Structure-based modeling of SARS-CoV-2 peptide/HLA-A02 antigens

Santrupti Nerli[1] and Nikolaos G. Sgourakis[2]
Email: nsgourak@ucsc.edu

[1]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA.
[2]Department of Chemistry and Biochemistry, University of California Santa Cruz, Santa Cruz, CA 95064, USA.

ABSTRACT
As a first step toward the development of diagnostic and therapeutic tools to fight the Coronavirus disease (COVID-19), we aim to characterize CD8+ T cell epitopes in the SARS-CoV-2 peptidome that can trigger adaptive immune responses. Here, we use RosettaMHC, a comparative modeling approach which leverages existing high-resolution X-ray structures from peptide/MHC complexes available in the Protein Data Bank, to derive physically realistic 3D models for high-affinity SARS-CoV-2 epitopes. We outline an application of our method to model 439 9mer and 279 10mer predicted epitopes displayed by the common allele HLA-A*02:01, and we make our models publicly available through an online database (https://rosettamhc.chemistry.ucsc.edu). As more detailed studies on antigen-specific T cell repertoires become available, RosettaMHC models of antigens from different strains and HLA alleles can be used as a basis to understand the link between peptide/HLA complex structure and surface chemistry with immunogenicity, in the context of SARS-CoV-2 infection.

An ongoing pandemic caused by the novel SARS coronavirus (SARS-CoV-2) has become the focus of extensive efforts to develop vaccines and antiviral therapies (1). Immune modulatory interferons, which promote a widespread antiviral reaction in infected cells, and inhibition of pro-inflammatory cytokine function through anti-IL-6/IL-6R antibodies, have been proposed as possible COVID-19 therapies (2, 3). However, stimulating a targeted T cell response against specific viral antigens is hampered by a lack of detailed knowledge of the immunodominant epitopes displayed by common Human Leukocyte Antigen (HLA) alleles across individuals ("public" epitopes). The molecules of the class I major histocompatibility complex (MHC-I, or HLA in humans) display on the cell surface a diverse pool of 8 to 15 amino acid peptides derived from the endogenous processing of proteins expressed inside the cell (4). This MHC-I restriction of peptide antigens provides jawed vertebrates with an essential mechanism for adaptive immunity: surveillance of the displayed peptide/MHC-I (pMHC-I) molecules by CD8+ cytotoxic T-lymphocytes allows detection of aberrant protein expression patterns, which signify viral infection and can trigger an adaptive immune response (5). A Recent study has shown important changes in T cell compartments during the acute phase of SARS-CoV-2 infection (6), suggesting that the ability to quantify antigen-specific T cells would provide new avenues for understanding

the expansion and contraction of the TCR repertoire in different disease cohorts and clinical settings. Given the reduction in breadth and functionality of the naïve T cell repertoire during aging (7), identifying a minimal set of viral antigens that can elicit a protective response will enable the design of diagnostic tools to monitor critical "gaps" in the T cell repertoire of high-risk cohorts, which can be addressed using peptide or epitope string DNA vaccines (8).

Human MHC-I molecules are extremely polymorphic, with thousands of known alleles in the classical HLA-A, -B and -C loci. Specific amino acid polymorphisms along the peptide-binding groove (termed A-F pockets) define a repertoire of 10^4-10^6 peptide antigens that can be recognized by each HLA allotype (9, 10). Several methods have been developed to predict the likelihood that a target peptide will bind to a given allele (reviewed in (11)). Generally these methods make use of available data sets in the Immune Epitope Database (12) to train artificial neural networks that predict binding scores, and their performance varies depending on peptide length and HLA allele representation in the database. Structure-based approaches have also been proposed to model the bound peptide conformation *de novo* (reviewed in (13)). These approaches utilize various MCMC-based schemes to optimize the backbone and side chain degrees of freedom of the peptide/MHC structure according to an all-atom scoring function derived from physical principles (14–16), that can be further enhanced using modified scoring terms (17) or mean field theory (18). While these methods do not rely on large training data sets, their performance is affected by bottlenecks in sampling of different backbone conformations, and any possible structural adaptations of the HLA peptide-binding groove.

