequence alignment and analysis of HSA and Llama serum albumin. Llama (Camelus Ferus) serum albumin sequence was fetched and aligned with HSA by tblastn (NCBI). The isoelectric point (pI) and hydropathy values for individual amino acids were obtained online from (www.peptide2.com/N_peptide_hydrophobicity_hydrophilicity.php). These values were normalized between 0 to 1.0 and the sequence variations between the two albumins were calculated for each aligned position (the pairwise differences of pI and hydropathy). For a specific aligned residue position, a value of 0 indicates identical residues were found between the two sequences, while 1.0 indicates the largest sequence variation, such as a charge reversion from the negatively charged residue glutamic acid 400 for HSA to the positively charged residue arginine at the corresponding aligned position for camelid albumin. A value of 0.5 was assigned at the position where an insertion or deletion of amino acid was identified. Sequence variations of both pI and hydropathy between HSA and Llama serum albumin were thus plotted. The plots were further smoothed by a gaussian function to generate FIG. 4A.

Analysis of relative abundance of amino acids on Nb CDRs. The amino acid frequencies at each CDR (including CDR1, CDR2 and CDR3 head) were calculated and normalized to generate the bar plots and the pie plots in FIG. 6, 7, 12 and 13. CDR3 head sequences were obtained by removing the semi-conserved C terminal four residues of CDR3s. The CDR residue frequencies of both high-affinity and low-affinity Nbs were normalized based on the sum of the CDR residues of each affinity group.

Analysis of amino acid positions on CDR3 heads. The relative position of a residue on a CDR3 head was calculated where a value of 0 indicates the very N terminus of a CDR3 head while 1.0 indicates the last residue. The CDR3 head sequences were then sliced into 20 bins with a bin width of 0.05. Within each bin, the occurrence of a specific type of amino acid (such as tyrosine, glycine, or serine) was counted and normalized to the sum of residues on CDR3 heads. The distributions of different amino acids including their relative positions and abundances were plotted in FIG. 5H and 12G.

Proteomics database search of Nb peptide candidates. Raw MS data was searched by Sequest HT embedded in the Proteome Discoverer 2.1 (Thermo Fisher) against an in-house generated Nb sequence database using the standard target-decoy strategy for FDR estimation. The mass accuracy was specified as 10 ppm and 0.02 Da for MS1 and MS2, respectively. Other search parameters included cysteine carbamidomethylation as a fixed modification and methionine oxidation as a variable modification. A maximum of one or two missed-cleavage sites was allowed for trypsin and chymotrypsin-processed samples respectively. The initial search results were filtered by percolator with the FDR of 0.01 (strict) based on the q-value (Kall, 2007). After database search, the peptide-spectrum-matches (PSMs) were exported, processed and analyzed by Augur Llama with following steps:

a. Nanobody Identification

i) Quality assessment of CDR3 fingerprints

Peptide candidates were first annotated as either CDR or FR peptides. To confidently identify CDR3 fingerprint peptides, we implemented a filter/algorithm requiring sufficient coverage of high-resolution CDR3 fragment ions in the PSMs (See illustration in FIG. 8B). The filter was evaluated using a target sequence database containing approximately 0.5 million unique Nb sequences and a non-overlapping decoy database of similar size. Target and decoy Nb sequence databases herein used were obtained from different llamas. Any peptide identification from the decoy database was considered as a false positive. The FDR was defined based on the % of peptide identifications from the decoy database compared with those from the target database. CDR3 length was also considered to enable development of a sensitive CDR3 peptide filter. The CDR3 fragmentation coverage was defined as the percentage of the CDR3 residues that were matched by fragment ions (either b ions or y ions) within the mass accuracy window. Spectra of the same peptide were combined for assessment. Only CDR3 peptides that passed this filter (5% FDR) were selected for the downstream Nb assembly.

ii) Nanobody sequence assembly

CDR peptides including the confident CDR3 peptides were used for Nb protein assemblies. Two additional criteria must be matched before a Nb can be identified. These include: 1) both CDR1 and CDR2 peptides must be available for a Nb assembly. 2) for any Nb identification, a minimum of 50% combined CDR coverage was mandated.

b. Quantification and classification of antigen-specific Nb repertoires

MS raw data was accessed by MSFileReader 3.1 SP4(ThermoFisher), and a python library of pymsfilereader (github.com/frallain/pymsfilereader). Reliable CDR3 peptides that passed the quality filter were quantified by label-free LC/MS.

i) CDR3 peptide quantification

To enable accurate label-free quantification of CDR3 peptide identification across different LC runs, different retention time windows for peptide peak extraction were specified. For peptides that can be directly identified by the search engine based on the MS/MS spectra, a small quantification window of +/- 0.5 minutes retention time (RT) shift was used for peak extractions. For peptides that were not directly identified from a particular LC run (due to the complexity of peptides and stochastic ion sampling), their RTs were predicted based on the RT of the adjacent LC and were adjusted using the median RT difference of the commonly identified peptides between the two LC runs. In this case, a relaxed RT window of +/- 2.0 minutes (for a typical 90 min LC gradient), in which approximately 95% of all the identified peptides can be matched between the two LC runs, was applied to facilitate extraction of the peptide peaks. Both m/z and z of a peptide were used for peak extractions with a mass accuracy window of +/- 10 ppm. The peptide peaks were extracted and smoothed using a Gaussian function. Their AUCs (area under the curve) were calculated and AUCs from the replicated LC runs were averaged to infer the CDR3 peptide intensities.

ii) Classifications of Nbs

To enable accurate classifications e.g., based on Nb affinities, relative ion intensities (AUCs) of the CDR3 fingerprint peptides among three different biochemically fractionated Nb samples (*F1*, *F2* and *F3*) were quantified as *I1*, *I2* and *I3*. Based on the quantification results, CDR3 peptides were arbitrarily classified into three clusters (*C1*, *C2*, and *C3*) using the following criteria:

1) For *C3* (high-affinity) cluster: $I3 > I1+I2$ (indicating Nbs were more specific to *F3*)

2) For C2 (mediocre-affinity) cluster: $I2 > I1+I3$ (indicating Nbs were more specific to *F2*)

3) For *C1* (low-affinity) cluster:

$I1 > I2+I3$ (indicating Nbs were either more specific to *F1* or likely nonspecific binders), alternatively, if $I1 < I2+I3$ and $I2 < I1+I3$ and $I3 < I1+I2$, these Nb identifications were likely nonspecifically identified and were grouped into *C1* as well. See illustration in FIG. 8C.

The above method was used to classify HSA and GST Nbs. Some modifications were made for quantification and characterization of high-affinity PDZ Nbs. Specifically, an additional control of MBP interacting Nbs "*F_control*" (ion intensity of *I_control*) was included for quantification. High-affinity cluster Nbs (represented by their unique CDR3 peptides) were defined when the sum intensities of *I2* and *I3* for a Nb CDR3 peptide were 20 fold higher than *I_control*(i.e. *20*I_control* < *I2+I3*). For Nbs where more than one unique CDR3 peptide was used for quantification, classification results among different CDR3 peptides from the same Nb must be consistent; otherwise, they were removed before the final results were reported.

Heatmap analysis of the relative intensities of CDR3 peptides. The identified CDR3 peptides were quantified based on their relative MS1 ion intensities and were subsequently clustered using scripts in Augur Llama. Z-scores were calculated based on the relative ion intensities and were used to generate a heatmap in FIG. 3A for visualization.

Structural modeling of antigen-Nb complexes. Structural models for Nbs were obtained using a multi-template comparative modeling protocol of MODELLER (Webb, B. & Sali, A, 2014). Next, we refine the CDR3 loop and select the top 5 scoring loop conformations for the downstream docking. Each Nb model is then docked to the respective antigen by an antibody-antigen docking protocol of PatchDock software that focuses the search to the CDRs (Schneidman-Duhovny, 2005). The models are then re-scored by a statistical potential SOAP (Dong, 2013). The antigen interface residues (distance <XÅ from Nb atoms) among the 10 best scoring models according to the SOAP score were used to determine the epitopes. Once the epitopes were defined, we clustered Nbs based on the epitope similarity using k-means clustering. The clusters reveal the most immunogenic surface patches on the antigens. Antigen-Nb complexes with CXMS data were modeled by distance-restrained based PatchDock protocol that optimizes restraints satisfaction (Schneidman-Duhovny, 2020; Russel, 2012). A restraint was considered satisfied if the Ca-Ca distance between the cross-linked residues was within 25Å and 20Å for DSS and EDC cross-linkers, respectively (Shi, 2014; Fernandez-Martinez, 2016). In the case of ambiguous restraints, such as the GST dimer, it is required that one of the cross-links is satisfied.

Machine learning analysis of Nb repertoires. A deep neural network was trained to distinguish between low- and high- affinity Nbs that were characterized by the accurate high-pH fractionation method and quantitative proteomics. This model consists of one convolutional layer with batch normalization and ReLU activation function, followed by a max pooling layer ending with a fully connected layer to integrate the features extracted into the logits layer that leads to the classifier prediction. The convolutional layer consists of 20 1D filters, representing local receptive fields with window size of 7 amino acids, long enough to capture the relevant CDRs and short enough to avoid data overfitting. During the forward pass, each filter slides along the protein sequence with a fixed stride performing an elementwise multiplication with the current sequence window, followed by summing it up to generate a filter response. The classification accuracy of the model was 92%.

To understand the physicochemical features learned by the network for distinguishing low- and high- affinity binders, the activation path was calculated through the network back from the prediction to the activated filter. Similar to the backpropagation algorithm, backward was iterated from the last two layers of fully connected network, extracting for each sequence the output signal and looking for the highest peaks which contribute the most weight to the classification. In the same way, upstream the contribution of each filter to those peaks was calculated. In addition, filter activity in CDRs was analyzed to extract region-specific dominant filters. This process of network interpretation results in a unique contribution per filter per sequence. Each filter is activated along the sequence downsampled in the max pooling layer. For each filter, its highest peak was then picked leading to classification. Finally, the most contributing filters per sequence was determined and there also we got an interesting filter out with more than 30% contribution in those regions of interest.

Computer Implemented Methods

It should be appreciated that the logical operations described herein with respect to the various figures may be implemented (1) as a sequence of computer implemented acts or program modules (i.e., software) running on a computing device (e.g., the computing device described in FIG. 14), (2) as interconnected machine logic circuits or circuit modules (i.e., hardware) within the computing device and/or (3) a combination of software and hardware of the computing device. Thus, the logical operations discussed herein are not limited to any specific combination of hardware and software. The implementation is a matter of choice dependent on the performance and other requirements of the computing device. Accordingly, the logical operations described herein are referred to variously as operations, structural devices, acts, or modules. These operations, structural devices, acts and modules may be implemented in software, in firmware, in special purpose digital logic, and any combination thereof. It should also be appreciated that more or fewer operations may be performed than shown in the figures and described herein. These operations may also be performed in a different order than those described herein.

Referring to FIG. 14, an example computing device 500 upon which the methods described herein may be implemented is illustrated. It should be understood that the example computing device 500 is only one example of a suitable computing environment upon which the methods described herein may be implemented. Optionally, the computing device 500 can be a well-known computing system including, but not limited to, personal computers, servers, handheld or laptop devices, multiprocessor systems, microprocessor-based systems, network personal computers (PCs), minicomputers, mainframe computers, embedded systems, and/or distributed computing environments including a plurality of any of the above systems or devices. Distributed computing environments enable remote computing devices, which are connected to a communication network or other data transmission medium, to perform various tasks. In the distributed computing environment, the program modules, applications, and other data may be stored on local and/or remote computer storage media.

In its most basic configuration, computing device 500 typically includes at least one processing unit 506 and system memory 504. Depending on the exact configuration and type of computing device, system memory 504 may be volatile (such as random access memory (RAM)), non-volatile (such as read-only memory (ROM), flash memory, etc.), or some combination of the two. This most basic configuration is illustrated in FIG. 14 by dashed line 502. The processing unit 506 may be a standard programmable processor that performs arithmetic and logic operations necessary for operation of the computing device 500. The computing device 500 may also include a bus or other communication mechanism for communicating information among various components of the computing device 500.

Computing device 500 may have additional features/functionality. For example, computing device 500 may include additional storage such as removable storage 508 and non-removable storage 510 including, but not limited to, magnetic or optical disks or tapes. Computing device 500 may also contain network connection(s) 516 that allow the device to communicate with other devices. Computing device 500 may also have input device(s) 514 such as a keyboard, mouse, touch screen, etc. Output device(s) 512 such as a display, speakers, printer, etc. may also be included. The additional devices may be connected to the bus in order to facilitate communication of data among the components of the computing device 500. All these devices are well known in the art and need not be discussed at length here.

The processing unit 506 may be configured to execute program code encoded in tangible, computer-readable media. Tangible, computer-readable media refers to any media that is capable of providing data that causes the computing device 500 (i.e., a machine) to operate in a particular fashion. Various computer-readable media may be utilized to provide instructions to the processing unit 506 for execution. Example tangible, computer-readable media may include, but is not limited to, volatile media, non-volatile media, removable media and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. System memory 504, removable storage 508, and non-removable storage 510 are all examples of tangible, computer storage media. Example tangible, computer-readable recording media include, but are not limited to, an integrated circuit (e.g., field-programmable gate array or application-specific IC), a hard disk, an optical disk, a magneto-optical disk, a floppy disk, a magnetic tape, a holographic storage medium, a solid-state device, RAM, ROM, electrically erasable program read-only memory (EEPROM), flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices.

In an example implementation, the processing unit 506 may execute program code stored in the system memory 504. For example, the bus may carry data to the system memory 504, from which the processing unit 506 receives and executes instructions. The data received by the system memory 504 may optionally be stored on the removable storage 508 or the non-removable storage 510 before or after execution by the processing unit 506.

It should be understood that the various techniques described herein may be implemented in connection with hardware or software or, where appropriate, with a combination thereof. Thus, the methods and apparatuses of the presently disclosed subject matter, or certain aspects or portions thereof, may take the form of program code (i.e., instructions) embodied in tangible media, such as floppy diskettes, CD-ROMs, hard drives, or any other machine-readable storage medium wherein, when the program code is loaded into and executed by a machine, such as a computing device, the machine becomes an apparatus for practicing the presently disclosed subject matter. In the case of program code execution on programmable computers, the computing device generally includes a processor, a storage medium readable by the processor (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. One or more programs may implement or utilize the processes described in connection with the presently disclosed subject matter, e.g., through the use of an application programming interface (API), reusable controls, or the like. Such programs may be implemented in a high level procedural or object-oriented programming language to communicate with a computer system. However, the program(s) can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language and it may be combined with hardware implementations.

As noted above, logical operations described herein, for example logical operations as described in Example 8, can be implemented with hardware, software or, where appropriate, with a combination thereof. For example, the logical operations can be implemented using one or more computing devices such as computing device 500 of FIG. 14. Logical operations described in Example 8 include, but are not limited to, methods for determining antigen affinity of nanobody peptide sequences, methods for training deep learning models, and deep learning-based methods for inferring antigen affinity of nanobody peptide sequences. These operations are described in detail above.

In some embodiments, a computer-implemented method includes:

receiving a nanobody peptide sequence;

identifying a plurality of CDR regions of the nanobody peptide sequence, the CDR regions including CDR3 regions;

applying a fragmentation filter to discard one or more false positive CDR3 regions of the nanobody peptide sequence;

quantifying an abundance of one or more non-discarded CDR3 regions of the nanobody peptide sequence; and

inferring an antigen affinity based on the quantified abundance of the one or more non-discarded CDR3 regions of the nanobody peptide sequence.

In some embodiments, a method for training a deep learning model includes:

creating a dataset that comprises a plurality of nanobody peptide sequences and corresponding antigen-affinity labels; and

training, using the dataset, a deep learning model to classify nanobody peptide sequences having low antigen affinity and nanobody peptide sequences having high antigen affinity.

In some embodiments, a method for determining antigen affinity of nanobody peptide sequences includes:

receiving a nanobody peptide sequence;

inputting the nanobody peptide sequence into a trained deep learning model; and

classifying, using the trained deep learning model, the nanobody peptide sequence as having low antigen affinity or high antigen affinity.

Table 1. Summary of GST Nbs and their biophysical and physiochemical properties

ID	Enzyme	Protein Sequence	SEQ ID NO	Salt Trend	LowpH Trend	HighpH Trend	Soluble	Binder by Beads-binding Assay (Fig S3C)	ELISA affinity (LogIC50 (oD450nm))	SPR ka (1/Ms)	SPR kd (1/s)	SPR KD (M)	Cross-linker	Cross-linked Peptides	CX residue on GST	CX residue on Nbs	CX Model Folder	CX Model Epitope
G1	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAAPGSTFSTNIIAWYRQPPGKQRELVAAIGGPGSTNYADSVKGRFTISRDNAKNTGYLQMKSLKPDDTAVYYCNMVTQRGNEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 1	0	0	2	Yes	/	2.93	/			/	/	/	/	/	/
G2	Trypsin/Chymo	MASMTGGQQMGRNSAEVQLVESGGDLVQAGGSLRLSCSASGNIFKINDMGWYRQAPGKQRELVARISSSGNTNYADSVKGRFTISRDNGKNTVYLQMNRVKPEDTAVYYCNADVQVSRNYEYEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 2	/	1	0	Yes	/	2.667	1.02E+03	2.04E-03	2.00E-06	/	/	/	/	/	/
G3	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASAGTFSTYAISWFRQAPGKERDFVAAINRISRSAYSPYYADSVKGRFTISEDNAKNTVNLQMNSLKPEDTAVYYCAAGSIFHTDVRYYAYWARGPRSPSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 3	/	/	0	Yes	Yes	/	/			/	/	/	/	/	/
G4	Trypsin/Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQPGGSLRLSCSASGRTLDSYGIGWFRQAPGKEREEVSCISSSGGNADYADSVMGRFTISRDNAKNTVYLHMNNLRPEDTAVYYCAAIAGLCALHYTDYKVVVIPGSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 4	/	0	/	Yes	Yes	/	/			/	/	/	/	/	/

G5	Trypsin	MASMTGGQQMGRNSAEVQLVESGGGVVQPGGSLTLSCAASGFAFRNYAMSWVRQAPGKGPEWVSQINGRGGYTSYADSVKGRFTISRDNTKNTLYLQMNNLKPDDTAVYYCAKDPTQLRWIPVPNYILGSTKGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 5	/	0	0	Yes	/	0, No binding	/			/	/	/	/	/	/
G6	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGRTISSYAMGWFRQAPGKERELVARITSSAGSTYYADSVKGRFTISRDNAKNTMYLQMNSLKPEDTAVYYCAVEIVRAQYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 6	/	/	0	Yes	/	0, No binding	/			/	/	/	/	/	/
G7	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQPGDSLRLSCAVSGQYVNMAAMGWFRQAPGKEREFVAGISWSDDTDIADSVKGRFTISRDHGKNTVDLQMNSLKPEDTGVYLCAGRFRRLAKDFGEYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 7	/	/	0	Yes	Yes	1.885	3.02E+03	8.09E-04	2.68E-07	/	/	/	/	/	/
G8	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGITSSIASMGWFRQAPGKEEEFVARIRWNTDNTYYADSVKGRFTISRDNAQNTVYLQMNRLKPEDTAVYYCVARRGWSDLLYDYRGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 8	1	0	1	Yes	/	4.658	/			/	/	/	/	/	/
G9	Trypsin	MASMTGGQQMGRNSADVQLVESGGGLVQPGGSLRLSCAASGLTLDNYDMAWFRQAPGKEREFVTAINYVGGRTYADSVRGRFTISRDDTKNTVYLQMNSLKPEDTAVYYCAAGLQYGITSLRTRNYNYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 9	1	/	2	Yes	Yes	8.05	2.00E+06	3.53E-04	1.77E-10	DSS	RIEAIPQIDKYLK (SEQ ID NO: 2537)(10)-NTVYLQMNSLKPEDTAVYYCAAGLQYGITSLR (SEQ ID NO: 2538)(11)	GST(191)	G9 (101)	Seq_17023	E1
G10	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGSIFSINSMGWYRQAPGIERELVAHMPTGGNTNYLDSVKGRFVISRDDDKKTVYLQMNSLTPEDTAVYYCHAVITTVGRTGVRTYSYWARGPRSPSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 10	/	/	2	No	/	/	/			/	/	/	/	/	/

G11	Trypsin/Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGRTFNSGILGWFRQAPGKDREFVAAIGWSAGSTYYSDSVKGRFTISRDITKNTVFLQMNSLKPEDTAVYYCADKKYYYGREASSNVYEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 11	/	0	/	No	/	/	/			/	/	/	/	/	/
G12	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASRSTFRINAAGWYRQAPGKERELVARISSGGSTNYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCNVPYYREDGYEYDAWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 12	1	0	2	Yes	/	4.207	/			DSS	YEEHLYERDEGDKWR (SEQ ID NO: 2539)(13)-DNAKNTVYLQMNSLKPEDTAVYYCNVPYYR (SEQ ID NO: 2540) (4)	GST(40)	G12 (90)	Seq_20204	E2
													EDC	VDFLSKLPEMLK (SEQ ID NO: 2541) (6)-EDGYEYDAWGQGTQVTVSSEPK (SEQ ID NO: 2542) (7)	GST(125)	G12 (123)		
														VDFLSKLPEMLK (SEQ ID NO: 2541)(6)-NTVYLQMNSLKPEDTAVYYCNVPYYREDGYEYDAWGQGTQVTVSSEPK (SEQ ID NO: 2543) (31)	GST(125)	G12 (121)		
														SDLEVLFQGPLGSPEFPGR (SEQ ID NO: 2544) (15)-ISSGGSTNYADSVKGR (SEQ ID NO: 2545) (14)	GST(233)	G12 (79)		
G13	Chymo	MASMTGGQQMGRNSADVQLVESGGGVVQAGGSLRLSCAASGRTFSDYAMGWFRQAPGKEREFVAGVSWSGVDTYYADSVKGRFTISRDNAKNTLYVQMNSLKPEDTAVYYCAAQRYYHGHAKNMRYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 13	/	/	2.00E+00	Yes	Yes	6.735	4.74E+05	1.70E-04	3.60E-10	DSS	RIEAIPQIDKYLK (SEQ ID NO: 2537) (10)-ASMTGGQQMGR (SEQ ID NO: 2546) (1)	GST(191)	G13 (2)	Seq_73309	E3
														KRIEAIPQIDK (SEQ ID NO: 2547) (1)-ASMTGGQQMGR (SEQ ID NO: 2546) (1)	GST(181)	G13 (2)		
														LLLEYLEEKYEEHLYER (SEQ ID NO: 2548) (9)-ASMTGGQQMGR (SEQ ID NO: 2546) (1)	GST(27)	G13 (2)		

														SSKYIAWPLQGWQATFGGGDHPPK (SEQ ID NO: 2549) (3)-ASMTGGQQMGR (SEQ ID NO: 2546) (1)	GST(197)	G13 (2)		
G14	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAASGSTFDTNPIGWYRQAPGKQRDLVAMITSGGHTNYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCTVPHYREDGYEYHFWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 14	2	0	2	Yes	/	5.274	/			DSS	DFETLKVDFLSK (SEQ ID NO: 2550) (6)-QAPGKQR (SEQ ID NO: 2551) (5)	GST(119)	G14 (58)	Seq_47378	E2
														IAYSKDFETLK (SEQ ID NO: 2552) (5)-DNAKNTVYLQMNSLKPEDTAVYYCTVPHYR (SEQ ID NO: 2553) (4)	GST(113)	G14 (90)		
													EDC	LSCAASGSTFDTNPIGWYR (SEQ ID NO: 2554) (11)-IAYSKDFETLK (SEQ ID NO: 2552) (5)	GST(113)	G14 (45)		
G15	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGSTFDTNPIGWYRQAPGKQRDLVAMITSGGHTNYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCTVPHYREDGYEYHCWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 15	2	0	2	Yes	Yes	4.606	/			/	/	/	/	/	/
G16	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQPGGSLRLSCAASGSTFSINAIGWYRQAPGKEREFVAALRWPGNIWYYADFVEGRITISRDNAKNTVYLQMNSLKPEDTAVYYCAARPENRGSYRDAATYDFWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 16	2	1	/	Yes	/	5.684	/			/	/	/	/	/	/
G17	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAASGVTISYWVMGWFRQAPGKEREFVARISWGGERTYYADSVKGRFAISRDNAKNTVYLQMNSLNAEDTAVYYCAADRTGWGHSNSRSEYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 17	2	2	2	Yes	Yes	9.81	1.34E+06	2.92E-05	2.17E-11	/	/	/	/	/	/

G18	Trypsin/Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGASLRLTCGPSGRSVGLYTMGWFRQAPGKEREFVAGVTYLGDTTTYSDAVKGRFTISRENNKNTVYLRMNSLKPEDTAVYYCTATATGWGSPIPSAPGRWDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 18	/	0	/	Yes	Yes	/	/	/	/	/	/	/	/
G19	Trypsin/Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGSTFSTNAVDWYRQAPGNQRDLVATITSGGHTNYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAVPHYREDGYEYRFWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 19	1	0	1	Yes	/	9	/	/	/	/	/	/	/
G20	Trypsin	MASMTGGQQMGRNSADVQLVESGGGLVQAGGSLRLSCAASERTFSRYMLGWFRQAPGKEREFVGVMGWSDSDTYYGDAVKGRFTISRDNVKNTIYLQMKSLKPEDTAVYYCAASAYGSTRNHKLYEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 20	1	/	/	Yes	/	2.434	/	/	/	/	/	/	/
G21	Trypsin/Chymo	MASMTGGQQMGRNSADVQLVESGGGSVQAGGSLRLSCAASGRTFSNYAMAWFRQAPGKEREFVAAVSRSGTNLYYADSVKGRFTISRDTAENTMYLQMNSLKPEDTAVYYCAAGLAERWGIGVQPRSEFLTTGARGPRSPSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 21	/	0	2	Yes	Yes	/	/	/	/	/	/	/	/
G22	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGSVQAGGSLRLSCAASGRTFSSYSMAWFRQAPGKEREFVAVMNCRYGDTDYPDSVKGRFTMSRDNAKNTLYLQMNSLKPEDTAVYYCAAKLIAYCGSGYYYRRNDYGYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 22	1	/	1	Yes	/	2.071	3.09E+04 1.25E-03 4.05E-08	/	/	/	/	/	/
G23	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGLVQPGGSLRLSCAASGFTFSVNTMSWVRQAPGKGREWVSGIESHGNTYYSDSVKGRFTISRDNAKNTLYLQMNSLKPEDTAVYYCATGIYGTTRNWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 23	1	/	/	Yes	/	2.543	/	/	/	/	/	/	/

G24	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGTFGSYVMGWFRQPPGKEREFVSGIMWNGTSTSTNYADSVKGRFTISRDNAKNTVFLQMNSLQPEDTAVYYCAASRSSALRTPVPLVEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 24	1	2	2	Yes	Yes	/	/	/	/	/	/	/	/
G25	Trypsin/Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAASGRTFSGRTFSDYPMAWFRQAPGKEREFLATISTSGSRTYYADSVKGRFTISRDNAKDTVYLQMNSLKPEDAAIYYCAARQGSYYSDYNRALPGEYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 25	0	0	1	Yes	/	1.575	/	/	/	/	/	/	/
G26	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLAQPGGSLRLSCAASGFTLDAYAIAWYRQAPGKDREEVACISSSGDSTNYAESVKGRFTISRDNAKKMGYLQMNSLKAEDTAIYYCAIDSRGCAWGGFAYYTFSHWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 26	/	/	0	Yes	Yes	/	/	/	/	/	/	/	/
G27	Trypsin	MASMTGGQQMGRNSADVQLVESGGDLVQAGGSLRLSCSASGNIFKINDMDWYRQAPGKQRELVARISSSGSTNYADSVKGRFTISRDNGKNTVYLQMNRVKPEDTAVYYCNADVQVSRNYEYEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 27	/	1	0	Yes	/	1.365	/	/	/	/	/	/	/
G28	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQPRGSLRLSCAASGFTWGDYAIGWFRQAPGKEREGVSCLSSSDGSTYYPDSVKDRFTISTDNAKNTVYLQMTNLKPDDTAIYYCAAREGPGASWYCSVNGYLTQPDSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 28	/	0	0	Yes	Yes	/	/	/	/	/	/	/	/
G29	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGLVQAGDSLLLSCGTSGRTFSSNTMGWFRQAPGKGREFVATITASGRGTNYGDSVRGRFTISRDNDKNTVYLQMNNLKPDDTGVYTCAASDSPYGSRWIEAYGYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 29	1	0	2	Yes	/	0, No binding	/	/	/	/	/	/	/

G30	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGRTINNYDMGWFRQAPGKEREFVAAITWSGRDTNYADSVKGRFTVSRDDAKNTVYLQMNTLSPEDTAVYYCASARIQFYRLVAATRTDYSYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 30	/	/	1	Yes	/	4.86	/			/	/	/	/	/	/
G31	Trypsin/Chymo	MASMTGGQQMGRNSAHVQLVESGGGLVQAGGSLRLSCKASESIFKFDAMAWFRQAPGKERELVACIDNKQRTTYGDSVKGRFTISGLDVKNTAYLEMNSLKPEDTAVYYCTADRSTCFSNYRLYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 31	0	0	0	Yes	Yes	/	/			/	/	/	/	/	/
G32	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVRAGDSLRLSCVVSGRPISSYAMAWFRQAPGKDREVVAGISANGDRTHYADSIKGRFTVSRDNAKNSMTLQMNKLKPEDTAVYYCAADSLTEGGYGLTGDFDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 32	/	/	0	Yes	/	0, No binding	/			/	/	/	/	/	/
G33	Chymo	MASMTGGQQMGRNSAEVQLVESGGSLRLSCSVSGGPFTSNGMGWYRQAPGKEREWVAAITNSGSANYADSVKGRFTVSMVNANNTMYLQMNNLKPDDTAVYYCNVAGWPHGYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 33	0	1	0	Yes	/	0, No binding	/			/	/	/	/	/	/
G34	Trypsin/Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGDSLRLSCAASGRTFSRYAMAWFRQAPGKEREFVAGISWTGRFTYYADSVKGRFTISRDDAKNTVYLQMNNLKPEDTGLYFCKVGDPYGVGLREYEWWGPGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 34	/	/	1	Yes	Yes	5.316	2.63E+04	4.62E-04	1.76E-08	/	/	/	/	/	/
G35	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGRTSRSFAMGWFRQAPGKGRDFVAAMTEFGTTYYADSVKGRFTISRDNAKNTVYLQMNVLQSEDTAVYYCAAHWDNTQWYVYEVGGYEHWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 35	/	1	/	Yes	Yes	/	/			/	/	/	/	/	/