Predicting the bound peptide conformation whose N- and C- termini are anchored within a fixed-length groove is a tractable modeling problem that can be addressed using standard comparative modeling approaches (19). In previous work focusing on the HLA-B*15:01 and HLA-A*01:01 alleles in the context of neuroblastoma neoantigens, we have found that a side chain optimization approach can yield accurate pMHC-I models for a pool of target peptides, provided that a reliable template of the same allele and peptide length can be identified in the database (20). In this approach (RosettaMHC), a local optimization of the backbone degrees of freedom is sufficient to capture minor (within 0.5 Å heavy atom RMSD) structural adaptations of the target peptide backbone relative to the conformation of the peptide in the template, used as a starting point. For HLA-A*02:01, the most common HLA allele among disease-relevant population cohorts (21), there is a large number of high-resolution X-ray structures available in the PDB (22), suggesting that a similar principle can be applied to produce models of candidate epitopes directly from the proteome of a pathogen of interest. Here, we apply RosettaMHC to derive structural models for all predicted HLA-A*02:01 epitopes from the ~30 kbp SARS-CoV-2 genome, and make our models publicly available through an online database. As detailed epitope mapping data from high-throughput tetramer staining (23–25) and T cell functional screens (26) become available for different clinical settings and disease cohorts, the models presented here can provide a toehold for understanding links between pMHC-I antigen structure and immunogenicity, with actionable value for the development of peptide vaccines to combat the disease.

Materials and Methods

Identification of SARS-CoV-2 peptide epitopes

The SARS-CoV-2 protein sequences were obtained from NCBI: (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2). From these sequences, we generated all possible peptides of lengths 9 and 10 (9,621 9mer and 9,611 10mer peptides). We applied NetMHCPan4.0 (2018 version) to derive binding scores to HLA-A*02:01, and retained only peptides classified as strong or weak binders (selected using the default percentile rank cut-off values).

Selection of PDB templates

To model SARS-CoV-2 epitopes on HLA-A*02:01, we identified 3D structures from the PDB that can be used as starting templates for comparative modeling. First, we selected all HLA-A*02:01 X-ray structures that are below 3.5 Å resolution and retained only those that have 100% identity to the HLA-A*02:01 heavy chain sequence (residues 1-180). We found 241 template structures bound to epitopes of lengths from 8 to 15 residues (of which 170 are 9mers and 61 are 10mers). For each SARS-CoV-2 target peptide, we select a set of candidate templates of the same length by matching the target peptide anchor positions (P2 and P9/P10) to peptides found in all template structures. Then, we utilized the BLOSUM62 (27) substitution matrix to score all remaining positions in the pairwise alignment of the target/template sequences, and the PDB template with the top alignment score was selected for modeling. For target peptides with no templates with matched peptide anchors in the database, we scored all positions in the pairwise alignment and select the top scoring template for modeling.

RosettaMHC modeling framework and database

RosettaMHC (manuscript in preparation) is a comparative modeling protocol developed using PyRosetta (28) to model pMHC-I complexes. The program accepts as input a list of peptide sequences, an HLA allele definition and a template PDB file (selected as described in the previous step). To minimize "noise" in the simulation from parts of the MHC-I fold that do not contribute to peptide binding, only the α_1 and α_2 domains are considered in all steps. For each peptide, a full alignment between the target and template peptide/MHC sequences is performed using clustal omega (29). The alignment is used as input to Rosetta's threading protocol (*partial_thread.<ext>*). From the threaded model, all residues in the MHC-I groove that are within a heavy-atom distance of 3.5 Å from the peptide are subjected to 10 independent all-atom refinement simulations using the FastRelax method (30) and a custom movemap file. Binding energies are extracted from the refined structures using interface analyzer protocol (*InterfaceAnalyzer.<ext>*). The top three models are selected based on the binding energies, and used to compute an average energy for each peptide in the input list. RosettaMHC models of SARS-CoV-2/HLA-A*02:01 epitopes are made available through an online database (see data availability). The website that hosts our database is constructed using the Django web framework.