G36	Trypsin/Chymo	MASMTGGQQMGRNSAQVQLVESGGSLRLSCAASGFALSNSYMKWVRQAPGKGPEWVSTIYADGSTYYTDSVKGRFITSRDNSKNTMYLQMSDLKPEDTAVYYCANPSAKGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 36	2	0	0	Yes	/	0, No binding	/	/	/	/	/	/	/
G37	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGDSLRLSCVASGDTFTSYTVGWFRQAPGKEQEFVAGISWSGRSTDYADFVKGRATISKDIAKVSLQMNALKPEDTAVYSCAAKKVDWSSDYVTNYDYDYRGRGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 37	/	0	/	Yes	/	1.785	/	/	/	/	/	/	/
G38	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCVASGHTDCISGMGWYRQAPGKERELVAVLIGGGNTYYGDSVKGRFTISKDKAKNTLYLQMKTLKPEDMAVYYCTADDHGSECPNKEMSSTATYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 38	/	0	/	Yes	/	0, No binding	/	/	/	/	/	/	/
G39	Trypsin/Chymo	MASMTGGQQMGRNSAHVQLVESGGELVQSGSSLRLSCAASGFDLDDYAIGWFRQAPGKEREGVSCTSTSDGPTSYLDSVKGRFTFSRDNAKNTLYLQMNSLKPEDTAVYYCAAISHIFAEDAPAMGLCWDQRSAFWYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 39	1	/	2	Yes	/	1.653	/	/	/	/	/	/	/
G40	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQPGGSLTLSCAASGFHLDNTAIAWFRQAPGKEREGVSCLSSRDGSTFYQYSLKDRFTISGDNAKNTVYLQMKGLKPEDTATYYCAAALGIDSQRTVIAGCPKRYFAAWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 40	1	0	0	Yes	Yes	/	/	/	/	/	/	/	/
G41	Chymo	MASMTGGQQMGRNSADVQLVESGGGLVQAGGSLRLSCVASGHTVSNYAMAWFRQAPGKEREFVAGISWRASITYYRDSVKGRFTISRDNAKNTVYLQMSSLKPEDTAVYYCASDKTHYVSRGTSLVEYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 41	1	0	1	Yes	/	1.02	/	/	/	/	/	/	/

G42	Trypsin/Chymo	MASMTGGQQMGRNSAHVQLVESGGGFVQAGGSLRLSCEASGRTFNVYTMGWFRQAPGKEREFVGSISWNGGSTYYADSVKGRFTISRDNAKNTVYLQMNSLEPEDTAVYYCAARRQSHLRLDLSVIDAWGKGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 42	1	0	2	Yes	/	2.63	/	/	/	/	/	/	/
G43	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCATSGRTSSTYAMGWFRQRPGKEREFVATIHWGVGSTIYADSVKGRFTLSRDNAQNTVYLQMNSLKPEDTAVYYCAASTYRIGSYDVSTSQGYNYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 43	/	2	2	Yes	/	4.32	/	/	/	/	/	/	/
G44	Chymo	MASMTGGQQMGRNSADVQLVESGGGLVQAGGSLRLSCVASGPIFSFSTGGWYRQAPGKQRELVAALTGGGNTNYADSVKGRFTISRDNAKNTVYLQMNLLKPEDTAVYYCQVMYYSGYDGYESTSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 44	/	2	/	Yes	/	3.058	/	DSS	DFETLKVDFLSK (SEQ ID NO: 2550) (6)-QAPGKQR (SEQ ID NO: 2551) (5)	GST(119)	G44 (58)	Seq_41521	E3
												IEAIPQIDKYLK (SEQ ID NO: 2555) (9)-ELVAALTGGGNTNYADSVKGR (SEQ ID NO: 2556) (19)	GST(191)	G44 (79)		
												YLKSSK(SEQ ID NO: 2557) (3)-ELVAALTGGGNTNYADSVKGR (SEQ ID NO: 2556) (19)	GST(194)	G44 (79)		
												KRIEAIPQIDK(SEQ ID NO: 2547)(1)-ELVAALTGGGNTNYADSVKGR (SEQ ID NO: 2556)(19)	GST(181)	G44 (79)		
G45	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQPGGSLRLSCAASGSIFSINSMGWYRQAPGKQRELVAAITSGGSTNYANSVKGRFTISRNNARNTVWLQMNSLKPEDTAVYYCNADLNVVRGYSGDYHGSSDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 45	1	1	2	Yes	/	3.171	/	/	/	/	/	/	/

G46	Chymo	MASMTGGQQMGRNSADVQLVESGGGLVQAGGSLRLSCAASGRTFSRYHMGWFRQAPGKERDVVAAISWSGDSTYYADSVKGRFTISKDNAKNTVYLQMDNLKPEDTAVYYCNVRGGVLRPYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 46	/	2	/	Yes	/	0, No binding	/			/	/	/	/	/	/
G47	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASERIFSNYAMGWFRQAPGKEREFVASIRGSGSQTSYADSVKGRFTISRDGAKDTVDLQMNSLKPEDTAVYYCSAKKYCGSTYNRAEGYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 47	1	0	1	Yes	Yes	3.86	3.91E+05	1.27E-02	3.24E-08	DSS	IAYSKDFETLK(SEQ ID NO: 2552) (5)-DTVDLQMNSLKPEDTAVYYCSAK(SEQ ID NO: 2558)(11)	GST(113)	G47 (102)	Seq_54055	E2
														IAYSKDFETLK (SEQ ID NO: 2552)(5)-KYCGSTYNR (SEQ ID NO: 2559)(1)	GST(113)	G47 (115)		
													EDC	SDLEVLFQGPLGSPEFPGR (SEQ ID NO: 2545)(4)-DGAKDTVDLQMNSLK (SEQ ID NO: 2560)(4)	GST(222)	G47 (91)		
														IKGLVQPTR(SEQ ID NO: 2561)(2)-AEGYDYWGQGTQVTVSSEPK(SEQ ID NO: 2562)(5)	GST(11)	G47 (128)		
G48	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGRTFSTLSMGWFRQAPGQGREFVGGINYDGSSVEYADSVKGRFTISRDNAKNMMYLQMNSLKPEDTAAYYCASSRGYNTGTNPLGYDVWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 48	/	0	0	Yes	Yes	1.601	5.93E+03	7.53E-04	1.26E-07	/	/	/	/	/	/
G49	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAASRSTFSINAAGWYRQAPGKQRELVAAISSGGSANYADSVKGRFIISRDNAKNTVYLQMNSLKPEDTAVYYCRVPYYRDDGYEYYSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 49	2	0	/	Yes	/	5.545	/			DSS	IEAIPQIDKYLK (SEQ ID NO: 2555) (9)-ELVAAISSGGSANYADSVKGR(SEQ ID NO: 2563)(19)	GST(191)	G49 (79)	Seq_24699	E3
													EDC	LERPHRD (SEQ IDNO: 2564)(2)-DNAKNTVYLQMNSLKPEDTAVYYCR (SEQ ID NO: 2565) (4)	GST(239)	G49 (90)		

												LERPHRD (SEQ ID NO: 2564) (7)-DNAKNTVYLQMNSLKPEDTAVYYCR (SEQ ID NO: 2565) (4)	GST(244)	G49 (90)		
												LERPHRD(SEQ ID NO: 2564)(7)-NTVYLQMNSLKPEDTAVYYCR (SEQ ID NO: 2566)(11)	GST(244)	G49 (101)		
												SDLEVLFQGPLGSPEFPGR(SEQ ID NO: 2545)(15)-ELVAAISSGGSANYADSVKGR (SEQ ID NO: 2563)(19)	GST(233)	G49 (79)		
												LERPHRD (SEQ ID NO: 2564)(2)-NTVYLQMNSLKPEDTAVYYCR (SEQ ID NO: 2566)(11)	GST(239)	G49 (101)		
G50	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAASGRTFSRYHMGWFRQAPGKERDVVAAISWSGDSTYYADSVKGRFTISKDNAKNTVYLQMDSLKPEDTAVYYCATLSGWDGDTIFPAGSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 50	/	1	/	Yes	/	1.091	/	/	/	/	/	/	/
G51	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLKLSCAASGITFSINTIGWYRQAPGKQREFVAHITSDSTTYYADSVKARFTISRDSAKNTVHLQMNNLKPEDTAVYYCNVNPTWPYGGEVDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 51	/	0	2	Yes	/	4.449	/	/	/	/	/	/	/
G52	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAASGSTFSSKPIGWYRQAPGKGRDLVAAIGGGSSTFYVDSVKGRFTMSRDNAKNTVALQMNSLKPEDTAVYYCNEYLGPKVLPIGSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 52	/	0	/	Yes	/	1	/	/	/	/	/	/	/
G53	Chymo	MASMTGGQQMGRNSAHVQLVESGGGLVQAGGSLRLSCVASGFTYSTYTMGWFRQAPGKEREIVAAKNWSGARIYYTESVKGRFTI	SEQ ID NO:	2	1	/	Yes	/	5.365	/	DSS	LLLEYLEEKYEEHLYER(SEQ ID NO: 2548)(9)-IYYTESVKGR	GST(27)	G53 (80)	Seq_55403	E3

		SRDSGSNTMYLQMDSLKPEDTAVYYCAARLTWTDTTTPTTYPYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH										(SEQ ID NO: 2567)(8)				
												LLLEYLEEKYEEHLYER(SEQ ID NO: 2548)(9)-EREIVAAKNWSGAR(SEQ ID NO: 2568)(8)	GST(27)	G53 (66)		
												IKGLVQPTR(SEQ ID NO: 2561)(2)-EIVAAKNWSGAR(SEQ ID NO: 2568)(6)	GST(11)	G53 (66)		
G54	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVKPGESLKLSCVASGETLSSYIMGWFRQAPGKEREFVAAVSWSGNQQDYADSVKGRFTISRDNAEKTVDLQMNSLNPEDTAVYYCAGDQIGFWSSRTQAHEYEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 54	0	1	0	Yes	/	1.537	/	/	/	/	/	/	/
G55	Chymo	MASMTGGQQMGRNSAHVQLVESGGGLVQAGGSLRLSCAASEDTFDNYAVAWFRQARGKEREFVAVISWGGGRSTDYTDSVKGRFSISRDNAKNTVDLQMSSLKPDDTAVYYCHAQYYYEDGYEHESWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 55	/	0	/	No	/	/	/	/	/	/	/	/	/
G56	Trypsin/Chymo	MASMTGGQQMGRNSAHVQLVESGFAFSSYAMSWVRQAPTYGREWVAGIYNDGSHIYYADSVKGRFSISRDNVGNTLYLQLNSLQPNDTALYRCVQEHARGFGGWGNPNPTDLVYRAWGRGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 56	1	/	0	No	/	/	/	/	/	/	/	/	/
G57	Trypsin	MASMTGGQQMGRNSAEVQLVESGGGLVQPGGSLRLSCAASGFTLDYYAIGWFRQAPGKEREGVSCISSSDGSTYYADSVKGRFTISRDNAKNTVDLQMDRLKPEDTAVYYCAADRGSLYSSGRARAQDYTYWGRGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 57	0	1	/	Yes	Yes	/	/	/	/	/	/	/	/
G58	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAGSGDTFSRYTLGWFRQAPGKEREFVAGISWSGGSTSYANSVKGRFTISRDNAKNTMYLQMNSLKPEDTAVYTCAAPGLPGTVVVGASDFYVYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 58	/	/	0	Yes	/	1.537	/	/	/	/	/	/	/

G59	Trypsin/Chymo	MASMTGGQQMGRNSADVQLVESGGSLRLSCAASGRTINTVGLAWFRQAPGQQRDFVAGIEIGGALRYADSVQGRFTVSRDNAKNTMYLQMNSLKPEDTAVYYCGASRGFNIGINPLGYGGWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 59	1	/	0	Yes	Yes	/	/			/	/	/	/	/	/
G60	Chymo	MASMTGGQQMGRNSAEVQLVESGGSLRLSCAASGSGFSSSIIAWYRQAPGKQRELVAAIGGPGSTNYADFVEGRFTISRDNAKNTGYLQMNNLNPEDTAVYYCNEVTRSGREYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 60	0	/	0	Yes	Yes	2.489	4.87E+03	5.86E-03	1.20E-06	/	/	/	/	/	/
G61	Chymo	MASMTGGQQMGRNSAQVQLVESGGSLRLSCVASGHTDCISGMGWYRQAPGKERELVAVLIGGGNTYYGDSVKGRFTISKDKAKNTLYLQMKTLKPEDTAVYYCTADDHGSECPNKEMSSTSTYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 61	/	0	0	No	/	/	/			/	/	/	/	/	/
G62	Trypsin	MASMTGGQQMGRNSAEVQLVESGGALVQAGGSLRLSCLVSGNIYNIKSVGWYRQAPGKEREDNVKNTVDLQMNSLKPEDAAVYYCNARDSSRPRSLPASPESLDGRMDVWGKGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 62	/	/	0	No	/	/	/			/	/	/	/	/	/
G63	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGLVQPGGSLRLSCKASGFAFSSYAMSWVRQAPRYGREWVAGIYNDGSHIYYADSVKGRFSISRDNVGNTLYLQLNSLQPNDTALYRCVQEHERGFGGWGNPNPTDLVYRAWGRGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 63	/	0	0	No	/	/	/			/	/	/	/	/	/
G64	Chymo	MASMTGGQQMGRNSAHVQLVESGGGLVQAGGSLRLSCKVSGTTFSNSAIGWYRQAPGNRRELVATINYGGSTNYADSGKGRFTISKDNAKNTVYLQMNSLKPEDTAVYYCKTTEWREDGYEYDVWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 64	2	0	2	Yes	/	5.676	/			/	/	/	/	/	/

G65	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCATSGRTFSTYALGWFRQRPGKEREFVATIHWSDGRTLYADSVKGRFTLSRDNAQNTVYLQMNSLKPEDTAIYYCAASIYRIGSYDVSTSQGYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 65	/	/	2	Yes	/	3.971	/	/	/	/	/	/	/
G66	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCATSGRTFSTYAMGWFRQRPGKEREFVATIHWSDGRTLYTDSVKGRFTLSRDNAQNTVYLQMNSLKPEDTAVYYCAAATYRIGSYDVSTSQGYNYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 66	/	2	2	Yes	/	4.291	/	DSS	IEAIPQIDKYLK (SEQ ID NO: 2555)(9)-QRPGKER (SEQ ID NO: 2575)(5)	GST(191)	G66 (58)	Seq_58516	E3
G67	Trypsin/Chymo	MASMTGGQQMGRNSAHVQLVESGGGLVQAGGSLRLSCVASGHTVSNYAMAWFRQAPGKEREFVAGISWRATLTYYRDSVKGRFTISRDNAKNTVYLQMSSLKPEDTAVYFCASDRTPYVSRGTSLVEYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 67	1	0	/	Yes	/	3.604	/	/	/	/	/	/	/
G68	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCTASGSIFSVNVMDWYRQAPGKQREFVATITGSGATNYADSVKGRFTISRGSAKNTVYLQMNSLKPDDTAVYYCHNADYREDGYEYDNWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 68	2	/	1	Yes	/	1.075	/	/	/	/	/	/	/
G69	Trypsin/Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCVDSGRTFSSNTMGWFRQAPGKDRDFVAAINRSGVITNYADSVKGRFTISRDNAKNTVYLQLNSLKPEDTAVYYCAARAGGWPSQIPVEYDRWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 69	/	/	2	Yes	/	3.687	/	DSS	QAPGKDRDFVAAINR(SEQ ID NO: 2569)(5)-YLKSSK (SEQ ID NO: 2557)(3)	GST(194)	G69 (58)	Seq_20239	E3
												QAPGKDRDFVAAINR (SEQ ID NO: 2569)(5)-RIEAIPQIDKYLK (SEQ ID NO: 2537)(10)	GST(191)	G69 (58)		
												NTVYLQLNSLKPEDTAVYYCAAR (SEQ ID NO: 2570)(11)-IAYSKDFETLK (SEQ ID NO: 2552)(5)	GST(113)	G69 (102)		
												SGVITNYADSVKGR (SEQ ID NO:	GST(18	G69		

												2571)(12)-KRIEAIPQIDK (SEQ ID NO: 2547)(1)				
G70	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQAGGSLRLSCAASGATFSINAIGWYRQAPGKQRELVAVIKSGNSINYADSVKGRFTISRDHAKNTVYLQMNNLKPEDTAVYYCHADQPPETGWGTWNDLWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 70	1	1	2	Yes	/	1.755	/	/	/	/	/	/	/
G71	Trypsin/Chymo	MASMTGGQQMGRNSAQVQLVESGGGSVQAGGSLRLSCAASGRTSVSYAMGWFRQAPGKEREFVAAVSRSGTNLYYADSVKGRFTISRHTAENTMYLQMNSLLPEDTALYYCAADEALRWGIGTQPRSEFFDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 71	1	2	/	No	/	/	/	/	/	/	/	/	/
G72	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQPGGSLRLSCATSGSTFSINGIGWYRQVPGIEREFVAGVSTDGKANYADSVAGRFTISINDGKNTAYLQMNSLKPEDTAVYYCNVDSTKGYYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 72	/	/	1	Yes	/	0, No binding	/	/	/	/	/	/	/
G73	Chymo	MASMTGGQQMGRNSAEVQLVESGGGLVQAGGSLRLSCAASGRTFSDDAMAWFRQAPGKEREFVAAISWHPENTFYADSVKGRFTISRDKTKNTEYLQMNSLKPEDTAVYYCAAGPRLEIGDYAQYKYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 73	/	/	1	Yes	/	0.9549	/	/	/	/	/	/	/
G74	Chymo	MASMTGGQQMGRNSAQVQLVESGGGLVQPGGSLTLSCAASGSTIDDGIGWFRQASGKEREGVSCIRLSDGSKYYRDIVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCANGPCTGPRAIAEILYGAWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 74	/	/	1	Yes	Yes	/	/	/	/	/	/	/	/
G75	Chymo	MASMTGGQQMGRNSAHVQLVESGGGLVQAGASLRLTCGPSGRSVGLYTMGWFRQAPGKEREFVAGVTYLGDTTTYSDAVKGRFTISRENNKNTVYLRMNSLKPEDTAVYYCTATATGWGSPIPSAPGRWGYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 75	/	/	0	Yes	Yes	/	/	/	/	/	/	/	/

G76	Trypsin	MASMTGGQQMGRNSAQVQLVESGGGLVQPGGSLRLSCVVSGFPFSEYAMSWVRQTPEKGREWVSGIYTDGSETLYENSVKGRFTISRDNTKNTLYLQMNNLKPEDTARYYCKLGDPYGVGLRDYEYLGHGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 76	/	/	0	Yes	/	0, No binding	/	/	/	/	/	/	/
G77	Chymo	MASMTGGQQMGRDPAQVQLVESGGGLVQAGGSLRLSCTASRSTFRVNPAGWYRQAPGKERELVARITSGGSTNYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCNVPYYMEDGYEHDAWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 77	2	1	2	Yes	/	6.012	/	/	/	/	/	/	/
G78	Chymo	MASMTGGQQMGRDPADVQLVESGGGLVQAGGSLRLSCTASQSILYINVMGWYRQAPGKQRELVAEIPTGGNTDYADSVKGRFTISRDNVKNTVSLQMNSLKPEDTAVYYCNVRGGVLSPYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 78	2	0	2	Yes	/	2.082	/	DSS	NTVSLQMNSLKPEDTAVYYCNVR (SEQ ID NO: 2572)(11)-IAYSKDFETLK (SEQ ID NO: 2552)(5)	GST(113)	G78 (101)	Seq_6584	E2
G79	Chymo	MASMTGGQQMGRDPADVQLVESGGGLVQAGGSLRLSCATSGRTFSTYAAGWFRQRPGKEREFVATIHWNDGRTLYADSVKGRFTLSRDNAQNTVYLQMNSLKPEDTAVYYCAAYTYRIGSYDVSTSQGYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 79	2	2	2	Yes	/	2.07	/	/	/	/	/	/	/
G80	chymo	MASMTGGQQMGRDPAQVQLVESGGGLVQAGASLRLSCAASRGTFSSYTMGWFRQAPGKERLFVASISRDGGTTYYADSVKGRFTISRDNAENILYLQMNSLKPEDTAVYYCAAASRHPSTWEVWGLEYYYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 80	2	1	2	Yes	/	4.428	/	EDC	DEGDKWR (SEQ ID NO: 2573)(2)-QAPGKER (SEQ ID NO: 2574)(5)	GST(37)	G80 (58)	Seq_51356	E3
												DGGTTYYADSVKGR (SEQ ID NO: 2576)(12)-DEGDKWR (SEQ ID NO: 2573)(2)	GST(37)	G80 (80)		
												DPAQVQLVESGGGLVQAGASLR (SEQ ID NO: 2577)(1)-KRIEAIPQIDK (SEQ ID NO: 2547)(1)	GST(181)	G80 (13)		
												DEGDKWR (SEQ ID NO: 2573)(1)-QAPGKER	GST(36)	G80 (58)		

														(SEQ ID NO: 2574)(5)				
G81	chymo	MASMTGGQQMGRDPAEVQLVESGGELVQAGGSLRLSCAASGRTDSVTRMAWFRQAPGKEREFVAAITWSSGYTYYPDSVKGRFTISRDNAKNTMYLQMNSLKAEDTAVYICAAAVGVISEYNSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 81	2	1	2	Yes	/	0, No binding	/			/	/	/	/	/	/
G82	Chymo	MASMTGGQQMGRDPAHVQLVESGGGLVQAGGSLRLSCAASGKIFSLSTMGWYRQAPGKQRELVAALTSGGSTNYADSVKGRFTISRDNAKYTTYLQMNSLKPEDTAVYYCNVRYYSGYDGYESNSWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 82	2	1	2	Yes	/	7.218	1.11E+06	5.61E-04	5.04E-10	EDC	DNAKYTTYLQMNSLKPEDTAVYYCNVR (SEQ ID NO: 2578)(4)-DEGDKWR (SEQ ID NO: 2573)(1)	GST(36)	G82(90)	Seq_1411	E2
														DPAHVQLVESGGGLVQAGGSLR (SEQ ID NO: 2579)(1)-IAYSKDFETLKVDFLSK (SEQ ID NO: 2580)(5)	GST(113)	G82(13)		
														DPAHVQLVESGGGLVQAGGSLR (SEQ ID NO: 2579)(1)-DFETLKVDFLSK (SEQ ID NO: 2550)(6)	GST(119)	G82(13)		
														KFELGLEFPNLPYYIDGDVK (SEQ ID NO: 2581)(7)-QAPGKQR(SEQ ID NO: 2551)(5)	GST(51)	G82(58)		
G83	Trypsin	MASMTGGQQMGRDPAQVQLVESGGGLVQAGGSLRLSCAASRRTFSIYNMGWFRQAPGKEREFVATITRYGDRTYTADSVKGRFTISSDQAKNTVYLQMNSLNPHDTAVYYCAADSAYSGPDFKHYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 83	2	1	2	Yes	/	8.716	9.87E+05	7.51E-04	7.61E-10	EDC	DPAQVQLVESGGGLVQAGGSLR (SEQ ID NO: 2582)(1)-YLKSSK (SEQ ID NO: 2557)(3)	GST(194)	G83(13)	Seq_22759	E1
														DPAQVQLVESGGGLVQAGGSLR (SEQ ID NO: 2582)(1)-RIEAIPQIDKYLK (SEQ ID NO: 2537)(10)	GST(191)	G83(13)		

G84	Chymo	MASMTGGQQMGRDPADVQLVESGGGLVQPGGSLRLSCAASGSTFSINAIGWYRQAPGKEREFVAALRWPGNIWYYADFVEGRITISRDNAKNTVYLQMNSLKPEDTAVYYCAATVGLDSPPRNEYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 84	2	/	2	No	/	/	/	/	/	/	/	/	/
G85	Trypsin	MASMTGGQQMGRDPAHVQLVESGGGLVQPGGSLRLSCAASGFTFSTYAMGWVRQAPGKGPEWVATIYSKGDTTHYANSAKGRFTISRDNARNTLYLQMNSLKPEDTAVYYCAKGISDSYLRVESNYRGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 85	2	0	/	Yes	/	0, No binding	/	/	/	/	/	/	/
G86	Trypsin	MASMTGGQQMGRDPAEVQLVESGGGLVQAGDSLRLSCAASGRTFSSYTMGWFRQAPGKEREFVAGIRWSGGSTYFTNYEDSVKGRFTISKDNAKNTVFLQMNSLRPEDTAVYYCAFTGHYSTYDSPQRYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 86	2	/	/	Yes	/	5.542	/	/	/	/	/	/	/
G87	Trypsin	MASMTGGQQMGRDPAEVQLVESGGGLVQAGDSLRLSCAASGRTFSSYNLGWFRQAPGKEREFVAVMNCRYGDTDYPDSVKGRFTMSRDNAKNTLYLEMNNLKPEDTAVYYCAAKVLAYCGSGYYYRRNDYGYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 87	0	1	1	Yes	/	0, No binding	/	/	/	/	/	/	/
G88	Chymo	MASMTGGQQMGRDPAEVQLVESGGGLVKPGESLKLSCVASGETLSSYIMGWFRQAPGKEREFVAAVSWSGNQQDYADSVKGQFTISRDNAEKTVDLQMNSLNPEDTAVYYCAGDQMGFWSSRTQAHEYEYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 88	0	1	0	No	/	/	/	/	/	/	/	/	/
G89	Chymo	MASMTGGQQMGRDPAEVQLVESGGGLVQAGGSLSLSCAASGSINSINAMGWFRQAPGKQRELVATITRGGSTNYADSVKGRFTISIDNAKNTVYLQMNSLKPEDTAVYYCNADRGTDDGWLYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 89	0	/	2	Yes	/	5.736	/	/	/	/	/	/	/

G90	Trypsin	MASMTGGQQMGRDPADVQLVESGGGLVQAGGSLRLSCAASGLTFSNYAMGWFRQAPGKEREFAAGITWNGGASHYADSVKGRFTISRDNAQNTVYLQMNSLKPEDTAVYYCAARLGSVAYPGLRYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 90	1	0	2	Yes	/	5.832	/			EDC	MASMTGGQQMGRDPADVQLVESGGGLVQAGGSLR (SEQ ID NO: 2583)(13)-SPILGYWK (SEQ ID NO: 2584)(1)	GST(2)	G90(13)	Seq_13998	E3
G91	Trypsin/Chymo	MASMTGGQQMGRDPAEVQLVESGGGLVQAGGSLRLSCAASGRTFSDYPMAWFRQALGKEREFLATISTSGSRTMYADSVKGRFTISRDNAKNMMYLQMNSLKPEDAAVYYCAARQGSYYSDYNRALPGEYHYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 91	0	/	/	Yes	/	1.145	/			/	/	/	/	/	/
G92	Chymo	MASMTGGQQMGRDPADVQLVESGGGLVKPGESLKLSCVASGETLSSYIMGWFRQAPGQGRKFVGGINYSGSSVEYADSVKGRFTISRDNAKNTMYLQMNSLKPEDTAAYYCASSRGYNTGTNPLGYNYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 92	0	/	0	No	/	/	/			/	/	/	/	/	/
G93	Chymo	MASMTGGQQMGRDPAQVQLVESGGGLVQPGGSLRLSCAASGSGFSSSIIGWHRQAPGKQRELVAAIGGPGSTNYADSVKGRFTISRDNAKNTAYLQMNNLKPEDSAVYYCEATTRSGREYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 93	0	/	/	Yes	/	2.16	/			/	/	/	/	/	/
G94	Chymo	MASMTGGQQMGRDPAHVQLVESGGGLVQPGGSLRLSCVASGFTFSAYAMSWVRQVPGKGREWISGIYNDGSNIYYTDSVKGRFSISRDNAKNTLYLQMNNLKPDDTAVYYCTKEHARGFGGRGNPNPSDLVYDAWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 94	0	0	0	No	/	/	/			/	/	/	/	/	/
G95	Trypsin/Chymo	MASMTGGQQMGRDPAHVQLVESGGGLVQAGGSLRLSCAASGRTFSSYAMAWFRQAVGKEREFVAAVSRSGTNLYYADSVKGRFTISRDTAKNTMYLQMNSLKPEDTALYYCAAGEALRWGIGQQPRSEFFDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 95	/	2	1	Yes	/	3.76	1.39E+04	1.07E-03	7.70E-08	/	/	/	/	/	/
G96	Chymo	MASMTGGQQMGRDPADVQLVESGGGLVQAGGSLKLSCAAFGVTFDINTIAWYRQAPGKQREFVAHITSGGTTYYADSVKARFTMSRDSAKNTVYLQMNNLK	SEQ ID NO: 96	/	0	/	Yes	/	6.769	/			EDC	DPADVQLVESGGGLVQAGGSLK (SEQ ID NO: 2585)(1)-YEEHLYERDEGDKWR	GST(40)	G96 (13)	Seq_17861	E1

		PEDTAVYYCNVNPTWPYSGEVDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH										(SEQ ID NO: 2539)(13)				
												DPADVQLVESGGGLVQAGGSLK (SEQ ID NO: 2585)(4)-YEEHLYERDEGDKWR (SEQ ID NO: 2539)(13)	GST(40)	G96 (16)		
												LLLEYLEEKYEEHLYERDEGDK (SEQ ID NO: 2586)(9)-DPADVQLVESGGGLVQAGGSLK (SEQ ID NO: 2585)(1)	GST(27)	G96 (13)		
												DPADVQLVESGGGLVQAGGSLK (SEQ ID NO: 2585)(1)-MSPILGYWKIK(SEQ ID NO: 2587)(9)	GST(9)	G96(13)		
G97	Chymo	MASMTGGQQMGRDPAEVQLVESGGGLVQAGGSLRLSCTASRSTFRVNPAGWYRQAPGKERELVARITSGGSTNYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCNVPCYMEDGYEHDAWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 97	/	2	/	Yes	/	3.298	/	/	/	/	/	/	/
G98	Chymo	MASMTGGQQMGRDPAQVQLVESGGGLVQAGDSLRLSCATSGRTFSTYAAGWFRQRPGKEREFVATIHWNDGRTLYADSVKGRFTLSRDNAQNTVYLQMNSLKPEDTAVYYCAASTYRIGSYDVSTSQGYDYWGQGTQVTVSSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 98	2	2	2	Yes	/	4.669	/	/	/	/	/	/	/

Table 2. Summary of HSA Nbs and their biophysical and physiochemical properties.

ID	Enzyme	Protein Sequence	SEQ ID NO	salt trend	lowpH trend	highpH trend	Soluble	ELISA affinity (LogIC50 (oD450nm))	Mutant Screening	SPR ka (1/Ms)	SPR kd (1/s)	SPR KD (M)	Cross-linker	Cross-linked Peptides	CX residue on Nbs	CX residue on HSA	CX Model Folder	CX Model Epitope
H1	Trypsin	MASMTGGQQMGRDPGSSSGSMAQVQLVESGGGLVQPGGSLRLSCVASGIMFDIYTMRWYRQAPGKQRELVAAITGAGRANYNDDSVKGRFTISRDNAKNTVYLQMNRMKPEDTALYECNTEILGGGPNYWGRGTQVTVSEPKTPKGGKGGGLEHHHHHH	SEQ ID NO: 99	2	1	2	Yes	4.916	/	9.73E+06	1.19E-03	1.22E-09	DSS	SLHTLFGDKLCTVATLR (SEQ ID NO: 2588)(9)-DNAKNTVYLQMNR (SEQ ID NO: 2589)(4)	H1 (98)	HSA(97)	Seq_16529	E1
														DNAKNTVYLQMNR (SEQ ID NO: 2589)(4)-FPKAEFAEVSK (SEQ ID NO: 2590)(3)	H1 (98)	HSA(249)		
														ANYNDDSVKGR (SEQ ID NO: 2591)(9)-KYLYEIAR (SEQ ID NO: 2592)(1)	H1 (86)	HSA(161)		
H2	Trypsin/Chymo	MASMTGGQQMGRDPENLYFQGAQVQLVESGGGLVQAGGSLRLSCTASGRTFTPYTIGWFRQAPGKEREFVASILWSGINTDYADSVKGRFAISRDNAKNAAYLQMSNLKPEDTAVYYCATGGGLGYYRSVSQYDYWGQGTQVTVSEPKTPKGGKGGGLEHHHHHH	SEQ ID NO: 100	2	/	2	Yes	5.883	Decreased	2.34E+05	3.99E-05	1.70E-10	DSS	SVSQYDYWGQGTQVTVSEPKTPK (SEQ ID NO: 2593)(20)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	H2 (127)	HSA(375)	Seq_8598	E2
														DNAKNAAYLQMSNLKPEDTAVYYCATGGGLGYYR (SEQ ID NO: 2595)(4)-KVPQVSTPTLVEVSR (SEQ ID NO: 2596)(1)	H2 (77)	HSA(438)		
H3	Chymo	MASMTGGQQMGRDPENLYFQGAQVQLVESGGGLVQAGGSLRLSCVASGRTFEPFVMGWFRQAPGKEREFVATISWSGGSLSYADSVKGRFTVSRDNAKNTVYLQM	SEQ ID NO:	1	0	2	Yes	7	Decreased	1.11E+06	5.04E-04	4.54E-10	EDC	YTFQYDYWGQGTQVTVSEPK (SEQ ID NO: 2597)(6)-VFDEFKPLVEEPQNLIK	H2 (134)	HSA(402)	Seq_14034	E2

		NSLKPEDTAVYYCAAAPGVGNYRYTFQYDYWGQGTQVTVSEPKTPKGGKGGGLEHHHHHH												(SEQ ID NO: 2598)(6)				
														YTFQYDYWGQGTQVTVSEPK (SEQ ID NO: 2597)(6)-ATKEQLK (SEQ ID NO: 2599)(3)	H3 (134)	HSA(565)		
														DPENLYFQGAQVQLVESGGGLVQAGGSLR (SEQ ID NO: 2600) (1)-TYETTLEKCCAAADPHECYAK (SEQ ID NO: 2601)(8)	H3 (13)	HSA(383)		
H4	Trypsin	MASMTGGQQMGRDPNSAHVQLVESGGGLVQTGGSLRLACAASGRAFSTYAMGWFRQAPGKEREFVASINRSGSSTYYADSVKGRFTISRDNGKDTVYLQMNRLIPEDTAVYYCAADSEGVGFRNMLEYDYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 102	0	0	2	Yes	5.4	No change	9.92E+05	2.64E-04	2.66E-10	DSS	VFDEFKPLVEEPQNLIK (SEQ ID NO: 2598)(6)-SGSSTYYADSVKGR (SEQ ID NO: 2602)(12)	H4 (82)	HSA(402)	Seq_29830	E2
														CCAAADPHECYAKVFDEFKPLVEEPQNLIK (SEQ ID NO: 2628)(13)-SGSSTYYADSVKGR (SEQ ID NO: 2602)(12)	H4 (82)	HSA(396)		
													EDC	VFDEFKPLVEEPQNLIK (SEQ ID NO: 2598)(6)-SGSSTYYADSVK (SEQ ID NO: 2603)(9)	H4 (79)	HSA(402)		
														VFDEFKPLVEEPQNLIK (SEQ ID NO: 2598)(10)-SGSSTYYADSVKGR (SEQ ID NO: 2602)(12)	H4 (82)	HSA(406)		
H5	Chymo	MASMTGGQQMGRDPNSAEVQLVESGGGLVQAGGSLRLSCAASGRTFIPYTTGWFRQTPGKEREFVATITWSGISTKYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCTKNPRALALNRDYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 103	1	0	2	No	/	/	/			/	/	/	/	/	/

H6	Chymo	MASMTGGQQMGRDPNSAEVQLVESGGGLVQVGGSLTLSCAAAGSTFTTNAMAWFRQFPGKERELVAAISWGGLGYVADSVRGRFTISRPTKNMMILQLNSLEREDTAIYYCAARKMSTVATEATMYAYWGHGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 104	2	1	/	Yes	2.905	No change	/			DSS	KMSTVATEATMYAYWGHGTQVTVSSEPK (SEQ ID NO: 2604)(1)-DVCKNYAEAK (SEQ ID NO: 2605)(4)	H6 (115)	HSA(341)	Seq_35308	E2
H7	Chymo	MASMTGGQQMGRDPNSADVQLVESGGGSVQAGGSLRLSCAASGGTFSSYAMGWYRQAPGKEREFVSGISWSGSSIDYVDSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCGAADPMGLGYGLGPRPVDRLLSAECDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 105	1	0	2	Yes	5.621	No change	9.35E+04	9.79E-05	1.09E-09	DSS	DNAKNTVYLQMNSLKPEDTAVYYCGAADPMGLGYGLGPRPVDR (SEQ ID NO: 2606)(4)-RDAHKSEVAHR (SEQ ID NO: 2607)(5)	H7 (93)	HSA(28)	Seq_45799	E3
														EREFVSGISWSGSSIDYVDSVKGR (SEQ ID NO: 2608)(22)-LKECCEK (SEQ ID NO: 2609)(2)	H7 (82)	HSA(300)		
H8	Chymo	MASMTGGQQMGRDPNSADVQLVESGGGLVQAGGSLRLSCAASGRTFSSYAMGWFRQAPGKEREFVSAISRSGGSTYYTDSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAAAEGLASGSYDYTPPLKSSWYDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 106	/	0	2	No	/	/	/			/	/	/	/	/	/
H9	Trypsin	MASMTGGQQMGRDPNSAEVQLVESGGGLVQAGGSLRLSCVASGRTFSYRAMGWFHQAPGKEREFVAAVGSSGLTTYYADSVKGRFTISRDNAKNTVYLQMNSLQLEDTAVYYCAAAKFGYVVVTAKEYEYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 107	2	/	2	No	/	/	/			/	/	/	/	/	/
H10	Chymo	MASMTGGQQMGRDPNSADVQLVESGGGLVQAGGSLRLSCRASGLPFGPYTMGWFRQTPGQEREFVAAITWSSMNTNYADSVKGRFTISRDSAKNTVYLQMNTLKPDDTAVYYCAADDRAVPMLGDFEDYIYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 108	2	/	/	Yes	1.462	/	/			/	/	/	/	/	/
H11	Trypsin	MASMTGGQQMGRDPNSAQVQLVESGGGLVQVGGSLRLSCAASGRTFSNYVMGWFRQAPGKEREFVAYIHWSGSSTSYADSVKGRFTISRDNTKNTMYLQMNSLKPEDTAVYYCTADQYASTLLR	SEQ ID NO: 109	2	/	2	Yes	6.272	No change	2.55E+05	1.55E-05	6.10E-11	DSS	EFVAYIHWSGSSTSYADSVKGR (SEQ ID NO: 2610)(20)-ATKEQLK (SEQ ID NO: 2599)(3)	H11 (82)	HSA(565)	Seq_20104	E4

		AAGEYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH												DNTKNTMYLQMNSLKPEDTAVYYCTADQYASTLLR (SEQ ID NO: 2611)(4)-AVMDDFAAFVEKCCK (SEQ ID NO: 2612)(12)	H11 (93)	HSA(581)		
														DNTKNTMYLQMNSLKPEDTAVYYCTADQYASTLLR (SEQ ID NO: 2611)(4)-CCKADDKETCFAEEGK (SEQ ID NO: 2613)(3)	H11 (93)	HSA(584)		
H12	Chymo	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCVSSGRTYRWNAMGWFRQAPGKEREFVAAIDWDGRNTDYADSVKGRFTISRDNAKNTVFLQMNRLKSEDTAVYSCALDRVVITSMRTNFDVWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 110	2	0	/	Yes	1.617	No change	/			/	/	/	/	/	/
H13	Trypsin	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCAASGRTFSTYHMGWFRQAPGKAREFVAAITGSGGITYYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAADTRAYGLVPSTTSSRYNYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 111	/	/	2	Yes	1.962	/	/			/	/	/	/	/	/
H14	Trypsin	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCTASGRTFTPYTMGWFRQAPGKEREFVASILWSGNNRDYADSVKDRFAISRDNAKNTAYLQMNSLKPEDTAVYYCAAGDGLGFYRSVNQYDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 112	2	1	2	Yes	5.964	Decreased	2.50E+05	5.57E-06	2.23E-11	/	/	/	/	/	/
H15	Trypsin	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGDSLRLSCAASERTSNYAMGWFRQAPGKEREFVADINHTGGRRKYGDSVKGRFTISRDNAENMVYLQMNNLQVEDTAVYYCATGLRYDVSGYAPDYRYWGRGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 113	2	0	0	Yes	No binding	/	/			/	/	/	/	/	/

H16	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQTGGSLTLSCAASGRTFSTKSMGWFRQAPGKEREFVADINWNGGITHYADSVEGRFTISRDNANDMVYLQMNSLKPEDTAVYYCAGGRYSTLFSKSEADYDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 114	/	0	2	Yes	5.986	No change	1.34E+06	1.16E-04	8.62E-11	/	/	/	/	/	/
H17	Trypsin	MASMTGGQQMGRDPNSAQVQLVESGGGLAQAGGSLRLSCAASGGTFSNSCMGWFRQAPGMEREFVVIIRSTGHTTYADSVEGRFTVSREIAKNTVYLEMNSLKPEDTAVYVCAAGVSDYGCYRTSGINYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 115	2	0	2	No	/	/	/			/	/	/	/	/	/
H18	Trypsin	MASMTGGQQMGRDPNSAEVQLVESGGGLVQAGGSLRLSCTASGPKDTPYTMGWFRQVPGKEREFVASVLWSGINTDYADSVKGRFAISRNNAKNTMYLQMNSLKPEDTAVYYCAAGYGLGFYRSISQYDYWGHGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 116	2	2	2	Yes	5.646	Decreased	1.71E+05	8.71E-05	5.02E-10	DSS	LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	H18 (45)	HSA(375)	Seq_45285	E2
														LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-KVPQVSTPTLVEVSR (SEQ ID NO: 2596)(1)	H18(45)	HSA(438)		
														LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-VTKCCTESLVNR (SEQ ID NO: 2615)(3)	H18(45)	HSA(499)		
														LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-ATKEQLK (SEQ ID NO: 2599)(3)	H18(45)	HSA(565)		
													EDC	LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-RPCFSALEVDETYVPK (SEQ ID NO: 2616)(8)	H18(45)	HSA(516)		
														DPNSAEVQLVESGGGLVQAGGSLR (SEQ ID NO: 2617)(1)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	H18(13)	HSA(375)		

														DPNSAEVQLVESGGGLVQAGGSLR (SEQ ID NO: 2617)(6)-VTKCCTESLVNR (SEQ ID NO: 2615)(3)	H18(18)	HSA(499)		
														DPNSAEVQLVESGGGLVQAGGSLR (SEQ ID NO: 2617)(1)-VTKCCTESLVNR (SEQ ID NO: 2615)(3)	H18(13)	HSA(499)		
														LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-VFDEFKPLVEEPQNLIK (SEQ ID NO: 2598)(11)	H18(45)	HSA(407)		
														LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-CCTESLVNR (SEQ ID NO: 2618)(4)	H18(45)	HSA(503)		
H19	Trypsin	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCTASGPKDTPYTMGWFRQVPGKEREFVASVLWSGINTDYADSVKGRFAISRNNAKNTMYLQMNSLKPEDTAVYYCAAGYGLGFYRSVSQHDYWGHGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 117	2	2	2	No	/	/	/			/	/	/	/	/	/
H20	Trypsin	MASMTGGQQMGRDPNSAEVQLVESGGGLVQAGGSLRLSCTASGPKDTPYTMGWFRQVPGKEREFVASVLWSGINTDYADSVKGRFAISRNNAKNTMYLQMNSLKPEDTAVYYCAAGYGLGFYRTVSQYDYWGHGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 118	2	2	2	Yes	5.759	Decreased	1.54E+05	3.80E-05	2.47E-10	DSS	LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	H20(45)	HSA(375)	Seq_45284	E2
														LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-VTKCCTESLVNR (SEQ ID NO: 2615)(3)	H20(45)	HSA(499)		
														LSCTASGPKDTPYTMGWFR (SEQ ID NO: 2614)(9)-KVPQVSTPTLVEVSR (SEQ ID NO: 2596)(1)	H20(45)	HSA(438)		
													EDC	LSCTASGPKDTPYTMGWFR	H20(	HSA		

														(SEQ ID NO: 2614)(9)-RPCFSALEVDETYVPK (SEQ ID NO: 2616)(8)				
														DPNSAEVQLVESGGGLVQAGGSLR (SEQ ID NO: 2617)(1)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	H20(13)	HSA(375)		
														DPNSAEVQLVESGGGLVQAGGSLR (SEQ ID NO: 2617)(6)-VTKCCTESLVNR (SEQ ID NO: 2615)(3)	H20(18)	HSA(499)		
														DPNSAEVQLVESGGGLVQAGGSLR (SEQ ID NO: 2617)(1)-VTKCCTESLVNR (SEQ ID NO: 2615)(3)	H20(13)	HSA(499)		
H21	Trypsin	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGGSLRLSCAASGYTSGNDAMGWFRQAPGKEREFVGAIRWSGVSTYYADSVKGRFTISRDGAKNTLYLQMNSLKPEDTAVYYCAAKFTGSAWYGVQKLESTYWDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 119	1	0	2	Yes	5.628	No change	6.83E+05	1.82E-04	2.66E-10	DSS	WSGVSTYYADSVKGR (SEQ ID NO: 2619)(13)-LKECCEK (SEQ ID NO: 2609)(2)	H21(82)	HSA(300)	Seq_6523	E5
														DGAKNTLYLQMNSLKPEDTAVYYCAAK (SEQ ID NO: 2620)(4)-AAFTECCQAADKAACLLPK (SEQ ID NO: 2621)(12)	H21(93)	HSA(198)		
H22	Trypsin	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCTASARTSNAMGWFRRAPGKERDFVAAISESGRTTDYADSVKGRFTISRDTAKNTVYLQMISLKPEDTAVYYCARKRVADAISSNYEFRYDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 120	1	0	2	Yes	4.211	No change	/			DSS	TPKGGCGGAAALEHHHHHH (SEQ ID NO: 2622)(3)-AFKAWAVAR (SEQ ID NO: 2623)(3)	H22(147)	HSA(236)	Seq_2558	E3
H23	Chymo	MASMTGGQQMGRDPNSADVQLVESGGGLVQAGGSLTLSCAASGRTFSSSTMGWFRRAPGKEREFVAAISGSARTTDYADSVKGRFTISRDNAKNTVYLQMISLKPEDTAIYYCARKRVVDVTTSNYELRYDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 121	1	0	2	Yes	1.625	/	/			/	/	/	/	/	/

H24	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGGSLRLSCVSSGRTYRWNAMGWFRQAPGKEREFVAAIDWDGRNTDYADSVKGRFTISRDNAKNTVYLQMNSLKVEDTAIYYCAAREWGSGGYSSIASYAYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 122	2	2	/	Yes	No binding	/	/			/	/	/	/	/	/
H25	Trypsin	MASMTGGQQMGRDPNSADVQLVESGGGLVQAGGSLRLSCAASGRTISDYGMAWFRQAPGKEREFVGVITSNSVTTYYADSVKGRFTISRDNTKNTVYLQMISLKPEDTAIYYCAARIPVGFYYNARNYDFWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 123	1	/	2	Yes	5.344	No change	9.36E+05	8.02E-04	8.58E-10	DSS	NYDFWGQGTQVTVSSEPKTPK (SEQ ID NO: 2623)(18)-KYLYEIAR (SEQ ID NO: 2592)(1)	H25(144)	HSA(161)	Seq_4162	E3
H26	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGGSLRLSCAASGRTPYVMGWFRQAPGNEREFVASISWTYGYTNYANSVKGRFRISKDNAKNTVLLQMNSLKPEDTAVYYCAARRGEDPEYDYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 124	1	0	2	Yes	2.813	No change	/			DSS	LAKTYETTLEK (SEQ ID NO: 2594)(3)-ISKDNAK (SEQ ID NO: 2625)(3)	H26(87)	HSA(375)	Seq_18634	E2
														VFDEFKPLVEEPQNLIK (SEQ ID NO: 2598)(6)-ISKDNAK (SEQ ID NO: 2625)(3)	H26(87)	HSA(402)		
														DNAKNTVLLQMNSLKPEDTAVYYCAAR (SEQ ID NO: 2626)(4)-VFDEFKPLVEEPQNLIK (SEQ ID NO: 2598)(6)	H26(91)	HSA(402)		
														NTVLLQMNSLKPEDTAVYYCAAR (SEQ ID NO: 2627)(11)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	H26(102)	HSA(375)		
														TYETTLEKCCAAADPHECYAK (SEQ ID NO: 2601)(8)-ISKDNAK (SEQ ID NO: 2625)(3)	H26(87)	HSA(383)		
														CCAAADPHECYAKVFDEFKPLVEEPQNLIK (SEQ ID NO: 2628)(13)-ISKDNAK (SEQ ID NO: 2625)(3)	H26(87)	HSA(396)		

												CCAAADPHECYAKVFDEFKPLVEEPQNLIK (SEQ ID NO: 2628)(13)-DNAKNTVLLQMNSLKPEDTAVYYCAAR (SEQ ID NO: 2626)(4)	H26(91)	HSA(396)		
												EFVASISWTYGYTNYANSVKGR (SEQ ID NO: 2629)(20)-VTKCCTESLVNR (SEQ ID NO: 2615)(3)	H26(80)	HSA(499)		
											EDC	RGEDPEYDYWGQGTQVTVSSEPK (SEQ ID NO: 2630)(6)-ATKEQLK (SEQ ID NO: 2599)(3)	H26(120)	HSA(565)		
												RGEDPEYDYWGQGTQVTVSSEPK (SEQ ID NO: 2630)(8)-ATKEQLK (SEQ ID NO: 2599)(3)	H26(122)	HSA(565)		
												DPNSAQVQLVESGGGLVQAGGSLR (SEQ ID NO: 2631)(1)-ATKEQLK (SEQ ID NO: 2599)(3)	H26(13)	HSA(565)		
H27	Chymo	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCIASGRTFSTYHMGWFREAPGKGREFVAAITQNGGTTYYADSVKGRFTISRDNAKNTVYLQMGSLKPEDTAVYYCAASPALIGRIYFGNENYSWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 125	2	/	0	Yes	2.575	No change	/	DSS	EFVAAITQNGGTTYYADSVKGR (SEQ ID NO: 2632)(20)-DAHKSEVAHR (SEQ ID NO: 2633)(4)	H27(82)	HSA(28)	Seq_10156	E3
												EFVAAITQNGGTTYYADSVKGR (SEQ ID NO: 2632)(20)-ECCEKPLLEK (SEQ ID NO: 2634)(5)	H27(82)	HSA(305)		
												EFVAAITQNGGTTYYADSVKGR (SEQ ID NO: 2632)(20)-FKDLGEENFK (SEQ ID NO: 2635)(2)	H27(82)	HSA(36)		
												FKDLGEENFK (SEQ ID NO: 2635)(2)-	H27(60)	HSA(36		

														EAPGKGR (SEQ ID NO: 2636)(5)				
H28	Trypsin	MASMTGGQQMGRDPNSADVQLVESGGGLAQAGGSLRLSCAASGRTFSNECMGWFRQAPGKEREFVATIRSTGHISYATSVQGRFTVSRDIAKNTVYLEMNNLKPEDTAVYSCGAGVSDYGCYRTSGYNYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 126	2	/	/	Yes	5.116	No change	5.24E+06	8.68E-03	1.66E-09	DSS	NTVYLEMNNLKPEDTAVYSCGAGVSDYGCYR (SEQ ID NO: 2637)(11)-ATKEQLK (SEQ ID NO: 2599)(3)	N28(103)	HSA(565)	Seq_14266	E2
														TSGYNYWGQGTQVTVSSEPKTPK (SEQ ID NO: 2638)(20)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	N28(143)	HSA(375)		
H29	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVPAGGSLRLSCAASGRTFSLYRMGWFRQAPGKEREFVAAIIWSSGSTYYADSVKGRFTISRDIAKNTVYLEMNSLKPEDTAVYSCGAGVSDYGCYRTSGYAYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 127	2	2	1	Yes	1.862	No change	/			/	/	/	/	/	/
H30	Trypsin	MASMTGGQQMGRDPNSAHVQLVESGGGLAQAGGSLRLSCAASGGTFSNSCMGWFRQAPGMEREFVAIIRSTGHTTYADSVEGRFTVSRDIAKNTVYLEMNSLKPEDTAVYSCVAGVSDYGCYRTSGIKYWGQGTQVTVSSEPKTPKGGCGGAAALEHHHHHH	SEQ ID NO: 128	/	0	2	Yes	5.895	No change	1.03E+06	1.68E-04	1.64E-10	/	/	/	/	/	/
H31	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQPGGSLRLSCTPSGFRLEDYPIAWFRQAPGKEREGLSCITSGDGRTYYEESVKGRFTISRDNAQNKVYLQMNKLTPEDTAVYHCATVPSDNLCGYLHRRPFASWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 129	0	0	0	Yes	1.075	/	/			/	/	/	/	/	/
H32	Chymo	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCAASDTIDNYARAWFRQAPGKEREFVAAITWTFGTPYYTDSVKGRFTISRDDAKNTVYLQMNSLKPEDTAVYYCAASLYLPVRTASGGYRLDTDRPQYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 130	/	0	0	Yes	5.335	/	/			/	/	/	/	/	/
H33	Trypsin	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCAASGRTLSSYDMGWFRQPPGKEREFVAAITRHDFNTFYRDSVKGRFTISRDNAKNTVYLQMNSLKSEDT	SEQ ID NO:	1	0	0	Yes	5.033	/	/			DSS	LAKTYETTLEK (SEQ ID NO: 2594)(3)-DSVKGR (SEQ ID NO: 2639)(4)	H33(82)	HSA(375)	Seq_28093	E2

		AVYFCAARLDPIFASNSEYAPLYDYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH										CCAAADPHECYAKVFDEFKPLVEEPQNLIK (SEQ ID NO: 2628)(13)-DNAKNTVYLQMNSLK (SEQ ID NO: 2640)(4)	H33(93)	HSA(396)		
												DNAKNTVYLQMNSLK (SEQ ID NO: 2640)(4)-LAKTYETTLEK (SEQ ID NO: 2594)(3)	H33(93)	HSA(375)		
H34	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGGSLRLSCAASGRTLSSYDMGWFRKAPGKEREFVAAITRHDYNTYYRDSVKGRFTISRDNAKNTVYLQMNSLKSEDTAVYFCAARLDPIFASNSAYSNLYDYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 132	0	0	0	Yes	5.065	/	/	DSS	PLVEEPQNLIKQNCELFEQLGEYK(11)-DSVKGR (SEQ ID NO: 2639)(4)	H34(82)	HSA(413)	Seq_9366	E2
												CCAAADPHECYAKVFDEFKPLVEEPQNLIK (SEQ ID NO: 2628)(13)-DNAKNTVYLQMNSLK (SEQ ID NO: 2640)(4)	H34(93)	HSA(396)		
											EDC	VFDEFKPLVEEPQNLIK (SEQ ID NO: 2598)(6)-HDYNTYYR (SEQ ID NO: 2641)(2)	H34(72)	HSA(402)		
H35	Trypsin	MASMTGGQQMGRDPNSAEVQLVESGGGLVQAGGSLRVSCAVSGISIYHSGWYRQAPGKERELVAGISRGGSTNYADSVKGRFTISRDSGENTVYLQMNSLKPEDTAVYYCKIDWDYRGVSQTAWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 133	0	/	0	Yes	No binding	/	/	/	/	/	/	/	/
H36	Chymo	MASMTGGQQMGRDPNSAEVQLVESGGGLVQAGGSLRLSCAAPAIALADYAIGWFRQGPGKEREGISCVASETDTTRYADSVKGRFTISRDNAKNLVYLQMNSLKPDDTAVYYCATEVMECRGLSYNAWGSWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHH	SEQ ID NO: 134	0	1	0	No	/	/	/	/	/	/	/	/	/

H37	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGGSLRLSCAASGLTFSNYALGWFRRAPGKERDFVAAISYSGGSTDYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAAAYLGWGTARTAYEYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 135	2	0	1	Yes	1.49	/	/	/	/	/	/	/	/
H38	Chymo	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCAASELTFSNYAMGWFRRAPGKERGFVAAISYSGGSTDYADSVKGRFTISRDNAKKTVYLQMNSLKPEDTAVYYCAAAYMGWGTARSAYEYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 136	0	/	1	Yes	2.435	/	/	/	/	/	/	/	/
H39	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGVSLRLSCAASERTFSSYIMGWFRQAPGKEREFIAAISWSGGNTDYAGSVQGRFTISRDNAQNTVYLQMNSLEPEDTAVYYCAADATHSWSYGSRWYDRNYNYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 137	/	/	1	Yes	5.196	/	/	EDC	SHCIAEVENDEMPADLPSLAADFVESK (SEQ ID NO: 2642)(15)-ASMTGGQQMGR (SEQ ID NO: 2546)(1)	H39(2)	HSA(325)	Seq_43495	E2
												RPCFSALEVDETYVPK (SEQ ID NO: 2616)(8)-ASMTGGQQMGR (SEQ ID NO: 2546)(1)	H39(2)	HSA(516)		
												SHCIAEVENDEMPADLPSLAADFVESK (SEQ ID NO: 2642)(22)-ASMTGGQQMGR (SEQ ID NO: 2546)(1)	H39(2)	HSA(332)		
												SHCIAEVENDEMPADLPSLAADFVESK (SEQ ID NO: 2642)(11)-ASMTGGQQMGR (SEQ ID NO: 2546)(1)	H39(2)	HSA(321)		
H40	Chymo	MASMTGGQQMGRDPNSAEVQLVESGGGLVQAGASLRLSCAASGGTFSSYIMGWFRQAPGKEREFVAAISWSGRSTHYADSVKGRFAISRDNDRVYLQMNSLKPEDTAVYSCAADPNYTWRDDRYYREEGYTYWGQGTQVTVSSEPKTPKGGCGGGAAALEHHHHHH	SEQ ID NO: 138	/	/	1	Yes	3.156	/	/	DSS	LAKTYETTLEK (SEQ ID NO: 2594)(3)-STHYADSVKGR (SEQ ID NO: 2643)(9)	H40(82)	HSA(375)	Seq_45710	E2

H41	Chymo	MASMTGGQQMGRDPNSAQVQLVESGGGLVQAGGSLRLSCAASGLTFSNYAMGWFRQAPGKEREFVVAISRGGNTY	SEQ ID NO:	1	/	1	No	/	/	/	/	/	/	/	/	/
H42	Chymo	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCAASGLTFSNYAMGWFRQAPGKEREFVVAISWSGANTYYSDSVKGRFTASRDNAKKTVY	SEQ ID NO: 140	0	/	1	Yes	4.922	/	/	DSS	KTVYLQMNSLKPEDTAVYYCAADYR (SEQ ID NO: 2644)(1)-HPEAKR	H42(94)	HSA(468)	Seq_44732	E2
H43	Chymo	MASMTGGQQMGRDPNSAHVQLVESGGGLVQAGGSLRLSCAASGLTFSNYALGWFRRAPGKERDFVAAISYSGGSTDYADSVKGRFTISRDNAKNTVYLQMNS	SEQ ID NO: 141	2	0	1	Yes	No binding	/	/	/	/	/	/	/	/
H44	Chymo	MASMTGGQQMGRDPNSAEVQLVESGGGLAQAGGSLRLSCAASGGTFSNSCMGWFRQAPGMEREFVAIIRSTGHTTYADSVEGRFTVSRDIAKNTVYL	SEQ ID NO: 142	2	1	1	Yes	5.622	/	/	/	/	/	/	/	/

Table 3. Summary of PDZ Nbs and their biophysical and physiochemical properties.