Results and Discussion

Template identification for structure modeling using RosettaMHC

Our full workflow for template identification and structure modeling is outlined in Figure 1a, with a flowchart shown in Figure 1b. To identify all possible regular peptide binders to HLA-A*02:01 that are expressed by SARS-CoV-2, we used a recently annotated version of all open reading frames (ORFs) in the viral genome (31) from NCBI which was made available through the UCSC genome browser (32). We used 9- and 10- residue sliding windows to scan all protein sequences, since these are the optimum peptide lengths for binding to the HLA-A*02:01 groove (33). While spliced peptide epitopes (34) are not considered in the current study, this functionality can be added to our method at a later stage. Using NetMHCPan4.0 (35), we identified all 439 9mer and 279 10mer epitopes that are predicted to yield positive (classified as both weak and strong) binders. To further validate this peptide set and derive plausible 3D models of the peptide/HLA-A*02:01 complexes, we used a structure-guided approach, RosettaMHC, which aims to derive a physically realistic fitness score for each peptide in the HLA-A*02:01 binding groove using an annotated database of high-resolution structures and Rosetta's all-atom energy function (36). RosettaMHC leverages a database of 241 HLA-A*02:01 X-ray structures encompassing a range of bound peptides, to find the closest match to each target epitope predicted from the SARS-CoV-2 proteome. To identify the best template for structure modeling, we use matching criteria which first consider the peptide anchors (positions P2 and P9/P10 for 9mer/10mer epitopes), followed by a sequence similarity metric calculated from the alignment between the template and target peptide sequences. The template assignment statistics for the four different classes of SARS-CoV-2 epitopes in our set is shown in Figure 2a. Using these criteria, we find that we can cover the entire set of 718 predicted binders using a subset of 114 HLA-A*02:01 templates in our annotated, PDB-derived database (Figure 2b). Each target peptide sequence is then threaded onto the backbone of the best identified template, followed by all-atom refinement of the side chain and backbone degrees of freedom using Rosetta's Ref2015 energy function (36).

RosettaMHC models recapitulate features of high-resolution X-ray structures

The sequence logos derived from all predicted 9mer and 10mer binders by NetMHCPan4.0 adhere to the canonical HLA-A*02:01 motif, with a preference for hydrophobic, methyl-bearing side chains at the peptide anchor residues P2 and P9 (Figure 3a). The anchor residue preferences are recapitulated in representative 9mer and 10mer models of the two top binders in our set as ranked by Rosetta's energy (Figure 3c and 3d), corresponding to epitopes TMADLVYAL and FLFVAAIFYL derived from the RNA polymerase and nsp3 proteins, respectively, which are both encoded by *orf1ab* in the viral genome (NCBI Reference YP_009724389.1). In accordance with features seen in high-resolution structures of HLA-A*02:01-restricted epitopes, the peptides adopt an extended, bulged backbone conformation. The free N-terminus of both peptides is stabilized by a network of polar contacts with Tyr 7, Tyr 159, Tyr 171 and Glu 63 in the A- and B- pockets of

161 the HLA-A*02:01 groove. The Met (9mer) or Leu (10mer) side chain of P2 is buried in a B-pocket
162 hydrophobic cleft formed by Met 45 and Val 67. Equivalently, the C-terminus is coordinated
163 through polar contacts with Asp 77 and Lys 145 from opposite sides of the groove, with the Leu
164 P9/P10 anchor nestled in the F-pocket defined by the side chains of Leu 81, Tyr 116, Tyr 123 and
165 Trp 147. Residues P3-P8 form a series of backbone and side chain contacts with pockets C, D and
166 E, while most backbone amide and carbonyl groups form hydrogen bonds with the side chains of
167 residues lining the MHC-I groove. These high-resolution structural features are consistent across
168 low-energy models of all target peptides in our input set, suggesting that, when provided with a
169 large set of input templates, a combined threading and side chain optimization protocol can derive
170 physically realistic models.
171
172
173 *Comparison of binding energies for different peptide sets*
174 To evaluate the accuracy of our models and fitness of each peptide within the HLA-A*02:01
175 binding groove, we computed Rosetta binding energies across all complexes modeled for different
176 peptide sets. High binding energies can be used an additional metric to filter low-affinity peptides
177 in the NetMHCPan4.0 predictions, with the caveat that high energies can be also due to incomplete
178 optimization of the Rosetta energy function as a result of significant deviations between the target
179 and template backbone conformations, not captured by our protocol. We performed 10
180 independent side chain optimization calculations for each peptide, and the 3 lower-energy models
181 were selected as the final ensemble and used to compute an average binding energy. The results
182 for all 9mer peptides are summarized in Figures 3e, f, while additional results for 10mers are
183 provided through our web-interface and Supplemental Table 1. As a positive reference, we used
184 the binding energies of the idealized and relaxed PDB templates, which are at a local minimum of
185 the Rosetta scoring function. As a negative control for poor binders, we modeled decoy structures
186 of poly alanine peptide sequences, threaded onto the same PDB templates. We observe a
187 significant (26 kcal/mol) energy gap between the average binding energies for PDB templates
188 (positives) and poly alanine negative controls. The binding energies for our modeled 9mers from
189 the SARS-CoV-2 genome fall between the distributions corresponding to positive and negative
190 control samples, and show a bimodal distribution with significant overlap with the refined PDB
191 template energies (Figure 1e). Inspection of the binding energies of 12 9mer peptides in our set
192 that are homologous to peptides in the SARS viral genome with confirmed binding to HLA-
193 A*02:01 in the IEDB (12, 37) shows a similar distribution, with the binding energies of 9/12
194 peptides falling within the distribution of the refined PDB templates (red dots in Figure 3e).
195 Comparison of the distributions between epitopes that are classified as strong versus weak binders
196 by NetMHCPan4.0 shows a moderate bias towards lower binding energies for the strong binders
197 and a larger spread in energies for weak binders, likely due to suboptimal residues at the P2 and
198 P9 anchor positions (Figure 3f).
199 Together, these results suggest that the high-resolution features seen in our models (Figure 3c, d)
200 yield optimal binding energies for a significant fraction of the epitopes predicted by

201 NetMHCPan4.0 (up to 45/33% of strong binders and 30/25% of weak binders for 9mers/10mers,
202 respectively), which are comparable to locally refined PDB structures. The average binding
203 energies for each peptide are provided in our web-interface, Supplemental Table 1 and can be used
204 to further select high-affinity peptides, by considering the structural complementarity of each
205 target epitope within the HLA-A*02:01 peptide-binding groove.
206
207
208 *Surface features of peptide/HLA-A*02:01 models for T cell recognition*
209 Visualization of our models through an interactive online interface provides direct information on
210 SARS-CoV-2 peptide residues that are bulging out of the MHC-I groove, and are therefore
211 accessible to interactions with complementarity-determining regions (CDRs) of T cell receptors
212 (TCRs). Given that αβ TCRs generally employ a diagonal binding mode to engage pMHC-I
213 antigens where the CDR3α and CDR3β TCR loops form direct contacts with key peptide residues
214 (38, 39), knowledge of the surface features for different epitopes adds an extra layer of information
215 to interpret sequence variability between different viral strains. For other important antigens with
216 known structures in the PDB, such features can be derived from an annotated database connecting
217 pMHC-I/TCR co-crystal structures with biophysical binding data (40), and were recently
218 employed in an artificial neural network approach to predict the immunogenicity of different HLA-
219 A*02:01 bound peptides in the context of tumor neoantigen display (41). A separate study has
220 shown that the electrostatic compatibility between self vs foreign HLA surfaces can be used to
221 determine antibody alloimmune responses (42). Given that antibodies and TCRs use a common
222 fold and similar principles to engage pMHC-I molecules (39), it is likely that electrostatics play
223 an important role in recognition of peptide/HLA surfaces by their cognate TCRs in the context of
224 SARS-CoV-2 infection.
225 Surface electrostatics calculated using a numerical solution to the Poisson-Boltzmann Equation
226 (43) for our modeled peptide/HLA-A*02:01 complexes allow us to compare important features
227 for TCR recognition between different high-affinity epitopes (Figure 4). We observe a moderate
228 electropositive character of the HLA-A*02:01 α_1 helix, and a moderate negative potential on the
229 α_2 helix, which is consistent between complexes with different bound peptides. However, due to
230 substantial sequence variability in surface-exposed residues at the P2-P8 positions, we observe a
231 range of electrostatic features ranging from negative (sequence TMADLVYAL), to neutral
232 (sequence NLIDSYFVV) or positively charged (sequence KLWAQCVQL). Further classification
233 and ranking of the top binders in our set on the basis of their molecular surface features would
234 enable the selection of a diverse panel of peptides for high-throughput pMHC tetramer library
235 generation (23-25). Such libraries can be used to screen T cells of COVID-19 patients, recovered
236 individuals and healthy donors from high-risk groups towards identifying a minimal panel of
237 epitopes to monitor T cell responses in patients, and to select the top candidates for vaccine
238 development.