ID	**Enzyme**	**Protein Sequences**	**SEQ ID NO**	**Soluble**	**Bind by beads-binding assay (Fig 10b)**	**WT ELISA affinity (LogIC50 (oD450nm))**	**Mutant ELISA affinity (LogIC50 (oD450nm))**	**Affinity fold change**
P1	Trypsin/Chymo	MASMTGGQQMGRNSADVQLVESGGGLVQPGGSLRLSCAASGFTLDDYAIGWFRQAPGKEREGVSCISSHGSTYYADSVKGRFTISRDNVKNTLYLQMNSLKPEDTALYYCAASYYSDYEVAVCRSDALDAWGQGTQVTVSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 143	Yes	Yes	/	/	/
P2	Trypsin/ Chymo	MASMTGGQQMGRNSADVQLVESGGGLVQAGGSLRLSCAASGHTFSSYTMGWFHQAPGKEREFVAEISGTGGNTGYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAAVIGSPTDSSDYRSSLDYDYWGQGTQVTVSEPKTPKGGCGGGLEHHHHHH	SEQ ID NO: 144	Yes	Yes	5.437	4.354	12.106
P3	Trypsin/ Chymo	MASMTGGQQMGRNSADVQLVESGGGLVQAGGSLRLSCAASGRTFSSYTMGWFHQAPGKEREFVAEIGGTGGNTGYADSVKGRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAAVIGSPTDSSDYRSSLDYDYWGQGTQVTVSEPKTPKGGCGGGLEHHHHHHH	SEQ ID NO: 145	Yes	Yes	5.264	4.781	3.04089

P14	P13	P12	P11	P10	P9	P8	P7	P6	P5	P4
Chymo	Chymo	Chymo	Trypsin/ Chymo	Chymo	Chymo	Trypsin	Trypsin	Trypsin/ Chymo	Trypsin/ Chymo	Trypsin/ Chymo
MASMTG GQQMGR NSAQVQL VESGGGL VQAGDSL RLSCTAS GRTFSTY TMAWFR QAPGKER EFVAAIS WSGTYYA DSVKGRF TISRDNA KNTVYLQ MNSLKPE DTAVYYC AAVIGST VDTYSPS DPLEYDY WGQGTQ VTVSSEP KTPKGGC GGGLEHH HHHH	MASMTG GQQMGR NSAQVQL VESGGGL VQAGGSL RLSCAAS GRTFSTY TMGWFR QAPGKER EFVAAVT WSETLYS DSVKGRF TISRDNA KNTVYLQ MNSLKPE DTAVYYC AAVQGSP VDTIVVL TTSEEYD YWGQGT QVTVSSE PKTPKGG CGGGLEH HHHHH	MASMTG GQQMGR NSAQVQL VESGGGL VQAGGSL RLSCTAS GRTFSTY TMAWFR QAPGKER EFVAAIS WSGAYY AESVKGR FTISRDNA KNTVYLQ MNSLKPE DTAVYYC AAVIGST VDSYSPS DPLEYDY WARGPRS PSEPKTPK GGCGGGL EHHHHHH	MASMTGG QQMGRNS ADVQLVE SGGGLVQ AGGSLRLS CVASGRT FSTYTMG WFRQAPG KEREFVA HIGWSGSS TYYADSV KGRFTISR DNAKNTM YLQMNSL KPEDTAV YYCAVAI GSPVDSY RHSDPLEY DYWGQGT QVTVSEP KTPKGGC GGGLEHH HHHH	MASMTG GQQMGR NSADVQL VESGGGL VQAGGSL RLSCTAS GRTFSTY TMAWFR QAPGKER EFVAAIT WSGTYYA DSVKGRF TISRDNA KNTMYLQ MNSLKPE DTAVYIC AAVIGST VDSYSPS DPLEYDY WGQGTQ VTVSEPK TPKGGCG GGLEHHH HHH	MASMTG GQQMGR NSADVQL VESGGGL VQAGGSL RLSCVAS VASGRTF GWYDMG WFRQAPG KEREFVA AISWSGG STYYADS VKGRSTIS RDNAKNT VYLQMNS LKPEDTA VYYCAAR GGGTSVD SDYDVGE FEYDYWG QGTQVTV SEPKTPK GGCGGGL EHHHHHH	MASMTG GQQMGR NSAEVQL VESGGGL VQPGGSL RLSCKAS GFDFEYY TIGWFRQ APGKERE GVSCINR GDGATYY RDSVKGR FTISRDNA KKTMYLE MNNLKPE DTAVYYC ATADSGW GCYGHRI QKNEFDH FGQGTQV TVSEPKTP KGGCGGG LEHHHHH H	MASMTG GQQMGR NSADVQL VESGGGL VKPGESL KLSCVAS GETLSSYI MGWFRQ APGKERE FVAAVSW SGNQQDY ADSVKGR FTISRDNA KNTVYLQ MNSLKPE DTAVYYC ANGPCTG PRAIAEVL YESWGQG TQVTVSE PKTPKGG CGGGLEH HHHHH	MASMTG GQQMGR NSAHVQL VESGGGL VQAGGSL RLSCAAS GRTFSRY TMGWFH QAPGKER EFVAEISG TGGNTGY ADSVKGR FTMSRDN AKNTVYL QMNSLKP EDTGVYY CAAVIGSP TDSSDYR SSLDYDY WGQGTQ VTVSEPK TPKGGCG GGLEHHH HHH	MASMTGG QQMGRNS ADVQLVES GGGLVQA GDSLRLSC AASGITFR WYTMAWF RQAPGKER DFVATINW SGSDTNYA DSVKGRFTI SRDNAKNT VTLQMNSL QPEDTAVY YCAGVPGT SLSGETDPR DYDYWGQ GTQVTVSE PKTPKGGC GGGLEHHH HHH	MASMTGG QQMGRNS AHVQLVE SGGGLVQ AGGSLRL SCAAAGR TSSDYAM GWFRQAP GKEREFV SAINWSGI STYYADS VKGRFTIS RDNAKNT VHLQMNS LKPEDTA VYYCAAE KLESLRN WHDPLM YDYWGQ GTQVTVS EPKTPKG GCGGGLE HHHHHH
SEQ ID NO: 156	SEQ ID NO: 155	SEQ ID NO: 154	SEQ ID NO: 153	SEQ ID NO: 152	SEQ ID NO: 151	SEQ ID NO: 150	SEQ ID NO: 149	SEQ ID NO: 148	SEQ ID NO: 147	SEQ ID NO: 146
Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
/	/	Yes	/	/	/	/	/	Yes	Yes	Yes
5.151	4.454	/	4.068	5.205	4.878	/	/	5.247	4.704	4.425
3.741	0.0071	/	0	3.834	2.61	/	/	4.726	0	4.578
25.704	27982.3	/	11695	23.4963	185.353	/	/	3.31895	50582.5	1

P15	Chymo	MASMTG GQQMGR NSAQVQL VESGGGL VQAGGSL RLSCVAS GRPFSSLD MGWFRQ RPGKERD VVATINW TGDSTYY LDSVKGR FTISRDNA KNTVFLQ MNSLKPE DTAVYYC AARGGGS SVDSEYD VGEFEYD YWGQGT QVTVSSE PKTPKGG CGGGLEH HHHHH	SEQ ID NO: 157	Yes	Yes	4.971	1.657	2060.63

Table 4. GST summary: amino acid sequence filters derived from a deep learning approach

Region of activity	Filter	Activity in Low affinity prediction	Activity in High affinity prediction
Cdr3	See FIG. 15A, SEQ ID NO: 2663	< 1%	56% (41% in 5-best contributors)
Cdr3	See FIG. 15B, SEQ ID NO: 2664	76% (69 % in 5-best contributors)	< 1%

Table 5. HSA summary: amino acid sequence filters derived from a deep learning approach

Region of activity	Filter	Activity in Low affinity prediction	Activity in High affinity prediction
Cdr3	See FIG. 16A, SEQ ID NO: 2665	79% (65%% in 5-best contributors)	20% (<10% in 5-best contributors)
Cdr3	See FIG. 16B; SEQ ID NO: 2666	< 1%	75%(50% in 5-best contributors) Most contributing
Cdr3	See FIG. 16C; SEQ ID NO: 2667	<1%	77% (27% in 5-best contributors) Most activated

References

1. Muyldermans, S. Nanobodies: natural single-domain antibodies. *Annu Rev Biochem* 82, 775-797 (2013).
2. Beghein, E. & Gettemans, J. Nanobody Technology: A Versatile Toolkit for Microscopic Imaging, Protein-Protein Interaction Analysis, and Protein Function Exploration. *Front Immunol* 8, 771 (2017).
3. Rasmussen, S.G. et al. Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. *Nature* 469, 175-180 (2011).
4. Jovcevska, I. & Muyldermans, S. The Therapeutic Potential of Nanobodies. *BioDrugs* 34, 11-26 (2020).
5. Lauwereys, M. et al. Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *The EMBO journal* 17, 3512-3520 (1998).
6. Pardon, E. et al. A general protocol for the generation of Nanobodies for structural biology. *Nature protocols* 9, 674-693 (2014).
7. McMahon, C. et al. Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nature structural & molecular biology* 25, 289-296 (2018).

8. Egloff, P. et al. Engineered peptide barcodes for in-depth analyses of binding protein libraries. *Nature methods* 16, 421-428 (2019).

9. Fridy, P.C. et al. A robust pipeline for rapid production of versatile nanobody repertoires. *Nature methods* 11, 1253-1260 (2014).

10. Savitski, M.M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Molecular & cellular proteomics : MCP* 14, 2394-2404 (2015).

11. DeKosky, B.J. et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature biotechnology* 31, 166-169 (2013).

12. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* 4, 207-214 (2007).

13. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H.J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research* 33, W363-W367 (2005).

14. Chait, B.T., Cadene, M., Olinares, P.D., Rout, M.P. & Shi, Y. Revealing Higher Order Protein Structure Using Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* 27, 952-965 (2016).

15. Rout, M.P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* 177, 1384-1403 (2019).

16. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Analytical Chemistry* 90, 144-165 (2018).

17. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in biochemical sciences* 41, 20-32 (2016).

18. Larsen, M.T., Kuhlmann, M., Hvam, M.L. & Howard, K.A. Albumin-based drug delivery: harnessing nature to cure disease. *Mol Cell Ther* 4, 3 (2016).

19. Zhu, W.H., Smith, J.W. & Huang, C.M. Mass Spectrometry-Based Label-Free Quantitative Proteomics. *J Biomed Biotechnol* (2010).

20. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* 26, 1367-1372 (2008).

21. Shi, Y. et al. Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Molecular & cellular proteomics : MCP* 13, 2927-2943 (2014).

22. Kim, S.J. et al. Integrative structure and functional anatomy of a nuclear pore complex. *Nature* 555, 475-482 (2018).

23. Pires, D.E.V., Ascher, D.B. & Blundell, T.L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics (Oxford, England)* 30, 335-342 (2014).

24. Finn, J.A. et al. Improving Loop Modeling of the Antibody Complementarity-Determining Region 3 Using Knowledge-Based Restraints. *PloS one* 11, e0154811 (2016).

25. Tiller, K.E. et al. Arginine mutations in antibody complementarity-determining regions display context-dependent affinity/specificity trade-offs. *The Journal of biological chemistry* 292, 16638-16652 (2017).

26. Mitchell, L.S. & Colwell, L.J. Analysis of nanobody paratopes reveals greater diversity than classical antibodies. *Protein Eng Des Sel* 31, 267-275 (2018).

27. Desmyter, A. et al. Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. *Nat Struct Biol* 3, 803-811 (1996).

28. Li, T. et al. Immuno-targeting the multifunctional CD38 using nanobody. *Scientific reports* 6 (2016).

29. Sheng, M. & Sala, C. PDZ domains and the organization of supramolecular complexes. *Annu Rev Neurosci* 24, 1-29 (2001).

30. Doyle, D.A. et al. Crystal structures of a complexed and peptide-free membrane protein-binding domain: Molecular basis of peptide recognition by PDZ. *Cell* 85, 1067-1076 (1996).

31. Niethammer, M. et al. CRIPT, a novel postsynaptic protein that binds to the third PDZ domain of PSD-95/SAP90. *Neuron* 20, 693-707 (1998).

32. Akram, A. & Inman, R.D. Immunodominance: A pivotal principle in host response to viral infections. *Clin Immunol* 143, 99-115 (2012).

33. Bar-On, Y.M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America* 115, 6506-6511 (2018).

34. Chaplin, D.D. Overview of the immune response. *J Allergy Clin Immun* 125, S3-S23 (2010).

35. Acharya, P. et al. Heavy chain-only IgG2b llama antibody effects near-pan HIV-1 neutralization by recognizing a CD4-induced epitope that includes elements of coreceptor- and CD4-binding sites. *J Virol* 87, 10173-10181 (2013).

36. Arabi, Y.M. et al. Middle East Respiratory Syndrome. *New Engl J Med* 376, 584-594 (2017).

37. Flajnik, M.F., Deschacht, N. & Muyldermans, S. A Case Of Convergence: Why Did a Simple Alternative to Canonical Antibodies Arise in Sharks and Camels? *PLoS biology* 9 (2011).

38. Sircar, A., Sanni, K.A., Shi, J. & Gray, J.J. Analysis and modeling of the variable region of camelid single-domain antibodies. *J Immunol* 186, 6357-6367 (2011).

39. Baran, D. et al. Principles for computational design of binding antibodies. *Proceedings of the National Academy of Sciences of the United States of America* 114, 10900-10905 (2017).

40. Chevalier, A. et al. Massively parallel de novo protein design for targeted therapeutics. *Nature* 550, 74-79 (2017).

41. Arbabi Ghahroudi, M., Desmyter, A., Wyns, L., Hamers, R. & Muyldermans, S. Selection and identification of single domain antibody fragments from camel heavy-chain antibodies. *FEBS letters* 414, 521-526 (1997).

42. Shi, Y. et al. A strategy for dissecting the architectures of native macromolecular assemblies. *Nature methods* 12, 1135-1138 (2015).

43. Chen, Z.L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nature communications* 10, 3404 (2019).

44. Dunbar, J. & Deane, C.M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics (Oxford, England)* 32, 298-300 (2016).

45. Lefranc, M.P. et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 27, 55-77 (2003).

46. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome research* 14, 1188-1190 (2004).

47. Sievers, F. & Higgins, D.G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods in molecular biology* 1079, 105-116 (2014).

48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics (Oxford, England)* 23, 127-128 (2007).

49. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)* 25, 1189-1191 (2009).

50. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods* 4, 923-925 (2007).

51. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* 47, 5 6 1-32 (2014).

52. Dong, G.Q., Fan, H., Schneidman-Duhovny, D., Webb, B. & Sali, A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics (Oxford, England)* 29, 3158-3166 (2013).

53. Schneidman-Duhovny, D. & Wolfson, H.J. Modeling of Multimolecular Complexes. *Methods in molecular biology* 2112, 163-174 (2020).

54. Russel, D. et al. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS biology* 10, e1001244 (2012).

55. Fernandez-Martinez, J. et al. Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* 167, 1215-1228 e1225 (2016).

CLAIMS

What is claimed is:

1. A method of identifying a group of complementarity determining region (CDR)3, 2 and/or 1 nanobody amino acid sequences (CDR3, CDR2 and/or CDR1 sequences) wherein a reduced number of the CDR3, CDR2 and/or CDR1 sequences are false positives as compared to a control, the method comprising:

   a. obtaining a blood sample from a camelid immunized with an antigen;
   b. using the blood sample to obtain a nanobody cDNA library;
   c. identifying the sequence of each cDNA in the library;
   d. isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen;
   e. digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products;
   f. performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data;
   g. selecting sequences identified in step c. that correlate with the mass spectrometry data;
   h. identifying sequences of CDR3, CDR2 and/or CDR1 regions in the sequences from step g.; and
   i. selecting from the CDR3, CDR2 and/or CDR1 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the fragmentation coverage percentage is determined by a formula $f(x,\text{chymotrypsin}) = 0.0023x^2-0.0497x+0.7723,x[5,30]$ when chymotrypsin is used in step e. or a formula $f(x,\text{trypsin})=0.00006x^2 - 0.00444x+0.9194, x[5,30]$ when trypsin is used in step e., and wherein x is the length of the CDR3, CDR2 or CDR1 region sequence, respectively; and
   j. wherein the selected sequences of step i. comprise a group having the reduced number of false positive CDR3, CDR2 and/or CDR1 sequences.

2. The method of claim 1, wherein the required fragmentation coverage percentage is about 30.

3. The method of claim 1, wherein the required fragmentation coverage percentage is about 50 and trypsin is used in step e.

4. The method of claim 1, wherein the required fragmentation coverage percentage is about 40 and chymotrypsin is used in step e.

5. The method of any one of claims 1-4, wherein step d. comprises obtaining plasma from the blood sample and isolating nanobodies using one or more affinity isolation methods.

6. The method of claim 5, wherein the one or more affinity isolation methods of step d. comprise one or more of protein G sepharose affinity chromatography and protein A sepharose affinity chromatography.

7. The method of any one of claims 1-6, wherein step d. further comprises a functional selection step comprising selecting antigen-specific nanobodies using an antigen-specific affinity chromatography and eluting the antigen-specific nanobodies under varying degrees of stringency thereby creating different nanobody fractions, and performing steps e. through i. on each fraction individually and estimating an affinity of each different step i. CDR3, CDR2 and/or CDR1 region sequence for the antigen based on a relative abundance of the CDR3, CDR2 and/or CDR1 region sequence in each of the nanobody fractions, respectively.

8. The method of claim 7, wherein the antigen-specific affinity chromatography is a resin conjugated to the antigen.

9. The method of claim 7, wherein the antigen-specific affinity chromatography is a resin coupled to maltose binding protein and the antigen.

10. The method of any one of claims 1-9, further comprising creating a CDR3, CDR2 and/or CDR1 peptide having a sequence identified in step i.

11. The method of any one of claims 1-9, further comprising creating a nanobody comprising a CDR3, CDR2 and/or CDR1 region having a sequence identified in step i.

12. A nanobody comprising an amino acid sequence selected from SEQ ID NOs: 1-2536 and SEQ ID NOs: 2665-2667.

13. A computer-implemented method, comprising:

receiving a nanobody peptide sequence;

identifying a plurality of complementarity-determining region (CDR) regions of the nanobody peptide sequence, the CDR regions including CDR3, CDR2 and/or CDR1 regions;

applying a fragmentation filter to discard one or more false positive CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence;

quantifying an abundance of one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence; and

inferring an antigen affinity based on the quantified abundance of the one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence.

14. The computer-implemented method of claim 13, further comprising classifying the one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence as having a low antigen affinity, mediocre antigen affinity, or high antigen affinity.

15. The method of claim 14, further comprising assembling the one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence classified as having the high antigen affinity into a nanobody protein.

16. The computer-implemented method of any one of claims 13-15, wherein the fragmentation filter is configured to require a minimum calculated fragmentation coverage percentage.

17. The computer-implemented method of claim 16, wherein the minimum calculated fragmentation coverage percentage is about 30.

18. The computer-implemented method of claim 17, wherein the minimum calculated fragmentation coverage percentage is about 50 for trypsin-treated samples and about 40 for chymotrypsin-treated samples.

19. The computer-implemented method of any one of claims 13-18, further comprising:

receiving a plurality of nanobody peptide sequences; and comparing each of the nanobody peptide sequences to a database to separate the nanobody peptide sequences into an excluded subgroup and a non-excluded subgroup, wherein the nanobody peptide sequences of the excluded subgroup are not found in the database, and wherein the CDR regions are only identified in the nanobody peptide sequences of the non-excluded subgroup.

20. The computer-implemented method of any one of claims 13-19, wherein the abundance of one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence is quantified based on relative MS1 ion signal intensities.

21. The computer-implemented method of any one of claims 13-20, wherein the antigen affinity is inferred using k-means clustering based on epitope similarity.

22. A method for training a deep learning model, comprising:

creating a dataset using the computer-implemented method of any one of claims 13-21; and

training, using the dataset, a deep learning model to classify nanobody peptide sequences having low antigen affinity and nanobody peptide sequences having high antigen affinity, wherein the dataset comprises a plurality of nanobody peptide sequences and corresponding antigen-affinity labels.

23. The method of claim 22, wherein the deep learning model is a convolutional neural network.

24. A method for determining antigen affinity of nanobody peptide sequences, comprising:

receiving a nanobody peptide sequence;

inputting the nanobody peptide sequence into a trained deep learning model; and

classifying, using the trained deep learning model, the nanobody peptide sequence as having low antigen affinity or high antigen affinity.

25. The method of claim 24, wherein the deep learning model is a convolutional neural network.

26. The method of claim 24 or 25, wherein the trained deep learning model is trained according to claim 22.

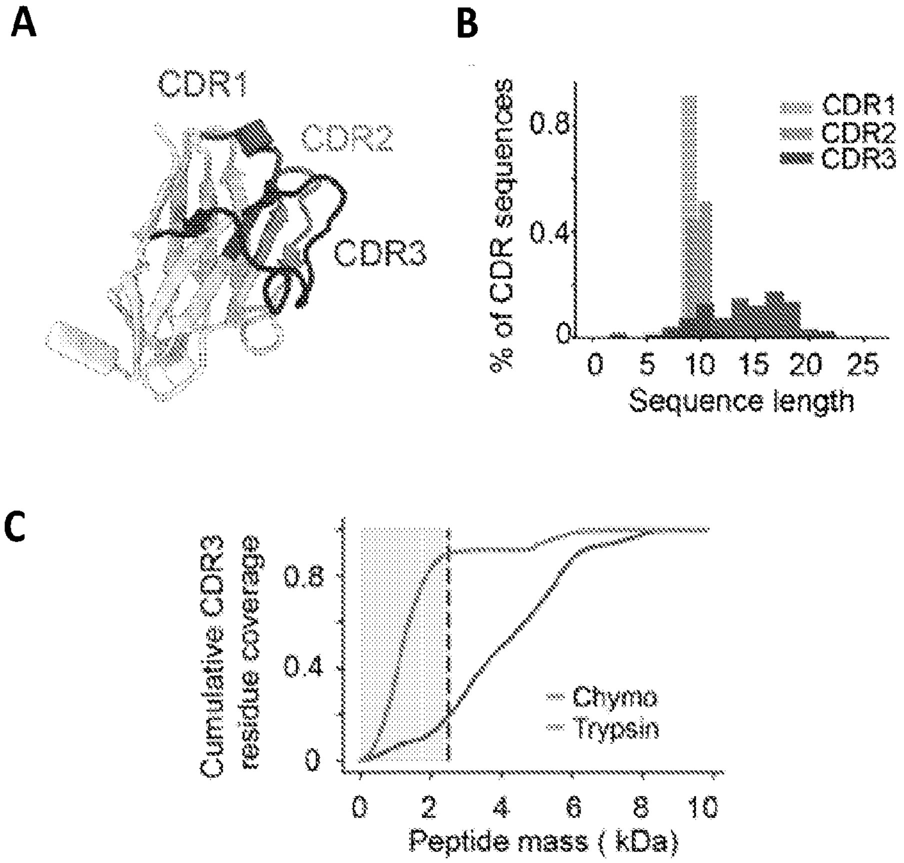

A
CDR1
CDR2
CDR3
B
% of CDR sequences
0.8
0.4
0
CDR1
CDR2
CDR3
0 5 10 15 20 25
Sequence length
C
Cumulative CDR3 residue coverage
0.8
0.4
0
Chymo
Trypsin
0 2 4 6 8 10
Peptide mass (kDa)

FIG. 1(A-C)

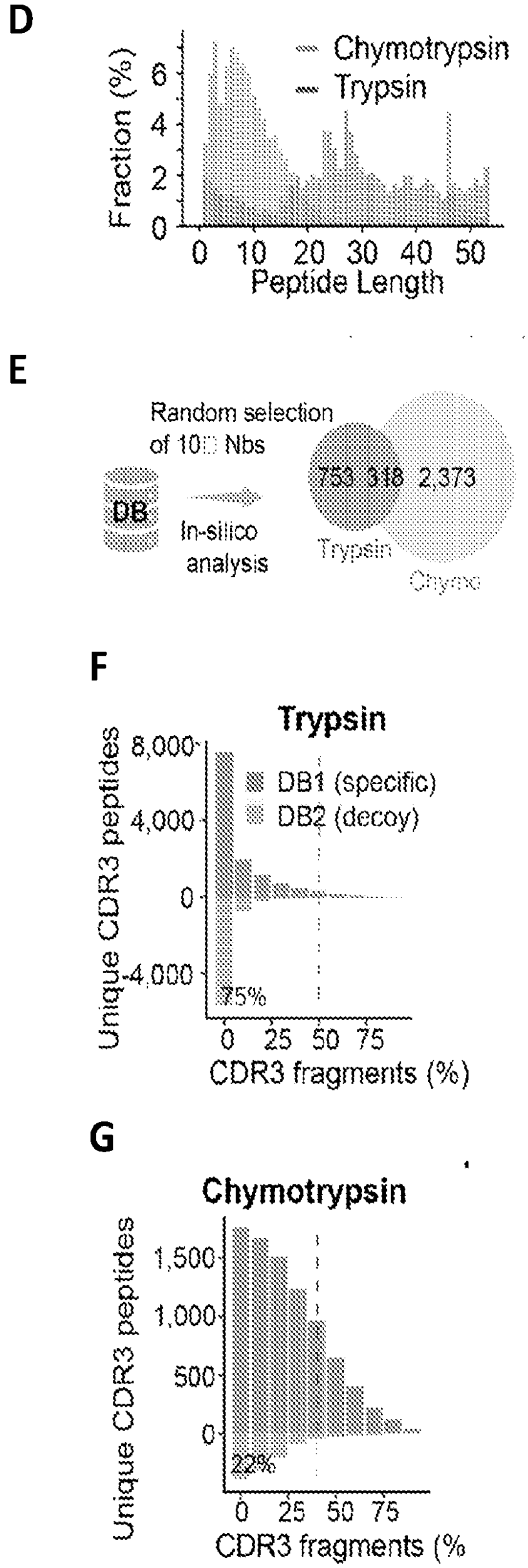

D
Fraction (%)
0
2
4
6
Chymotrypsin
Trypsin
0 10 20 30 40 50
Peptide Length
E
Random selection
of 10□ Nbs
DB
In-silico
analysis
753 318 2,373
Trypsin
Chymo
F
Trypsin
Unique CDR3 peptides
8,000
4,000
0
-4,000
DB1 (specific)
DB2 (decoy)
75%
0 25 50 75
CDR3 fragments (%)
G
Chymotrypsin
Unique CDR3 peptides
1,500
1,000
500
0
22%
0 25 50 75
CDR3 fragments (%

FIG. 1(D-G)

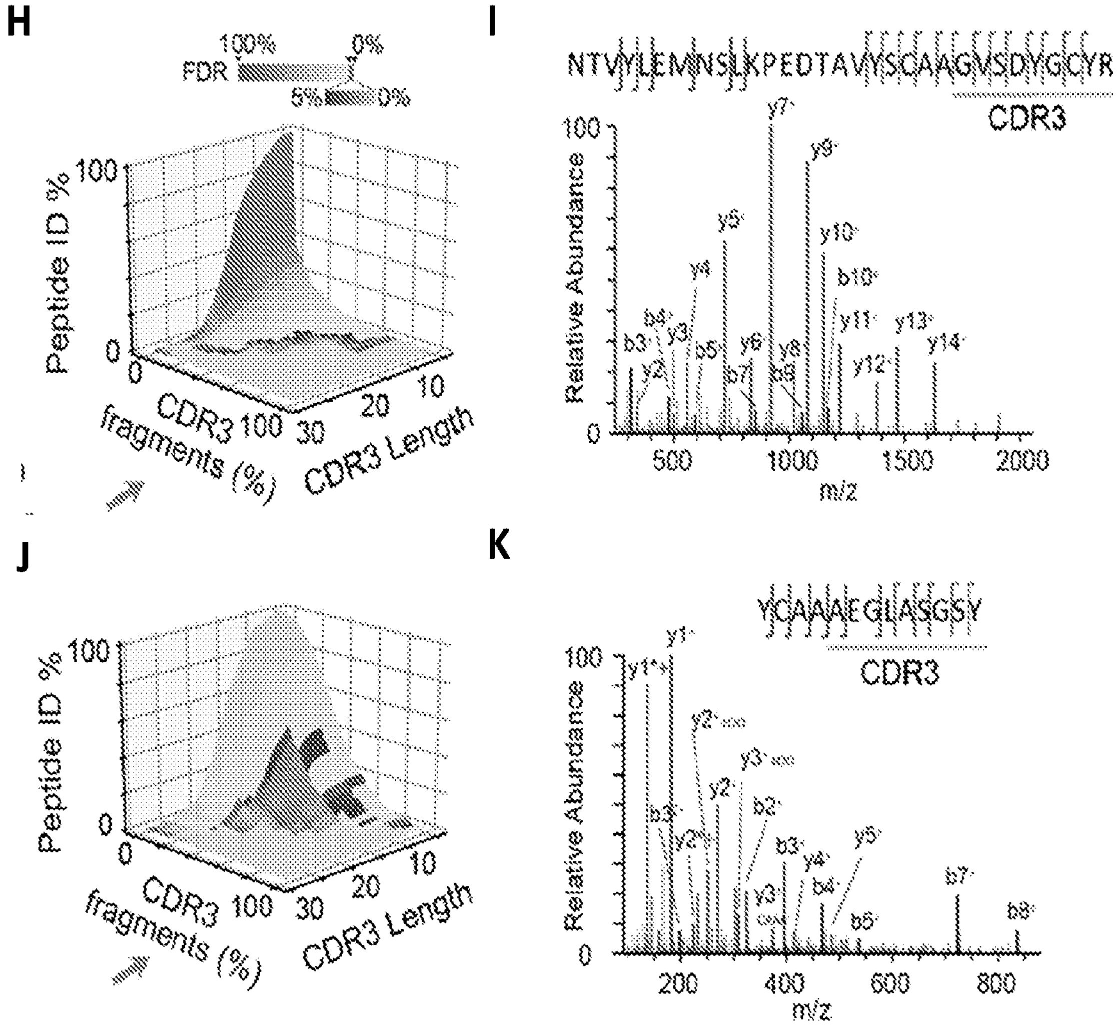

H
100%
0%
FDR
5%
0%
Peptide ID %
100
0
0
CDR3 fragments (%)
100
30
20
10
CDR3 Length
I
NTVYLEMNSLKPEDTAVYSCAAGVSDYGCYR
CDR3
Relative Abundance
100
0
y7
y9
y5
y10
y4
b10
b4
y3
y11
y13
b3
b5
y6
y8
y14
y2
b7
b9
y12
500
1000
1500
2000
m/z
J
Peptide ID %
100
0
0
CDR3 fragments (%)
100
30
20
10
CDR3 Length
K
YCAAAEGLASGSY
CDR3
Relative Abundance
100
0
y1
y1
y2
y3
y2
b2
b3
y2
b3
y4
y5
b7
y3
b4
b5
b8
200
400
600
800
m/z