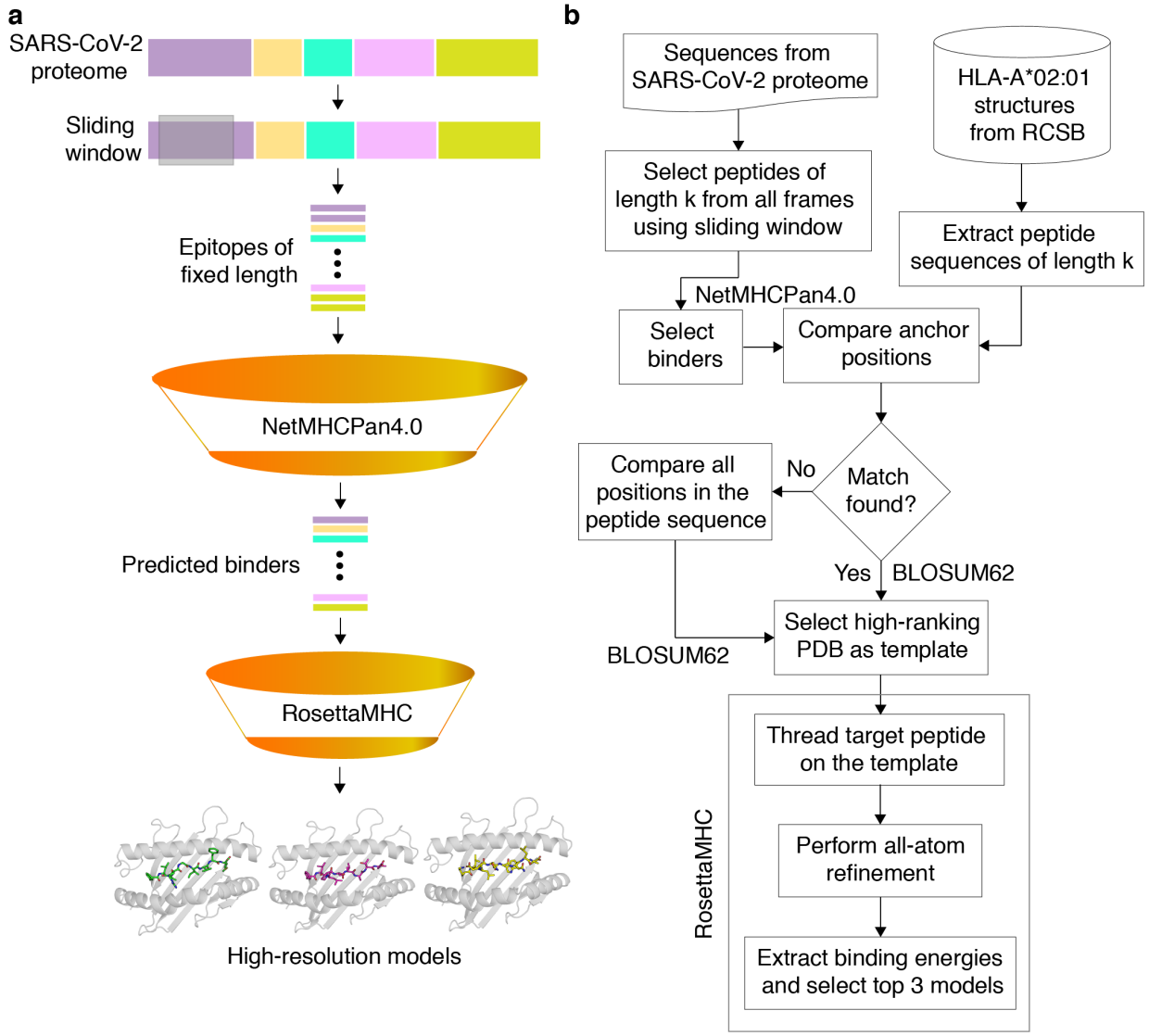

FIGURE 1. Structure-guided modeling of T cell epitopes in the SARS-CoV-2 proteome **(a)** General workflow, outlined in detail in (b). **(b)** Protein sequences from the annotated SARS-CoV-2 proteome are used to generate peptide epitopes with a sliding window covering all frames of a fixed length (9,621 9mer and 9,611 10mer possible peptides). Candidate peptides are first filtered by NetMHCPan4.0 (35) to identify all predicted strong and weak binders (439 9mer and 279 10mer epitopes). For template matching, we use a local database of all 241 HLA-A*02:01 X-ray structures with resolution below 3.5 Å in the Protein Data Bank (22). Each candidate