FIG. 1(H-K)

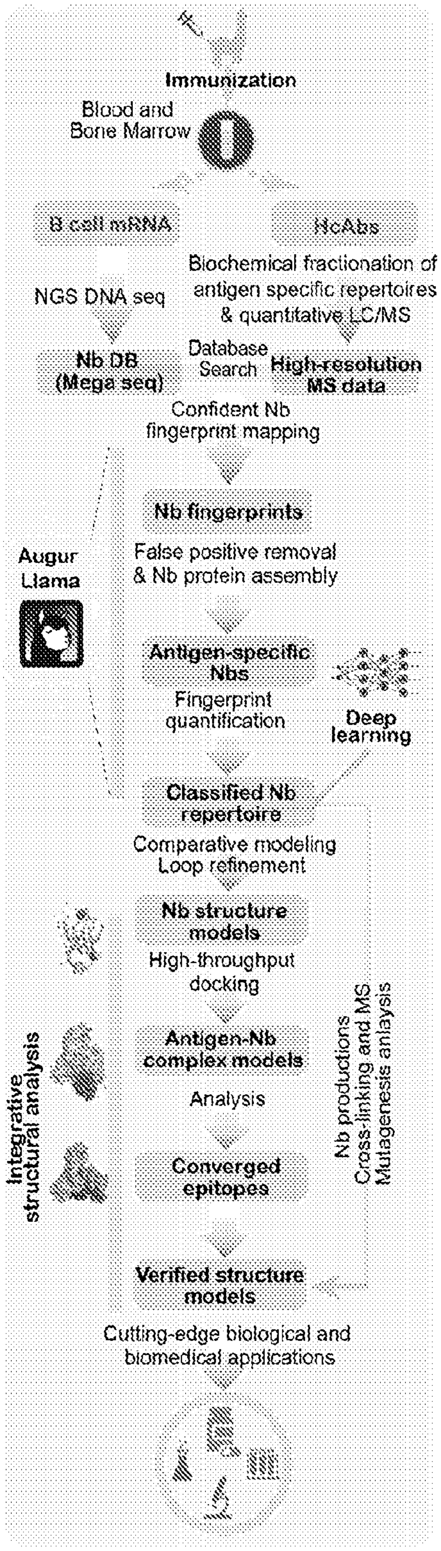

Immunization
Blood and
Bone Marrow
B cell mRNA
HcAbs
NGS DNA seq
Biochemical fractionation of
antigen specific repertoires
& quantitative LC/MS
Nb DB
(Mega seq)
Database
Search
High-resolution
MS data
Confident Nb
fingerprint mapping
Nb fingerprints
Augur
Llama
False positive removal
& Nb protein assembly
Antigen-specific
Nbs
Fingerprint
quantification
Deep
learning
Classified Nb
repertoire
Comparative modeling
Loop refinement
Nb structure
models
High-throughput
docking
Antigen-Nb
complex models
Analysis
Converged
epitopes
Integrative
structural analysis
Nb productions
Cross-linking and MS
Mutagenesis anlaysis
Verified structure
models
Cutting-edge biological and
biomedical applications

FIG. 2A

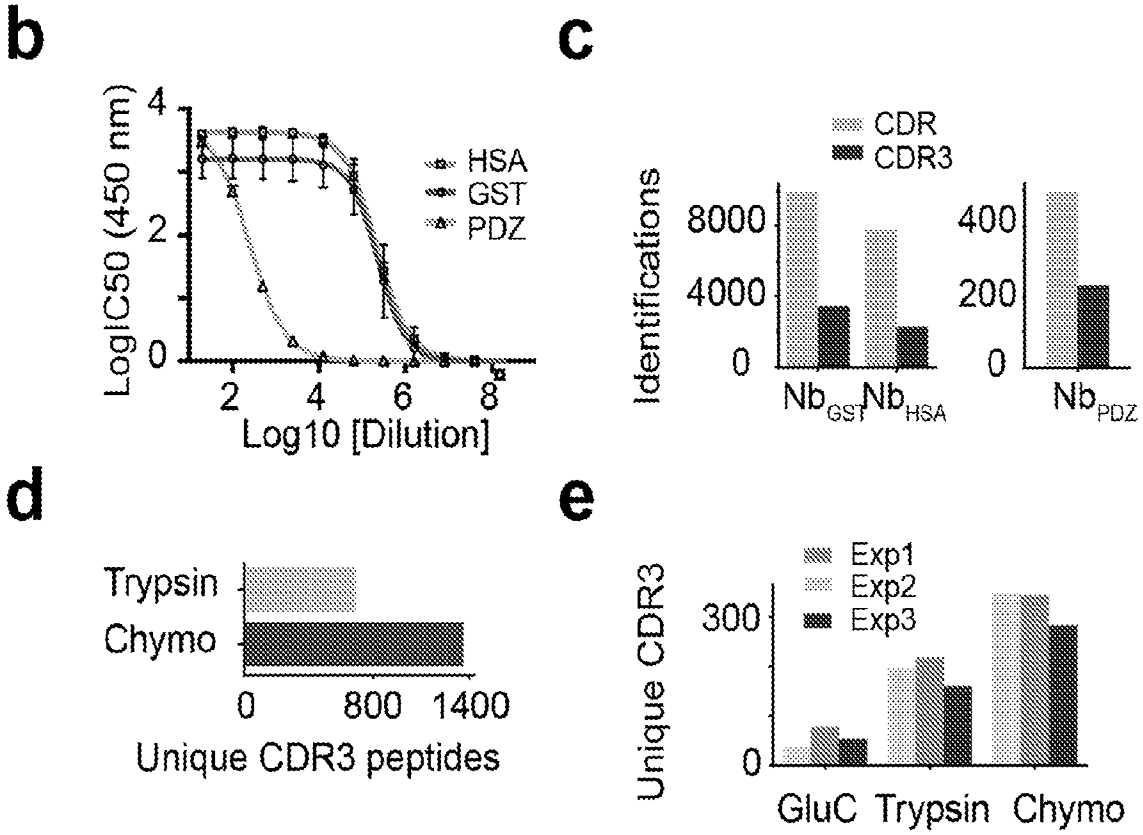
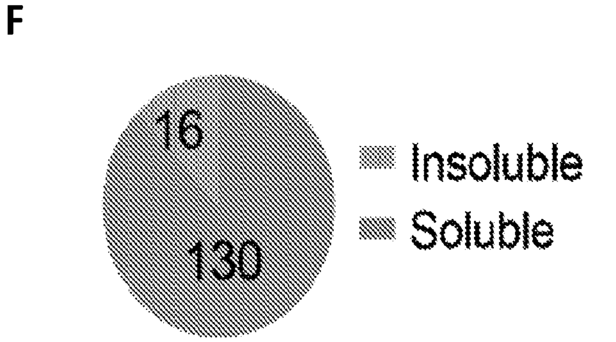

b
LogIC50 (450 nm)
4
2
1
0
2
4
6
8
Log10 [Dilution]
HSA
GST
PDZ
c
CDR
CDR3
Identifications
8000
4000
0
400
200
0
Nb$_{GST}$
Nb$_{HSA}$
Nb$_{PDZ}$
d
Trypsin
Chymo
0
800
1400
Unique CDR3 peptides
e
Exp1
Exp2
Exp3
Unique CDR3
300
0
GluC
Trypsin
Chymo

FIG. 2(b-e)


F
16
130
Insoluble
Soluble

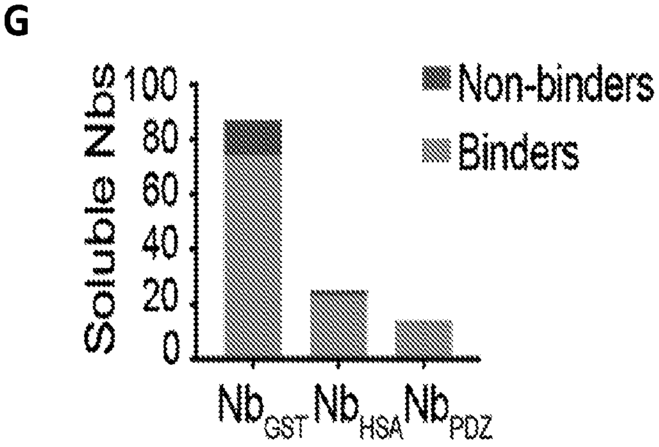

G
Soluble Nbs
100
80
60
40
20
0
Non-binders
Binders
Nb_{GST}
Nb_{HSA}
Nb_{PDZ}

FIG. 2(F-G)

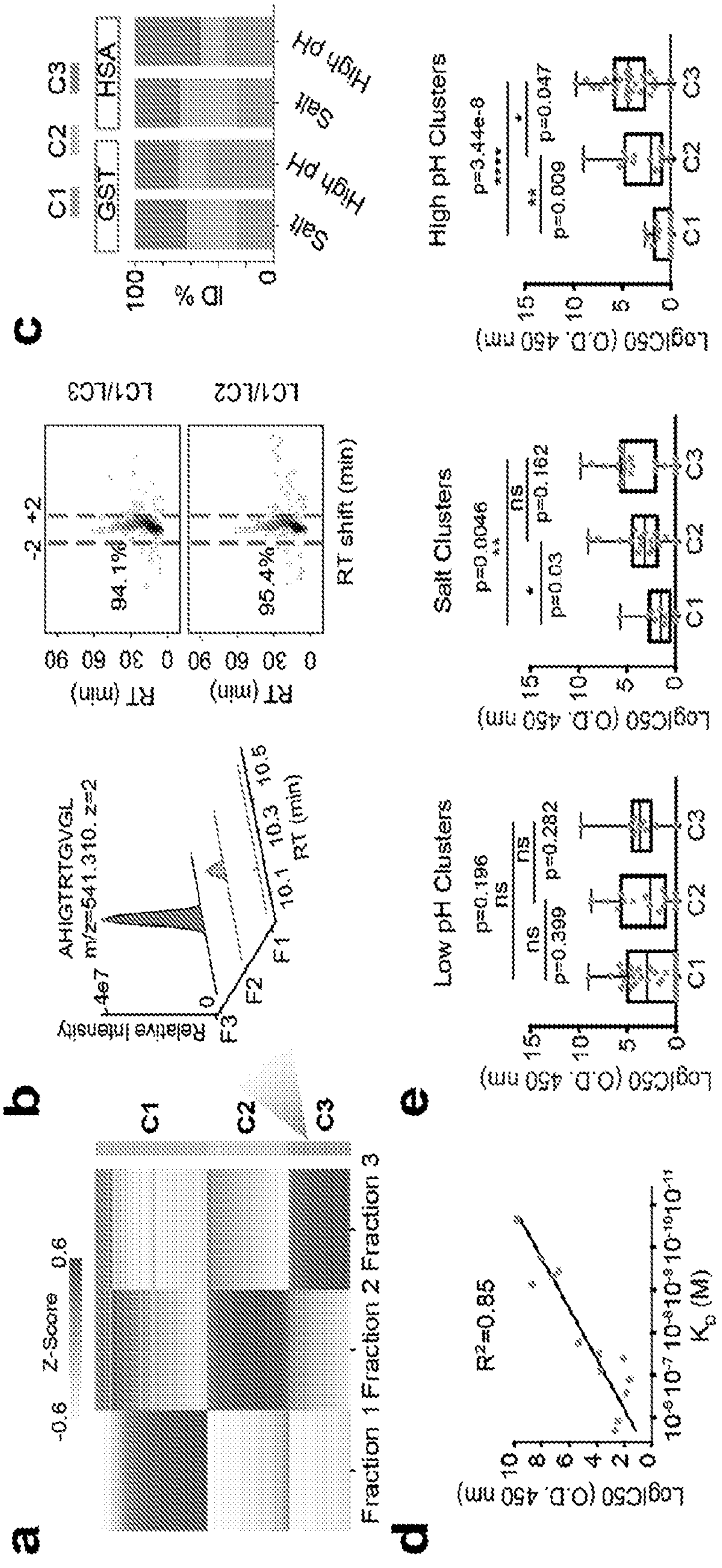

a
-0.6
Z-Score
0.6
Fraction 1
Fraction 2
Fraction 3
C1
C2
C3
b
AHIGTRTGVGL
m/z=541.310, z=2
Relative intensity
4e7
0
F3
F2
F1
10.1
10.3
10.5
RT (min)
-2 +2
94.1%
95.4%
RT (min)
0 30 60 90
RT shift (min)
LC1/LC3
LC1/LC2
c
C1
C2
C3
GST
HSA
100
ID %
0
Salt
High pH
Salt
High pH
d
R^2=0.85
LogIC50 (O.D. 450 nm)
10
8
6
4
2
0
10^{-6} 10^{-7} 10^{-8} 10^{-9} 10^{-10} 10^{-11}
K$_D$ (M)
e
Low pH Clusters
p=0.196
ns
ns
p=0.399
ns
p=0.282
LogIC50 (O.D. 450 nm)
15
10
5
0
C1
C2
C3
Salt Clusters
p=0.0046
**
*
p=0.03
ns
p=0.162
High pH Clusters
p=3.44e-8
****
**
p=0.009
*
p=0.047

FIG. 3(a-e)

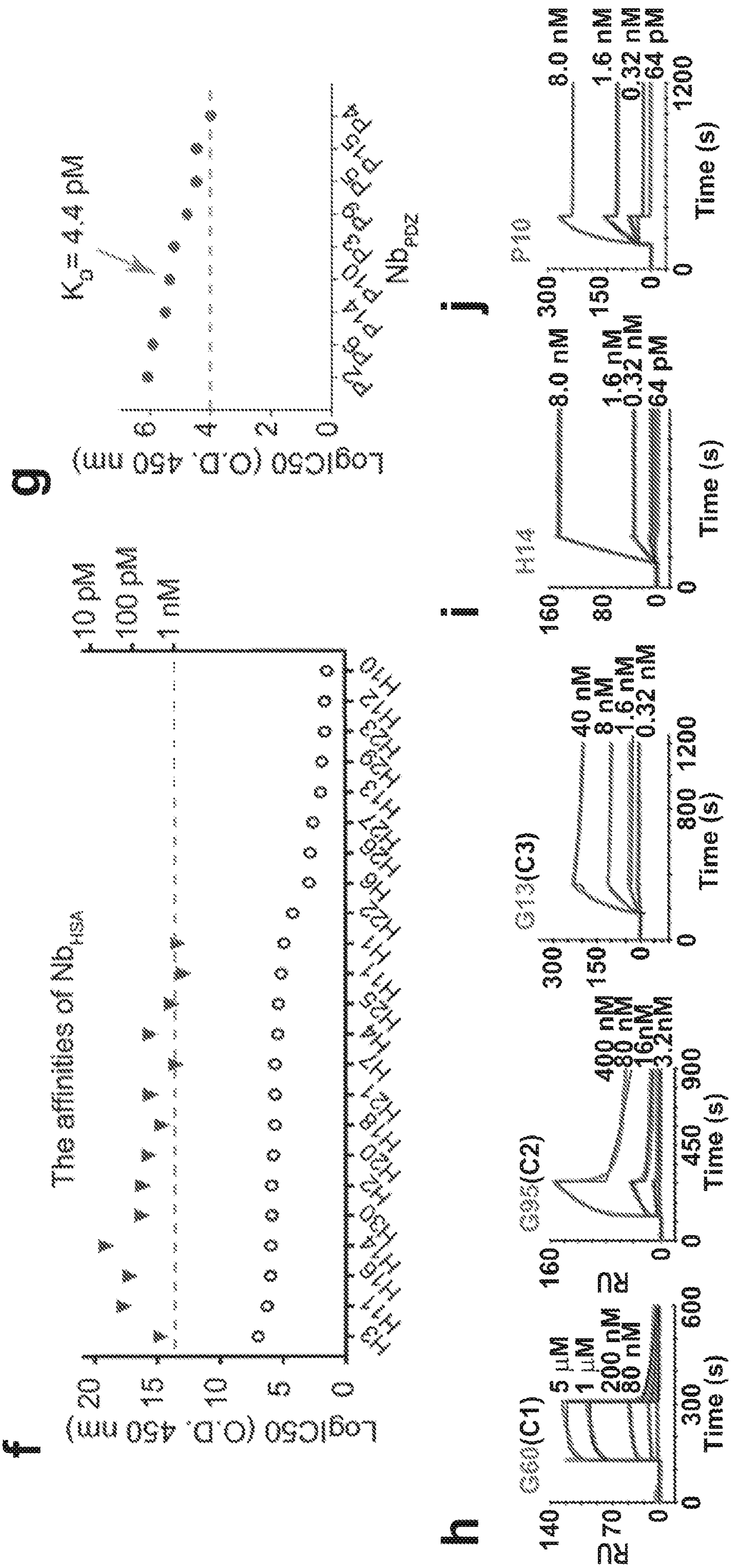

f
The affinities of Nb$_{HSA}$
LogIC50 (O.D. 450 nm)
20
15
10
5
0
10 pM
100 pM
1 nM
g
K$_D$ = 4.4 pM
LogIC50 (O.D. 450 nm)
6
4
2
0
Nb$_{PDZ}$
h
RU
140
70
0
G60(C1)
5 µM
1 µM
200 nM
80 nM
0
300
600
Time (s)
RU
160
0
G95(C2)
400 nM
80 nM
16nM
3.2nM
0
450
900
Time (s)
300
150
0
G13(C3)
40 nM
8 nM
1.6 nM
0.32 nM
0
800
1200
Time (s)
i
160
80
0
H14
8.0 nM
1.6 nM
0.32 nM
64 pM
0
Time (s)
j
300
150
0
P10
8.0 nM
1.6 nM
0.32 nM
64 pM
0
1200
Time (s)

FIG. 3(f-j)

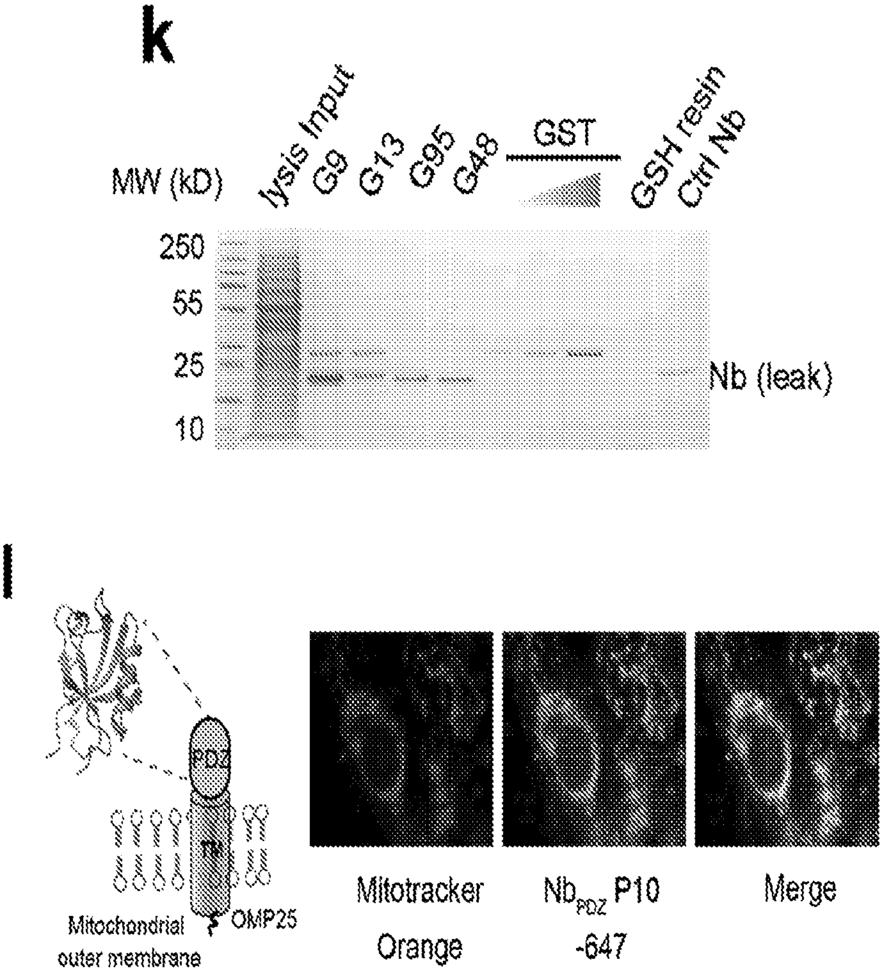

k
MW (kD)
lysis input
G9
G13
G95
G48
GST
GSH resin
Ctrl Nb
250
55
25
10
Nb (leak)
l
PDZ
TM
Mitochondrial
outer membrane
OMP25
Mitotracker
Orange
Nb$_{PDZ}$ P10
-647
Merge

FIG. 3(k-l)

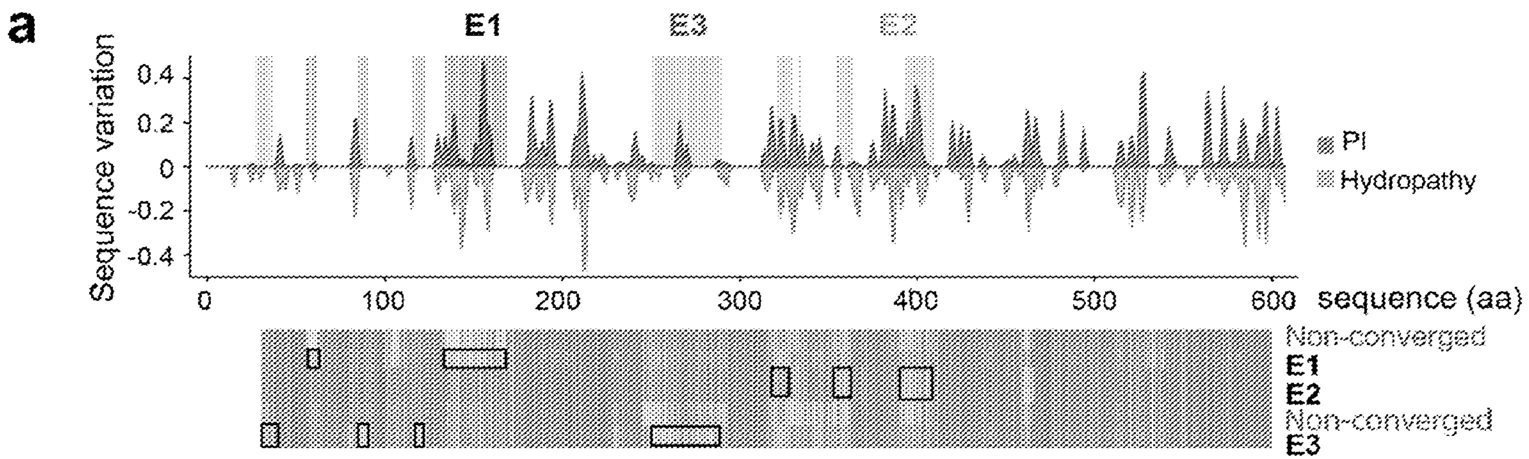

a
E1
E3
E2
Sequence variation
0.4
0.2
0
-0.2
-0.4
0
100
200
300
400
500
600
sequence (aa)
▨ PI
▨ Hydropathy
Non-converged
E1
E2
Non-converged
E3

FIG. 4a

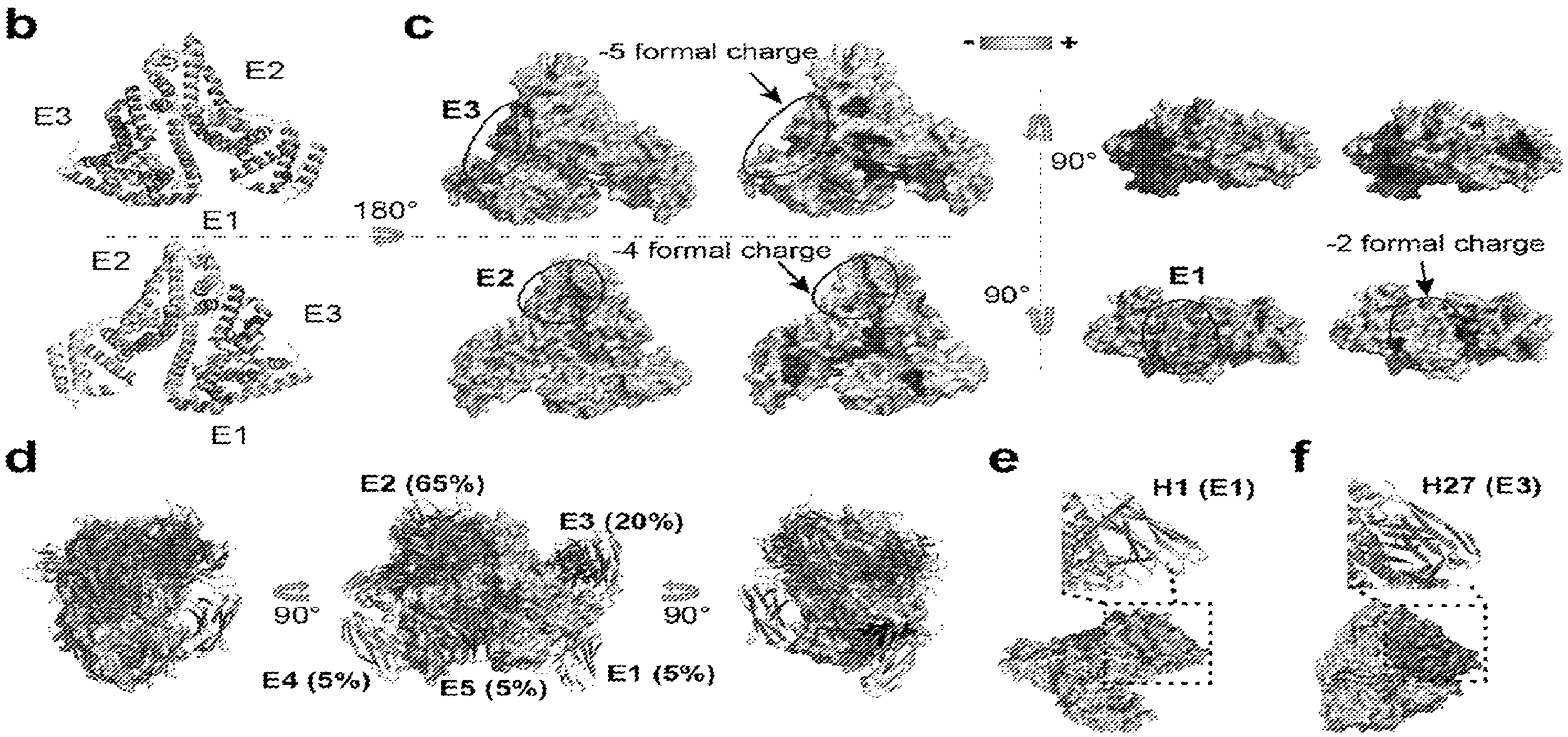

b
E3
E2
E1
E2
E3
E1
180°
c
E3
-5 formal charge
E2
-4 formal charge
- +
90°
90°
E1
-2 formal charge
d
E2 (65%)
E3 (20%)
90°
90°
E4 (5%)
E5 (5%)
E1 (5%)
e
H1 (E1)
f
H27 (E3)

FIG. 4(b-f)

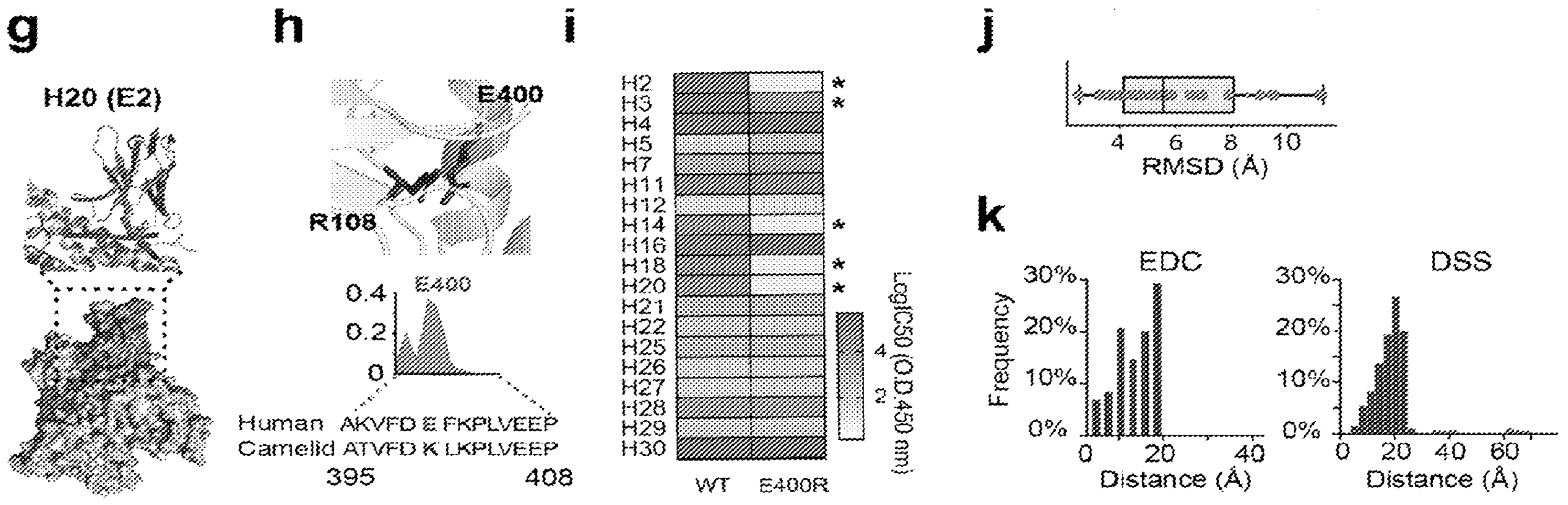

g
H20 (E2)
h
E400
R108
0.4
0.2
0
E400
Human AKVFD E FKPLVEEP
Camelid ATVFD K LKPLVEEP
395
408
i
H2
H3
H4
H5
H7
H11
H12
H14
H16
H18
H20
H21
H22
H25
H26
H27
H28
H29
H30
WT
E400R
*
*
*
*
*
LogIC50 (O.D.450 nm)
4
2
j
4
6
8
10
RMSD (Å)
k
EDC
DSS
Frequency
30%
20%
10%
0%
0
20
40
60
Distance (Å)

FIG. 4(g-k)

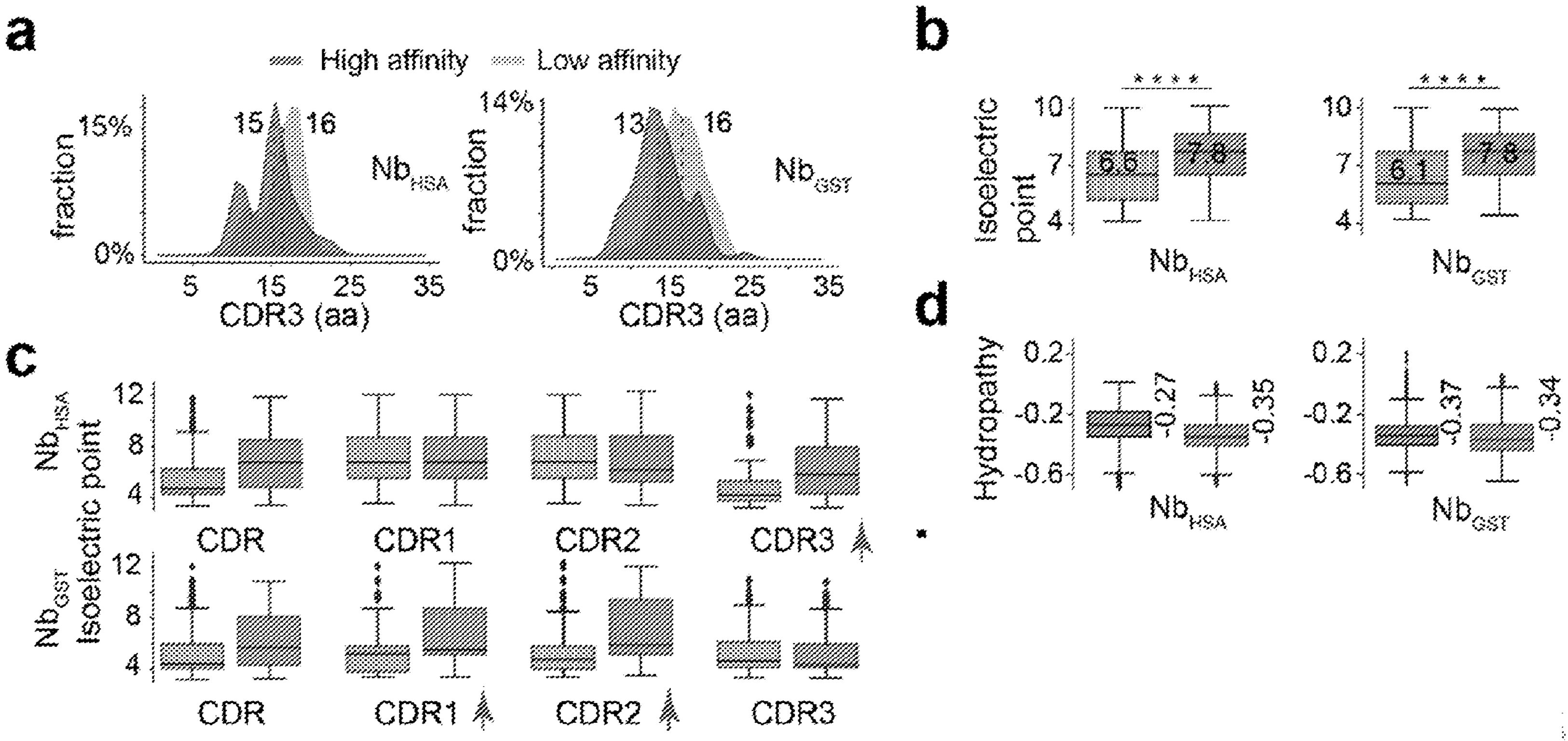

a
High affinity
Low affinity
fraction
15%
0%
15
16
Nb$_{HSA}$
5
15
25
35
CDR3 (aa)
fraction
14%
0%
13
16
Nb$_{GST}$
5
15
25
35
CDR3 (aa)
b
Isoelectric point
10
7
4
****
6.6
7.8
Nb$_{HSA}$
10
7
4
****
6.1
7.8
Nb$_{GST}$
c
Nb$_{HSA}$
Isoelectric point
12
8
4
CDR
CDR1
CDR2
CDR3
Nb$_{GST}$
Isoelectric point
12
8
4
CDR
CDR1
CDR2
CDR3
d
Hydropathy
0.2
-0.2
-0.6
-0.27
-0.35
Nb$_{HSA}$
0.2
-0.2
-0.6
-0.37
-0.34
Nb$_{GST}$

FIG. 5(a-d)

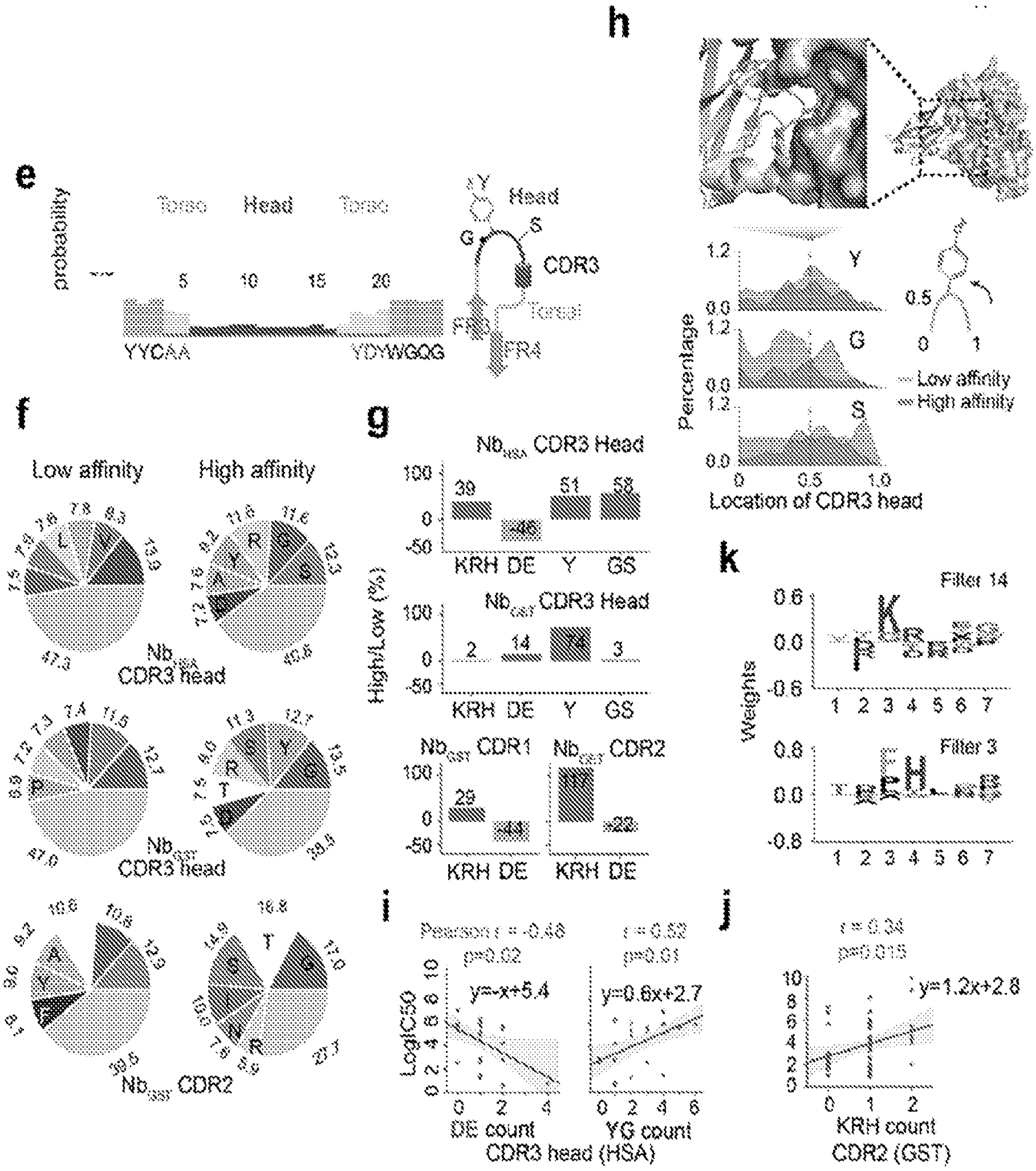

e
probability
Torso
Head
Torso
5
10
15
20
YYCAA
YDYWGQG
Y
Head
G
S
CDR3
Torsal
FR3
FR4
h
1.2
0.0
Y
G
S
Percentage
0.5
0
1
Low affinity
High affinity
Location of CDR3 head
f
Low affinity
High affinity
Nb_HSA CDR3 head
Nb_GST CDR3 head
Nb_GST CDR2
g
Nb_HSA CDR3 Head
Nb_GST CDR3 Head
Nb_GST CDR1
Nb_GST CDR2
High/Low (%)
KRH
DE
Y
GS
k
Filter 14
Filter 3
Weights
i
Pearson r = -0.48
p=0.02
r = 0.52
p=0.01
y=-x+5.4
y=0.6x+2.7
LogIC50
DE count
YG count
CDR3 head (HSA)
j
r = 0.34
p=0.015
y=1.2x+2.8
KRH count
CDR2 (GST)

FIG. 5(e-k)

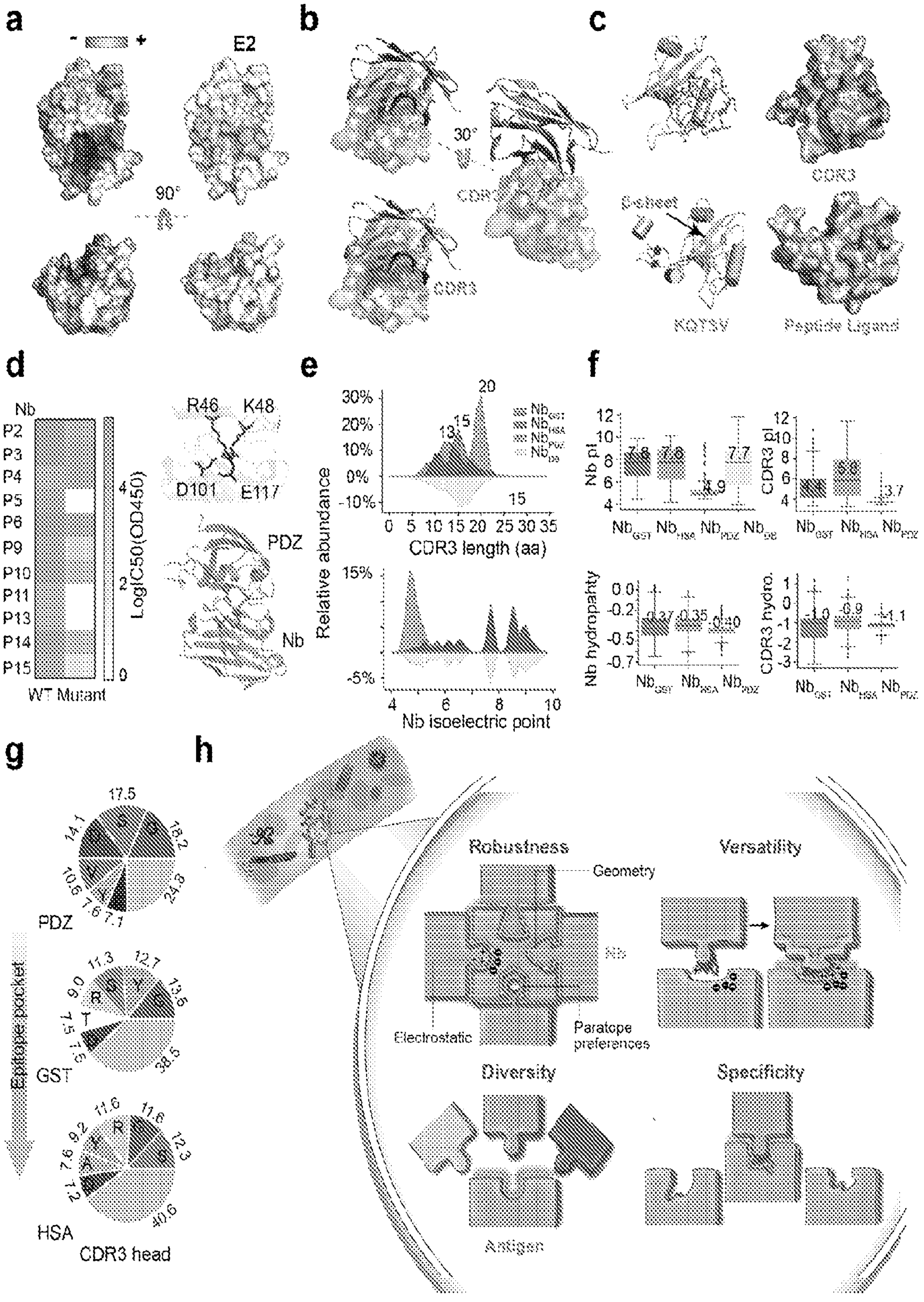

a
-
+
E2
90°
b
30°
CDR3
CDR3
c
CDR3
β-sheet
KQTSV
Peptide Ligand
d
Nb
*P2
P3
P4
*P5
P6
*P9
*P10
*P11
*P13
*P14
*P15
WT Mutant
LogIC50(OD450)
4
2
0
R46
K48
D101
E117
PDZ
Nb
e
30%
20%
10%
0%
-10%
20
15
13
15
Nb GST
Nb HSA
Nb PDZ
Nb DB
0 5 10 15 20 25 30 35
CDR3 length (aa)
Relative abundance
15%
-5%
4 6 8 10
Nb isoelectric point
f
12
10
8
6
4
Nb pI
7.8
7.8
4.9
7.7
Nb GST Nb HSA Nb PDZ Nb DB
CDR3 pI
5.8
3.7
Nb GST Nb HSA Nb PDZ
0.0
-0.2
-0.5
-0.7
Nb hydropathy
-0.37
-0.35
-0.40
Nb GST Nb HSA Nb PDZ
CDR3 hydro.
1
0
-1
-2
-3
-1.0
-0.9
-1.1
Nb GST Nb HSA Nb PDZ
g
17.5
14.1
18.2
S
G
R
10.6
7.6
7.1
24.8
PDZ
11.3
12.7
9.0
R
S
Y
13.6
G
T
7.5
7.5
38.5
GST
11.6
11.6
9.2
R
12.3
7.6
7.2
40.6
HSA
CDR3 head
Epitope pocket
h
Robustness
Geometry
Nb
Electrostatic
Paratope preferences
Versatility
Diversity
Specificity
Antigen

FIG. 6(a-h)

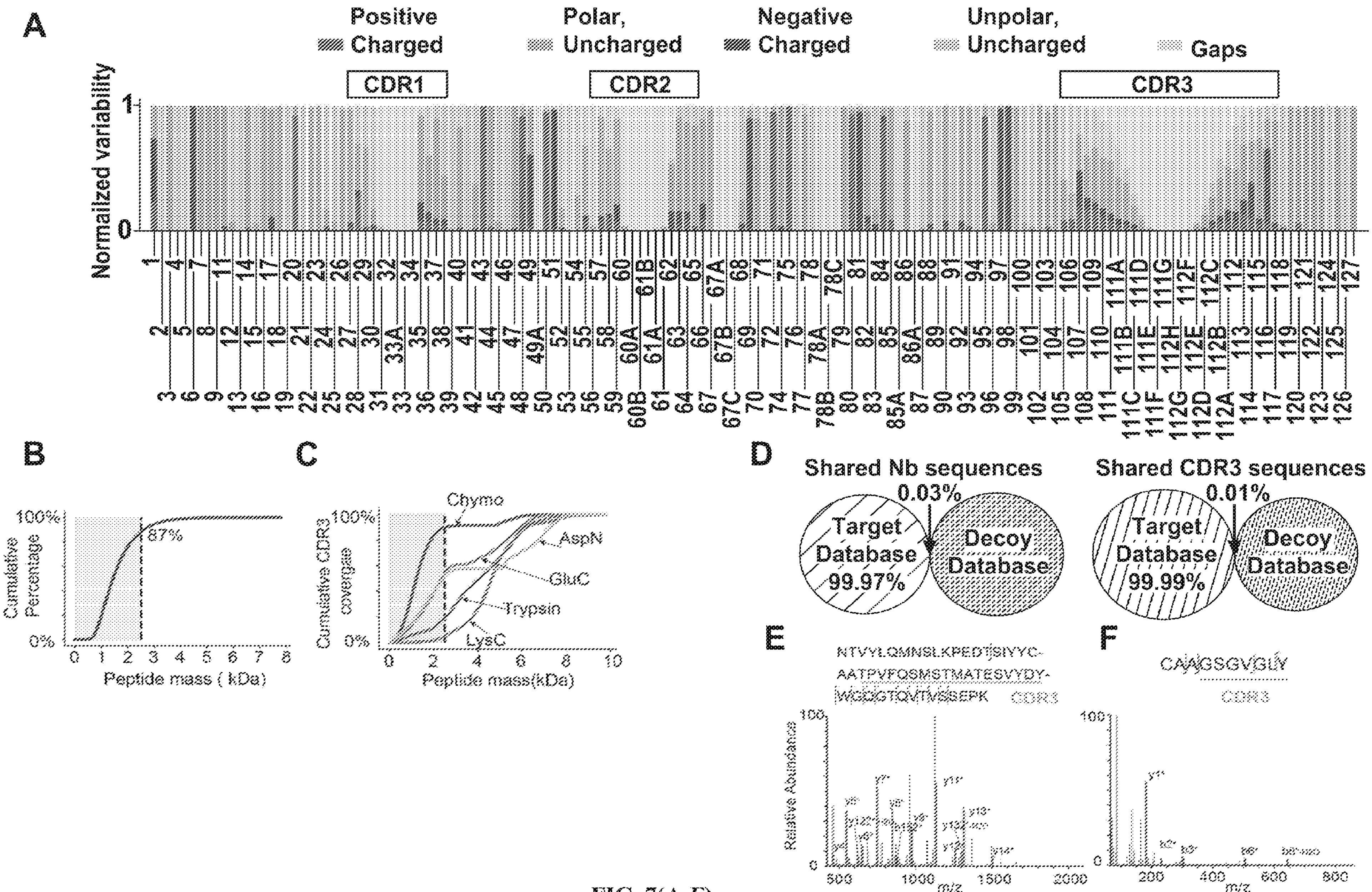

A
Positive Charged
Polar, Uncharged
Negative Charged
Unpolar, Uncharged
Gaps
CDR1
CDR2
CDR3
Normalized variability
1
0
1 2 3 4 5 6 7 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33A 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 49A 50 51 52 53 54 55 56 57 58 59 60 60A 60B 61B 61A 61 62 63 64 65 66 67 67A 67B 67C 68 69 70 71 72 74 75 76 77 78 78A 78B 78C 79 80 81 82 83 84 85 85A 86 86A 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 111A 111B 111C 111D 111E 111F 111G 112H 112G 112F 112E 112D 112C 112B 112A 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127
B
Cumulative Percentage
100%
0%
87%
0 1 2 3 4 5 6 7 8
Peptide mass (kDa)
C
Cumulative CDR3 covergae
100%
0%
Chymo
AspN
GluC
Trypsin
LysC
0 2 4 6 8 10
Peptide mass(kDa)
D
Shared Nb sequences
0.03%
Target Database 99.97%
Decoy Database
Shared CDR3 sequences
0.01%
Target Database 99.99%
Decoy Database
E
NTVYLQMNSLKPEDTSIYYC-
AATPVFQSMSTMATESVYDY-
WGQGTQVTVSSEPK
CDR3
Relative Abundance
100
0
500 1000 1500 2000
m/z
F
CAAGSGVGLY
CDR3
100
0
200 400 600 800
m/z

FIG. 7(A-F)

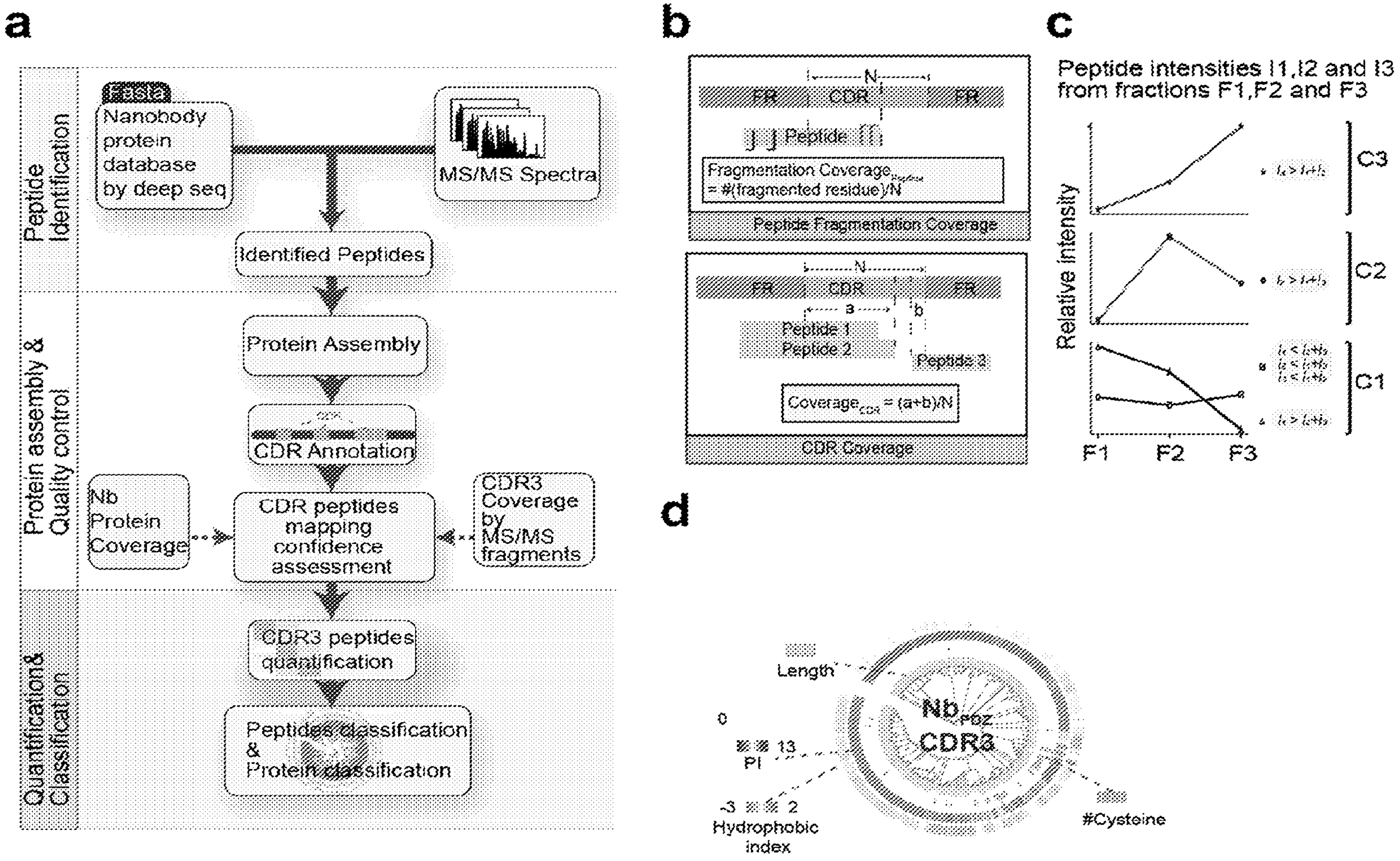

a
Peptide Identification
Fasta
Nanobody protein database by deep seq
MS/MS Spectra
Identified Peptides
Protein assembly & Quality control
Protein Assembly
CDR
CDR Annotation
Nb Protein Coverage
CDR peptides mapping confidence assessment
CDR3 Coverage by MS/MS fragments
Quantification& Classification
CDR3 peptides quantification
Peptides classification & Protein classification
b
N
FR
CDR
FR
Peptide
Fragmentation Coverage$_{Peptide}$ = #(fragmented residue)/N
Peptide Fragmentation Coverage
N
FR
CDR
FR
a
b
Peptide 1
Peptide 2
Peptide 3
Coverage$_{CDR}$ = (a+b)/N
CDR Coverage
c
Peptide intensities I1,I2 and I3 from fractions F1,F2 and F3
Relative intensity
F1
F2
F3
$I_3 > I_1+I_2$
C3
$I_2 > I_1+I_3$
C2
$I_1 < I_2+I_3$
$I_2 < I_1+I_3$
$I_3 < I_1+I_2$
$I_1 > I_2+I_3$
C1
d
Length
0
13
PI
-3
2
Hydrophobic index
Nb$_{PDZ}$
CDR3
#Cysteine

FIG. 8(a-d)

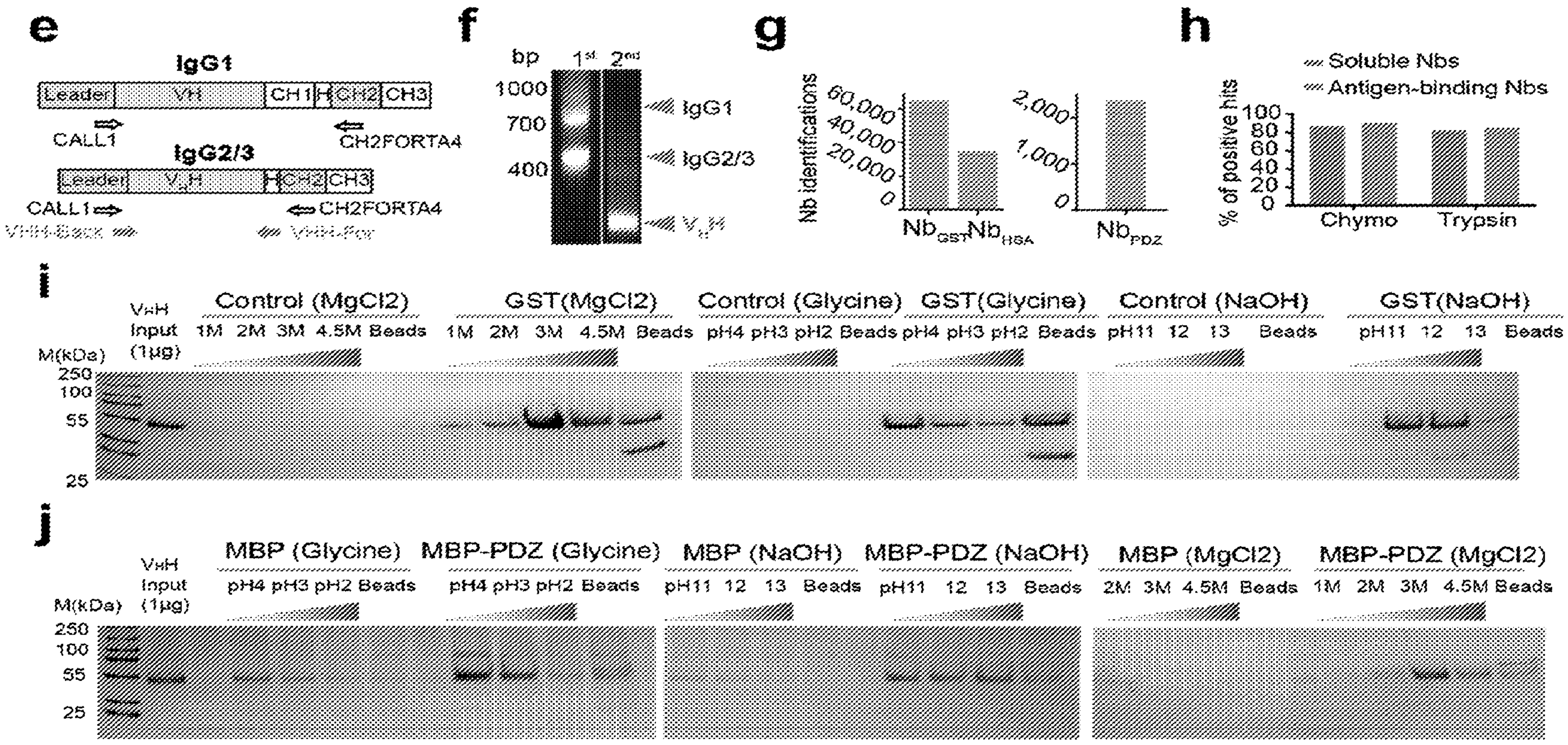
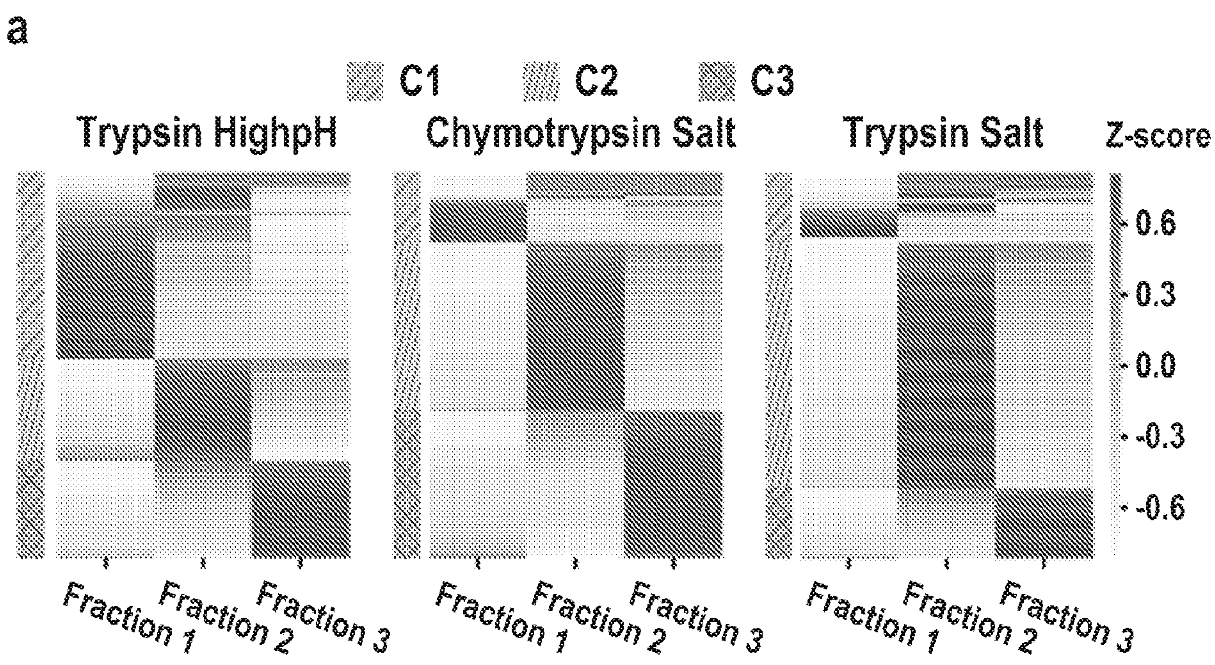

e
IgG1
Leader
VH
CH1
H
CH2
CH3
CALL1
CH2FORTA4
IgG2/3
Leader
V$_H$H
H
CH2
CH3
CALL1
CH2FORTA4
VHH-Back
VHH-For
f
bp
1st
2nd
1000
700
400
IgG1
IgG2/3
V$_H$H
g
Nb identifications
60,000
40,000
20,000
0
Nb$_{GST}$
Nb$_{HSA}$
2,000
1,000
0
Nb$_{PDZ}$
h
Soluble Nbs
Antigen-binding Nbs
% of positive hits
100
80
60
40
20
0
Chymo
Trypsin
i
V$_H$H Input (1µg)
M(kDa)
250
100
55
25
Control (MgCl2)
1M 2M 3M 4.5M Beads
GST(MgCl2)
1M 2M 3M 4.5M Beads
Control (Glycine)
pH4 pH3 pH2 Beads
GST(Glycine)
pH4 pH3 pH2 Beads
Control (NaOH)
pH11 12 13 Beads
GST(NaOH)
pH11 12 13 Beads
j
V$_H$H Input (1µg)
M(kDa)
250
100
55
25
MBP (Glycine)
pH4 pH3 pH2 Beads
MBP-PDZ (Glycine)
pH4 pH3 pH2 Beads
MBP (NaOH)
pH11 12 13 Beads
MBP-PDZ (NaOH)
pH11 12 13 Beads
MBP (MgCl2)
2M 3M 4.5M Beads
MBP-PDZ (MgCl2)
1M 2M 3M 4.5M Beads

FIG. 8(e-j)


a
C1
C2
C3
Trypsin HighpH
Chymotrypsin Salt
Trypsin Salt
Z-score
0.6
0.3
0.0
-0.3
-0.6
Fraction 1
Fraction 2
Fraction 3
Fraction 1
Fraction 2
Fraction 3
Fraction 1
Fraction 2
Fraction 3

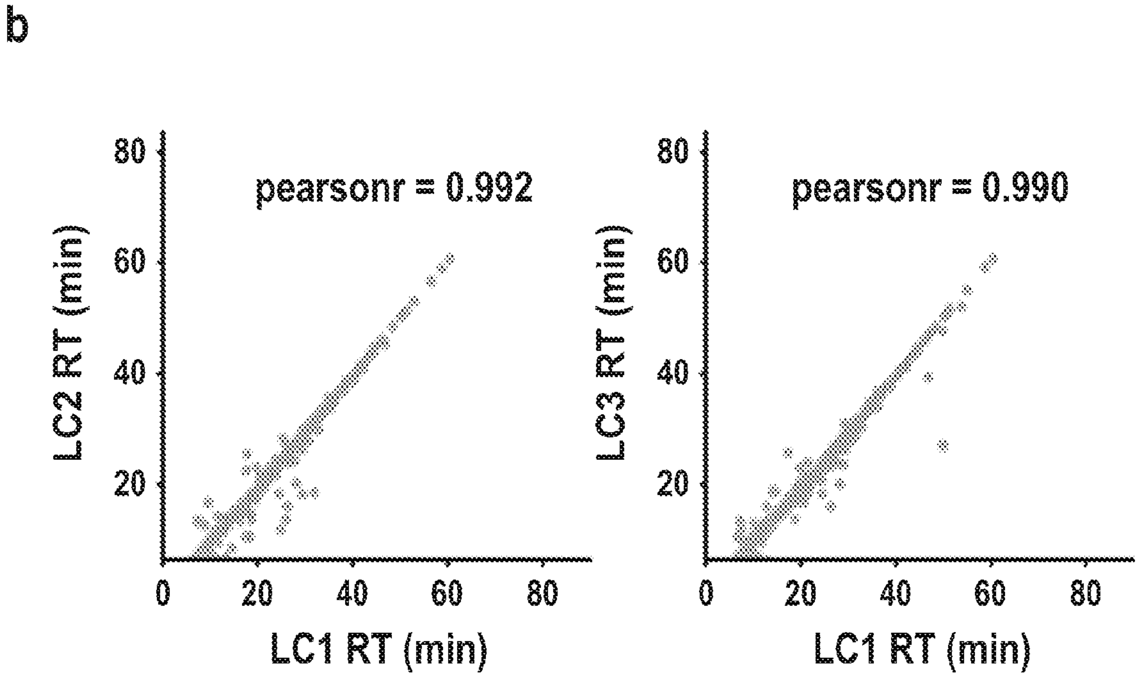

b
pearsonr = 0.992
LC2 RT (min)
LC1 RT (min)
80
60
40
20
0
20
40
60
80
pearsonr = 0.990
LC3 RT (min)
LC1 RT (min)
80
60
40
20
0
20
40
60
80

FIG. 9(a-b)

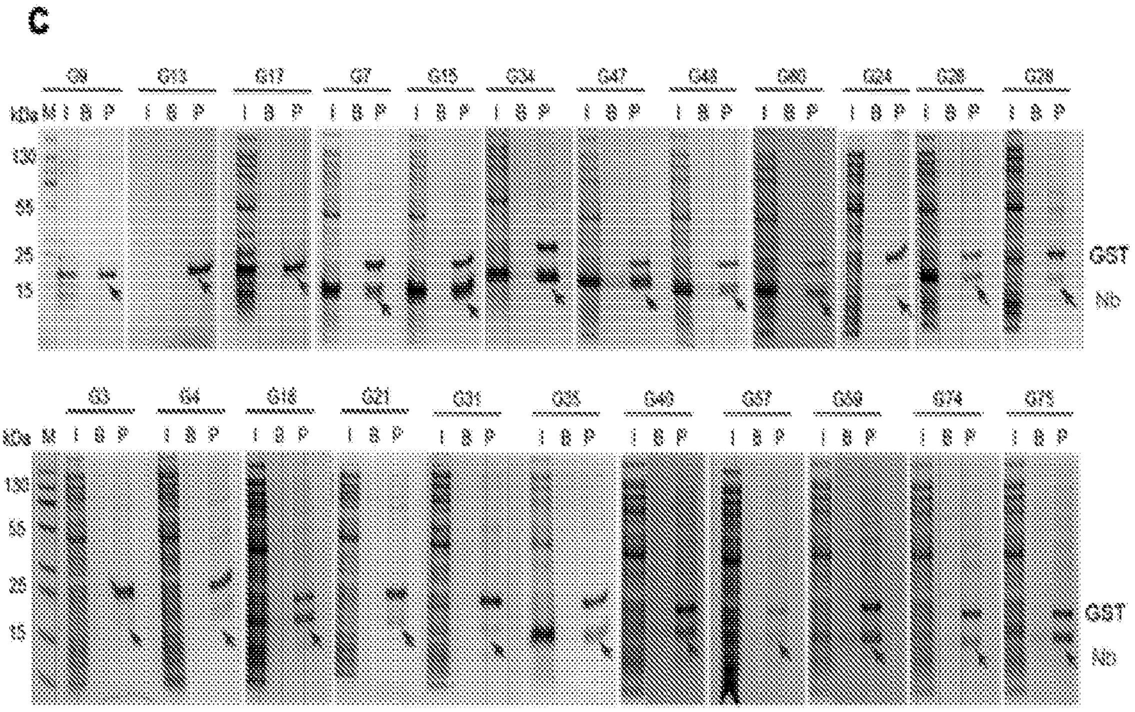

C
G9
G13
G17
G7
G15
G34
G47
G48
G60
G24
G26
G29
kDa M I S P
I S P
130
55
25
15
GST
Nb
G3
G4
G18
G21
G31
G35
G40
G57
G59
G74
G75
kDa M I S P
130
55
25
15
GST
Nb

FIG. 9(c)

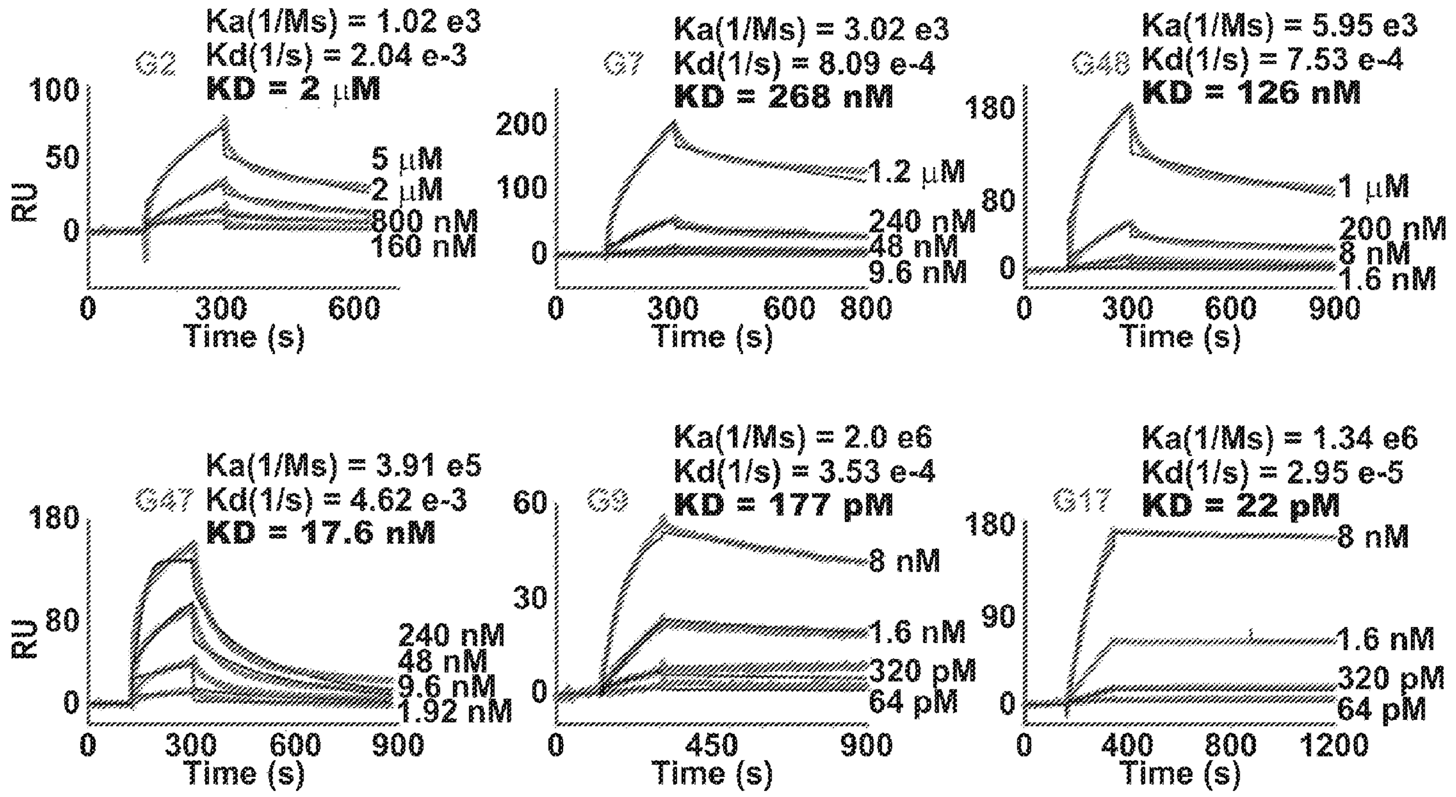

G2
Ka(1/Ms) = 1.02 e3
Kd(1/s) = 2.04 e-3
KD = 2 µM
RU
100
50
0
5 µM
2 µM
800 nM
160 nM
0
300
600
Time (s)
G7
Ka(1/Ms) = 3.02 e3
Kd(1/s) = 8.09 e-4
KD = 268 nM
200
100
0
1.2 µM
240 nM
48 nM
9.6 nM
0
300
600
800
Time (s)
G48
Ka(1/Ms) = 5.95 e3
Kd(1/s) = 7.53 e-4
KD = 126 nM
180
80
0
1 µM
200 nM
8 nM
1.6 nM
0
300
600
900
Time (s)
G47
Ka(1/Ms) = 3.91 e5
Kd(1/s) = 4.62 e-3
KD = 17.6 nM
RU
180
80
0
240 nM
48 nM
9.6 nM
1.92 nM
0
300
600
900
Time (s)
G9
Ka(1/Ms) = 2.0 e6
Kd(1/s) = 3.53 e-4
KD = 177 pM
60
30
0
8 nM
1.6 nM
320 pM
64 pM
0
450
900
Time (s)
G17
Ka(1/Ms) = 1.34 e6
Kd(1/s) = 2.95 e-5
KD = 22 pM
180
90
0
8 nM
1.6 nM
320 pM
64 pM
0
400
800
1200
Time (s)

FIG. 9(d)

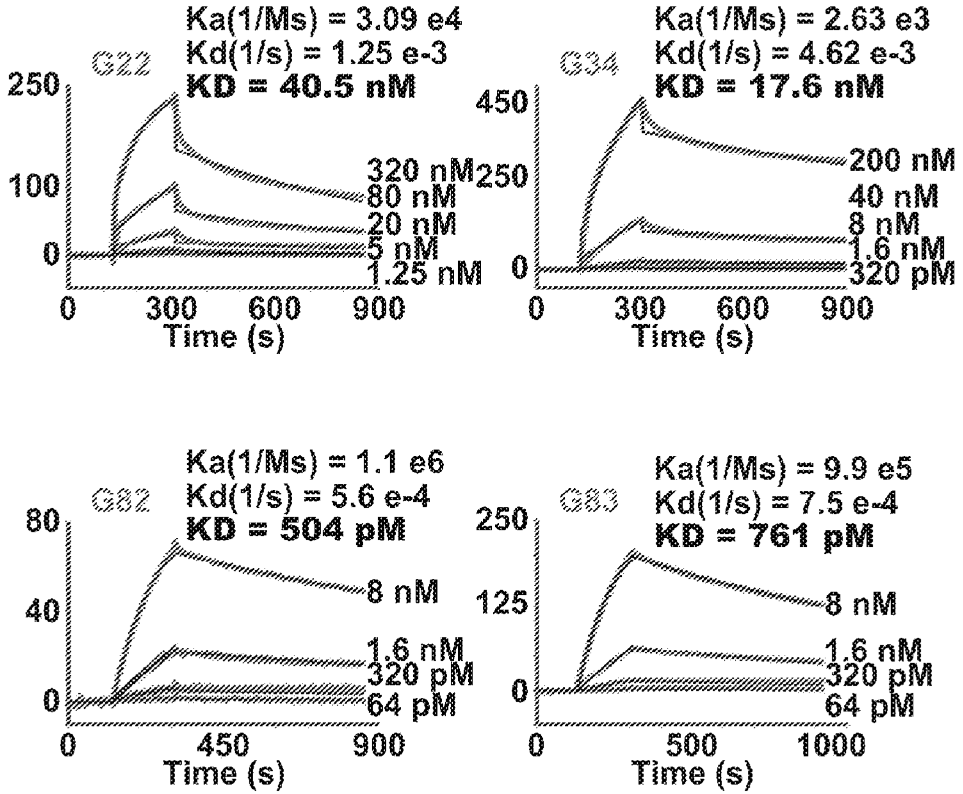

G22
Ka(1/Ms) = 3.09 e4
Kd(1/s) = 1.25 e-3
KD = 40.5 nM
250
100
0
320 nM
80 nM
20 nM
5 nM
1.25 nM
0
300
600
900
Time (s)
G34
Ka(1/Ms) = 2.63 e3
Kd(1/s) = 4.62 e-3
KD = 17.6 nM
450
250
0
200 nM
40 nM
8 nM
1.6 nM
320 pM
0
300
600
900
Time (s)
G82
Ka(1/Ms) = 1.1 e6
Kd(1/s) = 5.6 e-4
KD = 504 pM
80
40
0
8 nM
1.6 nM
320 pM
64 pM
0
450
900
Time (s)
G83
Ka(1/Ms) = 9.9 e5
Kd(1/s) = 7.5 e-4
KD = 761 pM
250
125
0
8 nM
1.6 nM
320 pM
64 pM
0
500
1000
Time (s)

FIG. 9(d) CONT.

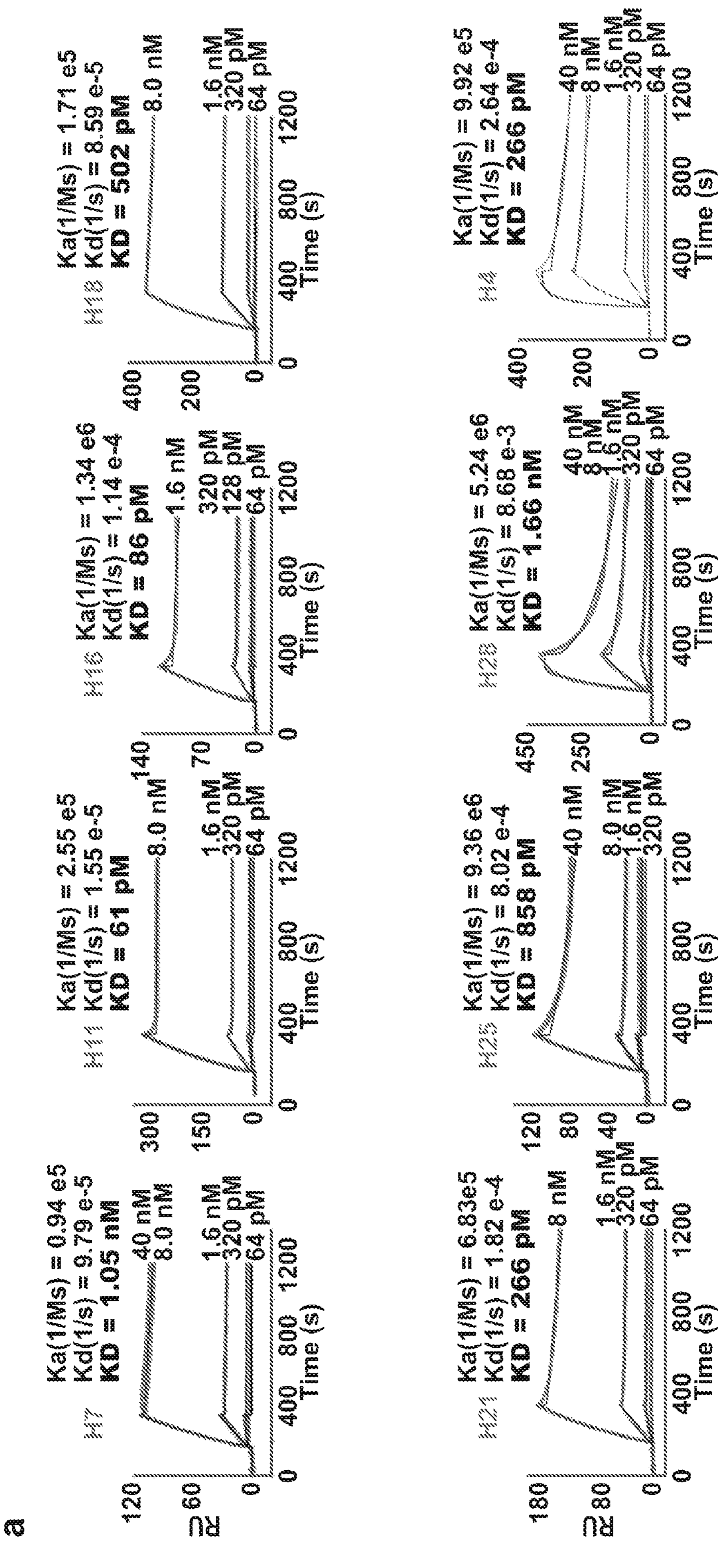

a
RU
H7
Ka(1/Ms) = 0.94 e5
Kd(1/s) = 9.79 e-5
KD = 1.05 nM
40 nM
8.0 nM
1.6 nM
320 pM
64 pM
120
60
0
0
400
800
1200
Time (s)
H11
Ka(1/Ms) = 2.55 e5
Kd(1/s) = 1.55 e-5
KD = 61 pM
8.0 nM
1.6 nM
320 pM
64 pM
300
150
0
H16
Ka(1/Ms) = 1.34 e6
Kd(1/s) = 1.14 e-4
KD = 86 pM
1.6 nM
320 pM
128 pM
64 pM
140
70
0
H18
Ka(1/Ms) = 1.71 e5
Kd(1/s) = 8.59 e-5
KD = 502 pM
8.0 nM
1.6 nM
320 pM
64 pM
400
200
0
H21
Ka(1/Ms) = 6.83e5
Kd(1/s) = 1.82 e-4
KD = 266 pM
8 nM
1.6 nM
320 pM
64 pM
180
80
0
H25
Ka(1/Ms) = 9.36 e6
Kd(1/s) = 8.02 e-4
KD = 858 pM
40 nM
8.0 nM
1.6 nM
320 pM
120
80
40
0
H28
Ka(1/Ms) = 5.24 e6
Kd(1/s) = 8.68 e-3
KD = 1.66 nM
40 nM
8 nM
1.6 nM
320 pM
64 pM
450
250
0
H4
Ka(1/Ms) = 9.92 e5
Kd(1/s) = 2.64 e-4
KD = 266 pM
40 nM
8 nM
1.6 nM
320 pM
64 pM
400
200
0

FIG. 10(a)

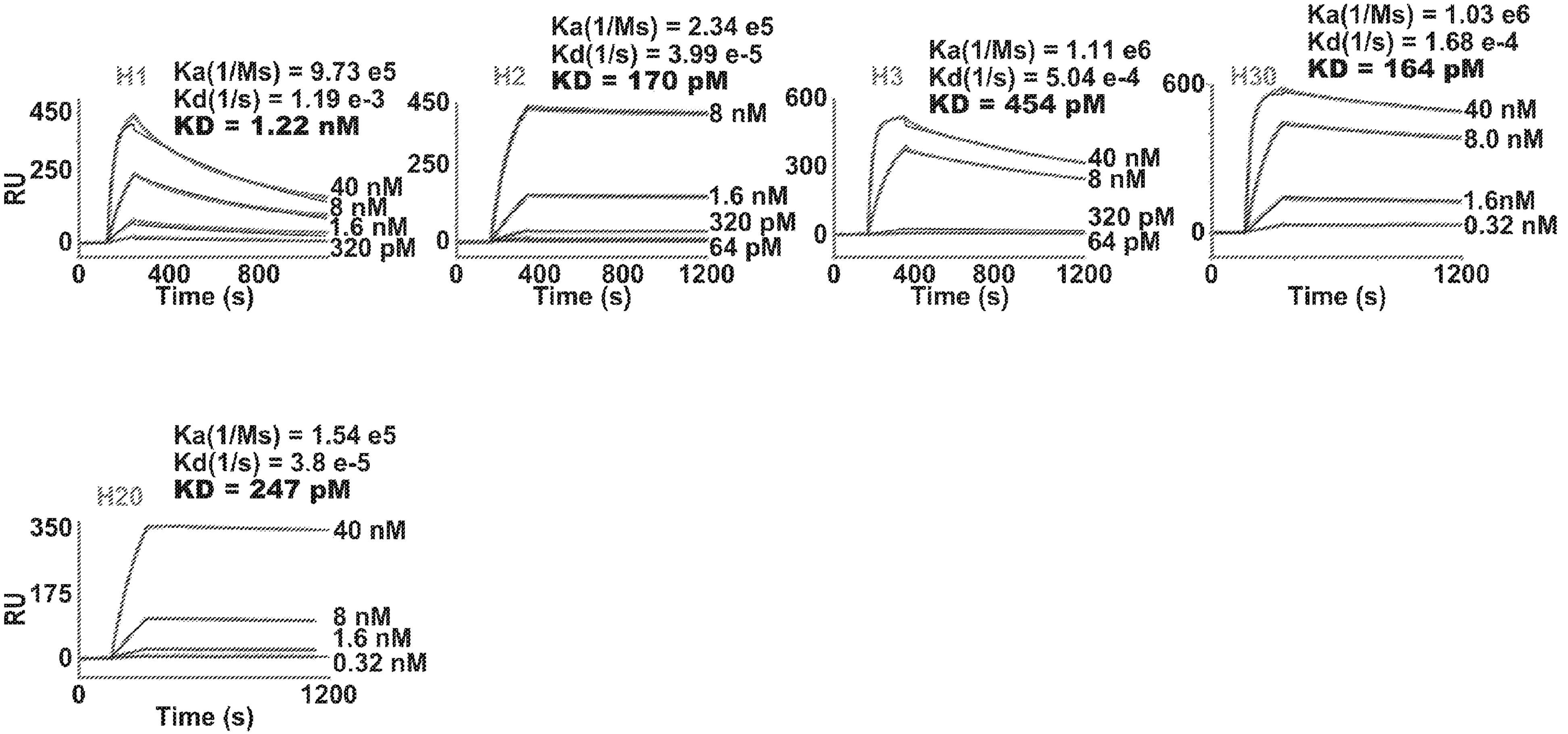

H1
Ka(1/Ms) = 9.73 e5
Kd(1/s) = 1.19 e-3
KD = 1.22 nM
450
250
0
RU
40 nM
8 nM
1.6 nM
320 pM
0
400
800
Time (s)
H2
Ka(1/Ms) = 2.34 e5
Kd(1/s) = 3.99 e-5
KD = 170 pM
450
250
0
8 nM
1.6 nM
320 pM
64 pM
0
400
800
1200
Time (s)
H3
Ka(1/Ms) = 1.11 e6
Kd(1/s) = 5.04 e-4
KD = 454 pM
600
300
0
40 nM
8 nM
320 pM
64 pM
0
400
800
1200
Time (s)
H30
Ka(1/Ms) = 1.03 e6
Kd(1/s) = 1.68 e-4
KD = 164 pM
600
0
40 nM
8.0 nM
1.6nM
0.32 nM
0
1200
Time (s)
H20
Ka(1/Ms) = 1.54 e5
Kd(1/s) = 3.8 e-5
KD = 247 pM
350
175
0
RU
40 nM
8 nM
1.6 nM
0.32 nM
0
1200
Time (s)

FIG. 10(a) CONT.

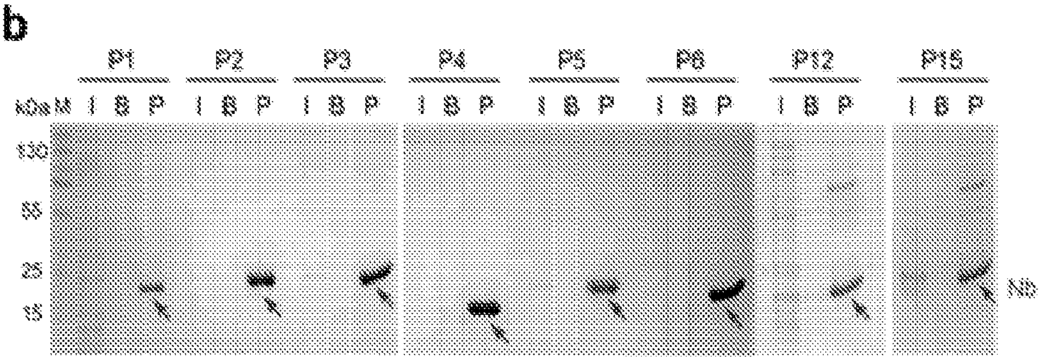


FIG. 10(b)

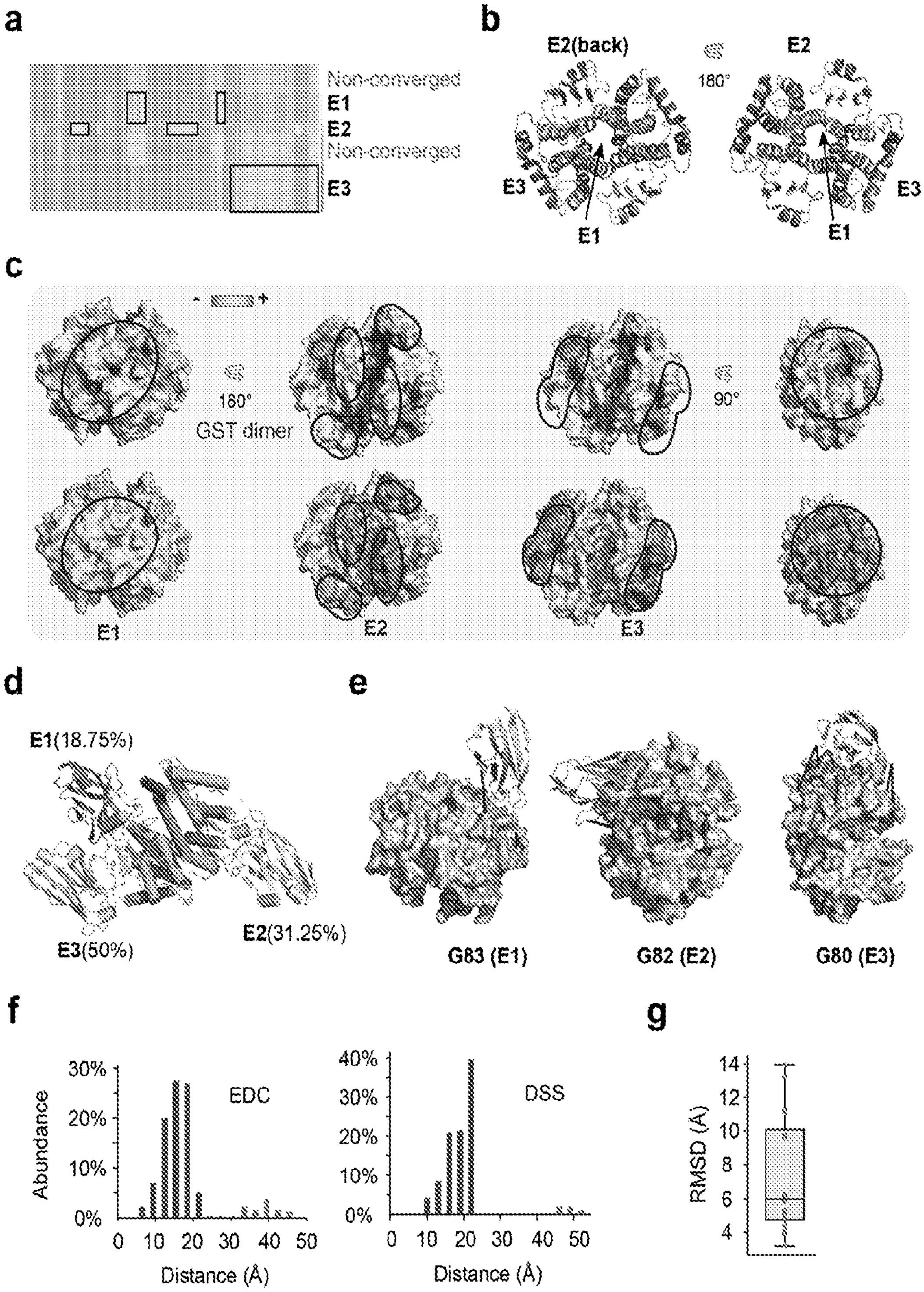

a
Non-converged
E1
E2
Non-converged
E3
b
E2(back)
180°
E2
E3
E1
E3
E1
c
- +
180°
GST dimer
90°
E1
E2
E3
d
E1(18.75%)
E3(50%)
E2(31.25%)
e
G83 (E1)
G82 (E2)
G80 (E3)
f
30%
20%
10%
0%
Abundance
EDC
0
10
20
30
40
50
Distance (Å)
40%
30%
20%
10%
0%
DSS
0
10
20
30
40
50
Distance (Å)
g
14
12
10
8
6
4
RMSD (Å)

FIG. 11(a-g)

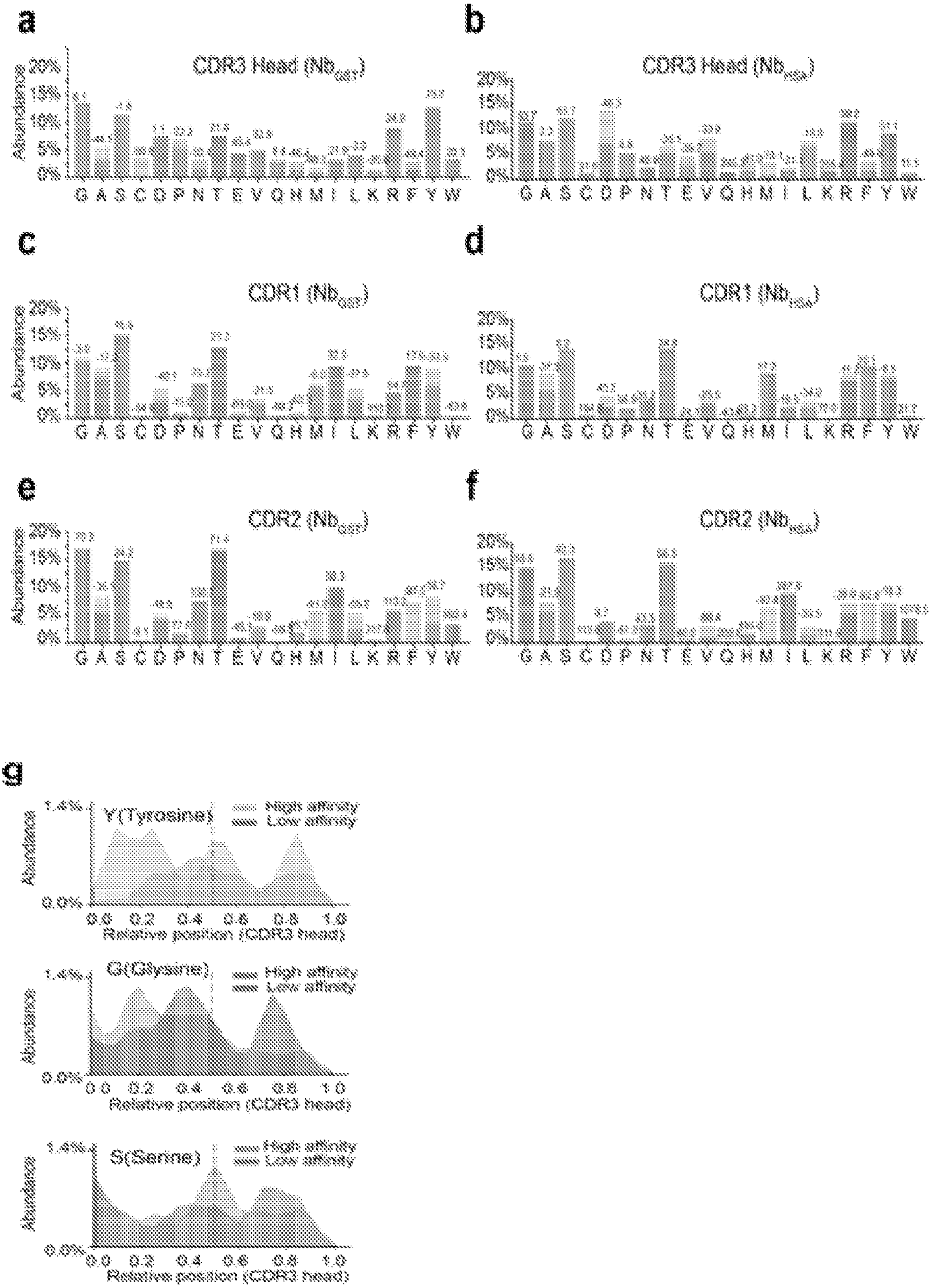

a
CDR3 Head (Nb_{GST})
Abundance
20%
15%
10%
5%
0%
G A S C D P N T E V Q H M I L K R F Y W
b
CDR3 Head (Nb_{HSA})
c
CDR1 (Nb_{GST})
d
CDR1 (Nb_{HSA})
e
CDR2 (Nb_{GST})
f
CDR2 (Nb_{HSA})
g
Y(Tyrosine)
High affinity
Low affinity
1.4%
0.0%
0.0 0.2 0.4 0.6 0.8 1.0
Relative position (CDR3 head)
G(Glysine)
S(Serine)

FIG. 12(a-g)

h

Score	Identities	Positives	Gaps
975 bits(2520)	462/608(76%)	530/608(87%)	1/608(0%)

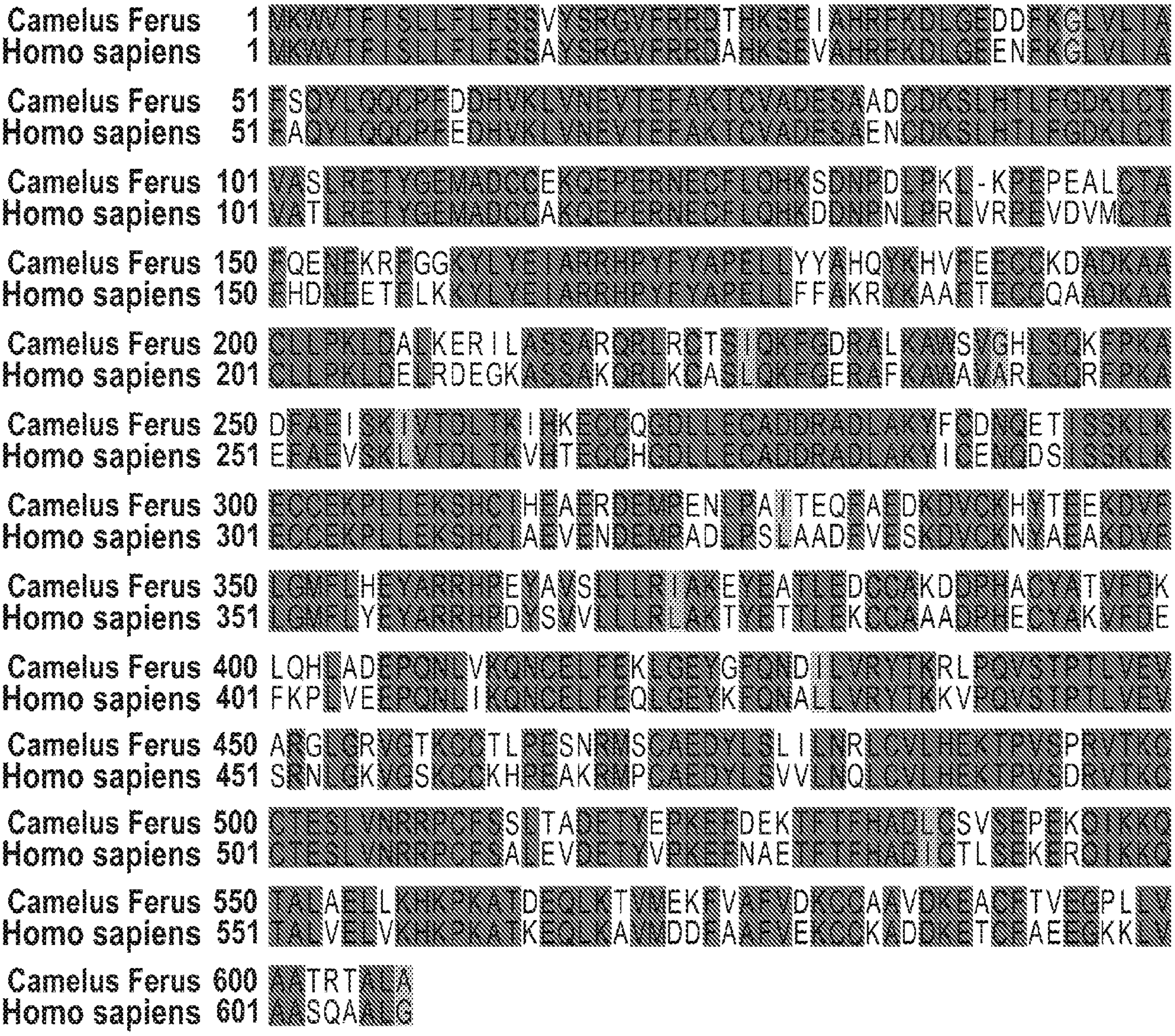


FIG. 12(h)

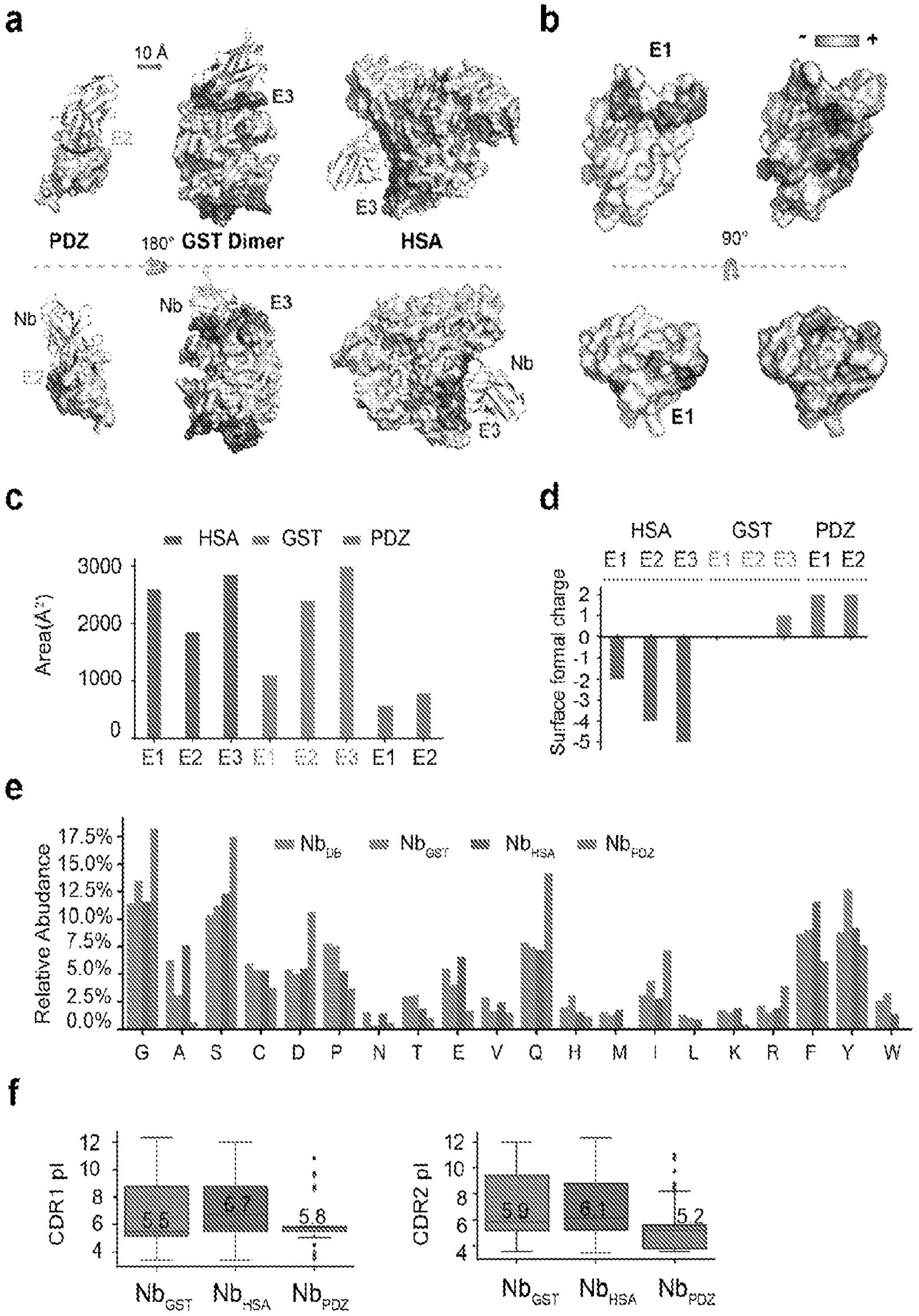

a
10 Å
E3
E2
PDZ
180°
GST Dimer
HSA
Nb
E3
E2
E3
Nb
E3
b
E1
- +
90°
E1
c
HSA
GST
PDZ
Area(Å²)
3000
2000
1000
0
E1 E2 E3 E1 E2 E3 E1 E2
d
HSA GST PDZ
E1 E2 E3 E1 E2 E3 E1 E2
Surface formal charge
2
1
0
-1
-2
-3
-4
-5
e
Relative Abudance
17.5%
15.0%
12.5%
10.0%
7.5%
5.0%
2.5%
0.0%
Nb_{DB}
Nb_{GST}
Nb_{HSA}
Nb_{PDZ}
G A S C D P N T E V Q H M I L K R F Y W
f
CDR1 pI
12
10
8
6
4
5.5
6.7
5.8
Nb_{GST} Nb_{HSA} Nb_{PDZ}
CDR2 pI
12
10
8
6
4
5.9
6.1
5.2
Nb_{GST} Nb_{HSA} Nb_{PDZ}

FIG. 13(a-f)

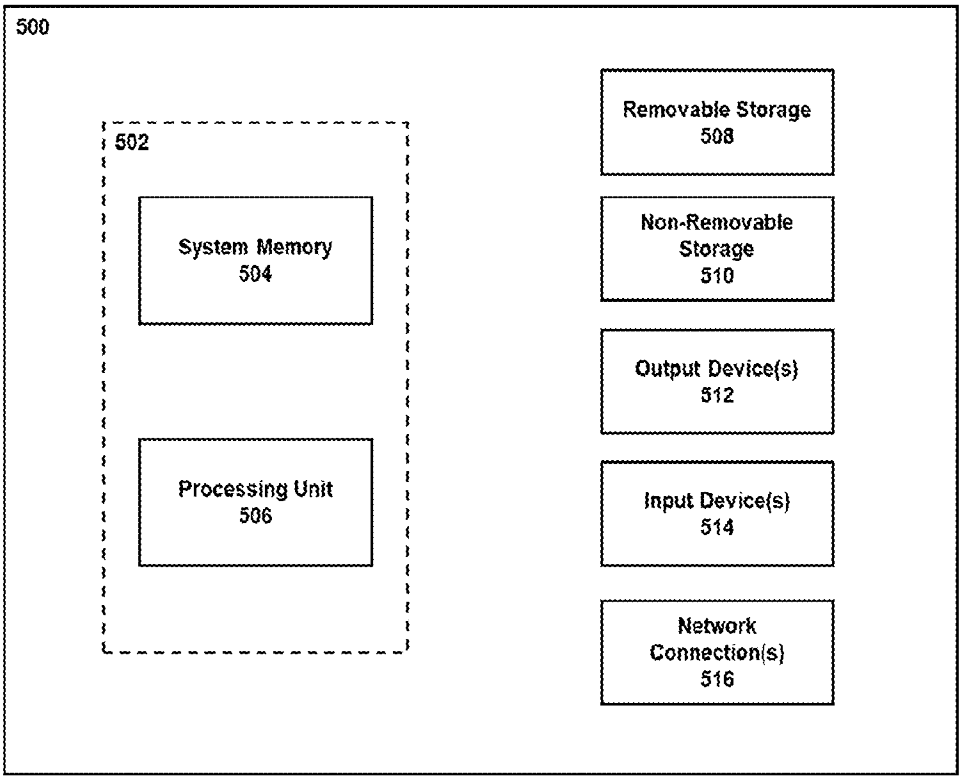

500
502
System Memory
504
Processing Unit
506
Removable Storage
508
Non-Removable
Storage
510
Output Device(s)
512
Input Device(s)
514
Network
Connection(s)
516

FIG. 14

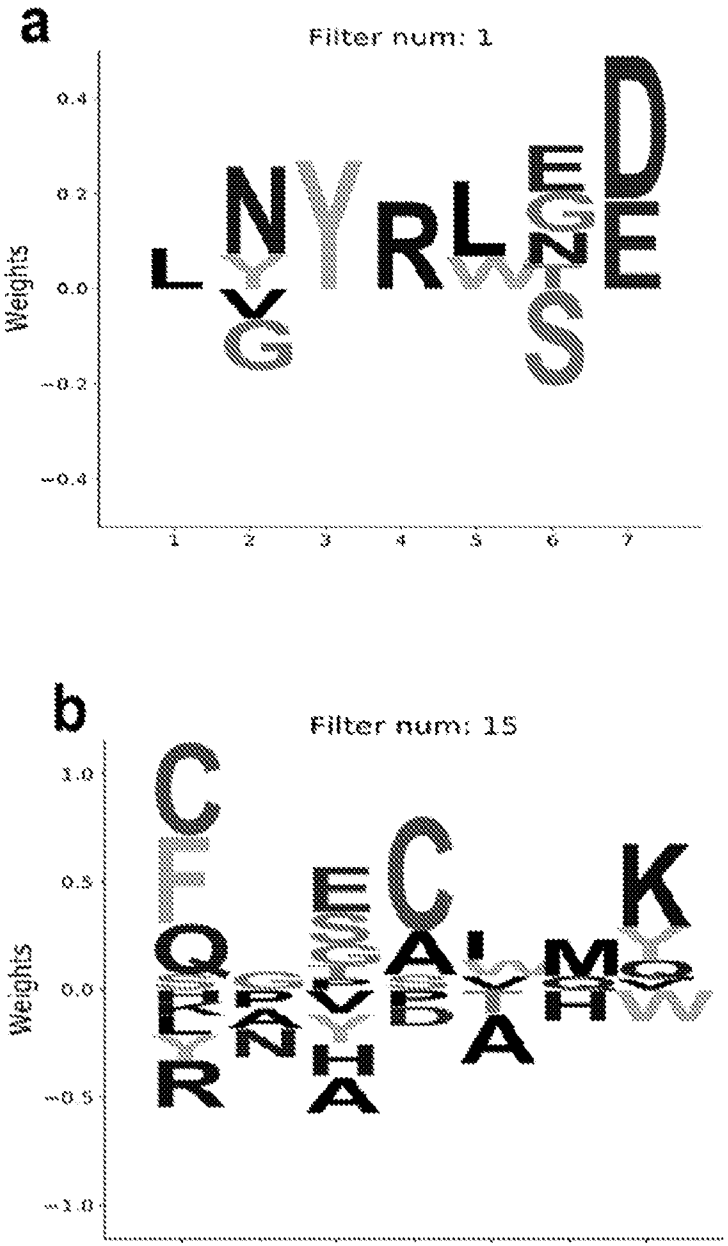

a
Filter num: 1
Weights
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
6
7
b
Filter num: 15
Weights
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
5
6
7

FIG. 15(a-b)

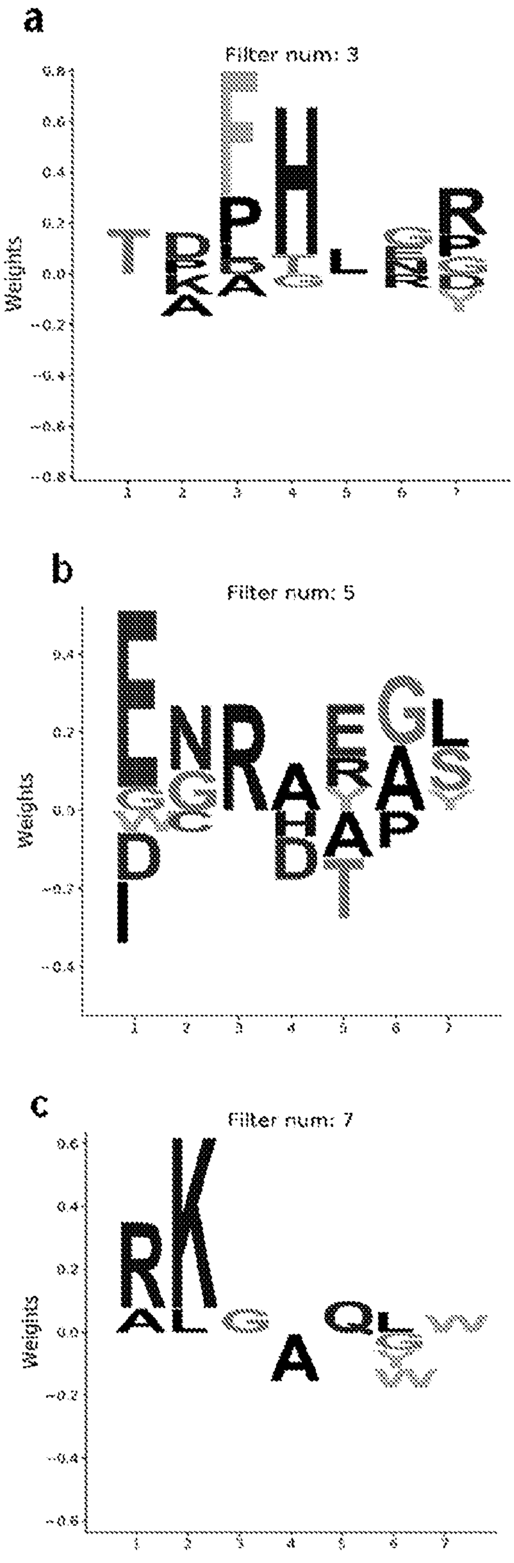

a
Filter num: 3
Weights
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
1
2
3
4
5
6
7
T
F
H
P
L
R
A
b
Filter num: 5
Weights
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
6
7
E
N
R
A
G
L
S
D
I
H
P
T
c
Filter num: 7
Weights
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
1
2
3
4
5
6
7
R
K
A
L
G
Q
W

FIG. 16(a-c)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 21/29869

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G01N 33/68, C07K 16/00 (2021.01)

CPC - C07K 16/44, G16B 30/00, C07K 2317/569, C07K 2317/92, G01N 33/6857, C07K 2317/22, C07K 16/00, C12Q 1/6876

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2019/0391159 A1 (The Rockefeller University), 26 December 2019 (26.12.2019), entire document, especially Abstract; para [0012], [0032]-[0033] -	1-6
A	US 2016/0347826 A1 (Bio-Rad Laboratories, Inc.), 01 December 2016 (01.12.2016), entire document, especially Abstract; para [0018], [0062]-[0064]	1-6
A	US 2014/0314832 A1 (Shanghai Pulmonary Hospital et al.), 23 October 2014 (23.10.2014), entire document, especially Abstract; para [0045]-[0047], [0124]-[0127]-	1-6
A	US 2016/0068600 A1 (THE UNIVERSITY COURT OF THE UNIVERSITY OF ABERDEEN), 10 March 2016 (10.03.2016), entire document	1-6
A	WO 2017/210104 A1 (PIERCE BIOTECHNOLOGY INC.), 07 December 2017 (07.12.2017), entire document	1-6
A	US 2014/0206579 A1 (Syndecion, LLC), 24 July 2014 (24.07.2014), entire document	1-6

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 15 September 2021	Date of mailing of the international search report OCT 04 2021
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Kari Rodriquez Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 21/29869

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. □ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. □ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 7-11, 19-23 and 26
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

--- (See Continuation in Supplemental Box) ---

1. □ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. □ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. □ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Claims 1-6

Remark on Protest

□ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

□ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

□ No protest accompanied the payment of additional search fees.

Continuation of:
Box III. Observations where unity of invention is lacking

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I- Claims 1-6 are directed to a method of identifying a group of complementarity determining region (CDR)3, 2 and/or 1 nanobody amino acid sequences (CDR3, CDR2 and/or CDR1 sequences).

Group II - Claims 12 is directed to a nanobody comprising an amino acid sequence selected from SEQ ID NOs: 1-2536 and SEQ ID NOs:2665-2667.

Group III - Claims 13-18 are directed to inferring an antigen affinity based on the quantified abundance of the one or more nondiscarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence.

Group IV - Claims 24-25 are directed to a method for determining antigen affinity of nanobody peptide sequences using trained deep learning model.

The inventions listed as Groups I-IV do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Special Technical Features:

The invention of Group I included the features of a. obtaining a blood sample from a camelid immunized with an antigen; b. using the blood sample to obtain a nanobody cDNA library; c. identifying the sequence of each cDNA in the library; d. isolating nanobodies from the same or a second blood sample from the camelid immunized with the antigen; e. digesting the nanobodies with trypsin or chymotrypsin to create a group of digestion products; f. performing a mass spectrometry analysis of the digestion products to obtain mass spectrometry data; g. selecting sequences identified in step c. that correlate with the mass spectrometry data; i. selecting from the CDR3, CDR2 and/or CDR1 region sequences of step h. those sequences having equal to or more than a required fragmentation coverage percentage; wherein the fragmentation coverage percentage is determined by a formula f(x,chymotrypsin) = 0.0023x2-0.0497x+0.7723,x[5,30] when chymotrypsin is used in step e. or a formula f(x,trypsin)=0.00006xpower2 - 0.00444x+0.9194, x[5,30] when trypsin is used in step e., and wherein x is the length of the CDR3, CDR2 or CDR1 region sequence, respectively; and j. wherein the selected sequences of step i. comprise a group having the reduced number of false positive CDR3, CDR2 and/or CDR1 sequences, not required by any other group.

The invention of Group II included the features of a nanobody comprising an amino acid sequence selected from SEQ ID NOs: 1-2536 and SEQ ID NOs: 2665-2667, not required by any other group.

The invention of Group III included the features of applying a fragmentation filter to discard one or more false positive CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence; quantifying an abundance of one or more non-discarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence; and inferring an antigen affinity based on the quantified abundance of the one or more nondiscarded CDR3, CDR2 and/or CDR1 regions of the nanobody peptide sequence, not required by any other group.

The invention of Group IV included the features of inputting the nanobody peptide sequence into a trained deep learning model; classifying, using the trained deep learning model, the nanobody peptide sequence as having low antigen affinity or high antigen affinity, not required by any other group.

Common Technical Features

Groups I and II share the features of identifying sequences of CDR3, CDR2 and/or CDR1 regions in the sequences.

Groups II-IV share the features of a Nanobody.

Groups III-IV share the features of receiving a nanobody peptide sequence.

However, the shared technical features do not represent a contribution over prior art as being obvious over US 2016/0347826 A1 (Bio-Rad Laboratories, Inc.), 01 December 2016 (01.12.2016) in view of US 2014/0314832 A1 to Shanghai Pulmonary Hospital et al. (hereinafter Shanghai), 23 October 2014 (23.10.2014).

Bio-Rad Laboratories, Inc teaches identifying sequences of CDR3, CDR2 and/or CDR1 regions in the sequences (para [0018], [0062]-[0064] - determining regions CDR1, CDR2, and CDR3 sequences); Nanobody (para [0052]-[0053]- nanobody).
Shanghai teaches a Nanobody (para [0063]-[0066]- nanobodies); receiving a nanobody peptide sequence (para [0045]-[0047], [0124]-[0127]- synthetic method was used to obtain the polypeptide of the nanobody).

As the common features were known in the art at the time of the invention, this cannot be considered a common technical feature that would otherwise unify the groups. Therefore, Groups I-IV lack unity under PCT Rule 13.