peptide is scanned against all peptide sequences of the same length in the database, and the top-scoring template is used to guide the RosettaMHC comparative modeling protocol.

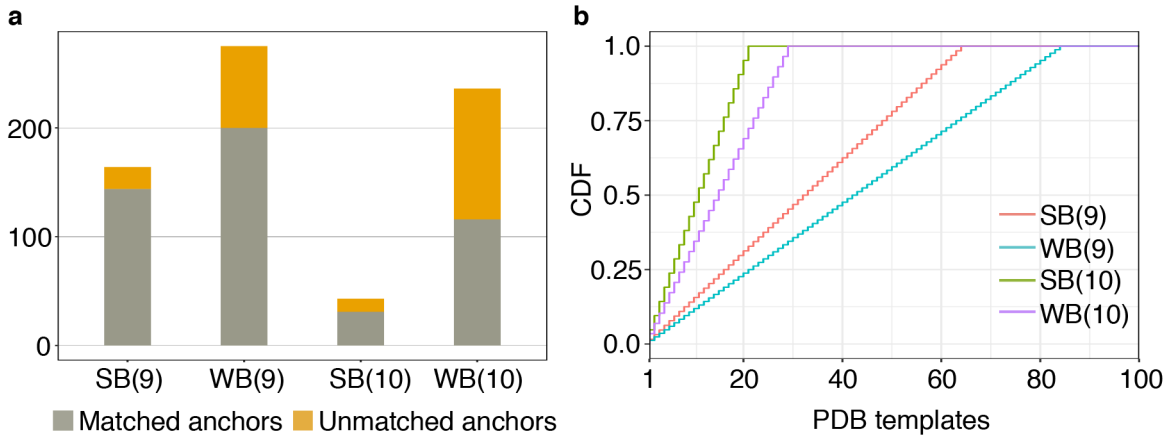

FIGURE 2. Coverage of predicted HLA-A02 epitopes by structural templates in the PDB **(a)** Peptide anchor matching statistics of all predicted SARS-CoV-2 strong (SB) and weak binders (WB) of lengths 9 and 10 to a database of 241 high-resolution HLA-A*02:01 X-ray structures **(b)** Plot showing cumulative distribution (CDF) of strong and weak binder peptides of lengths 9 and 10, as a function of the total number of matching templates from the Protein Data Bank (22).

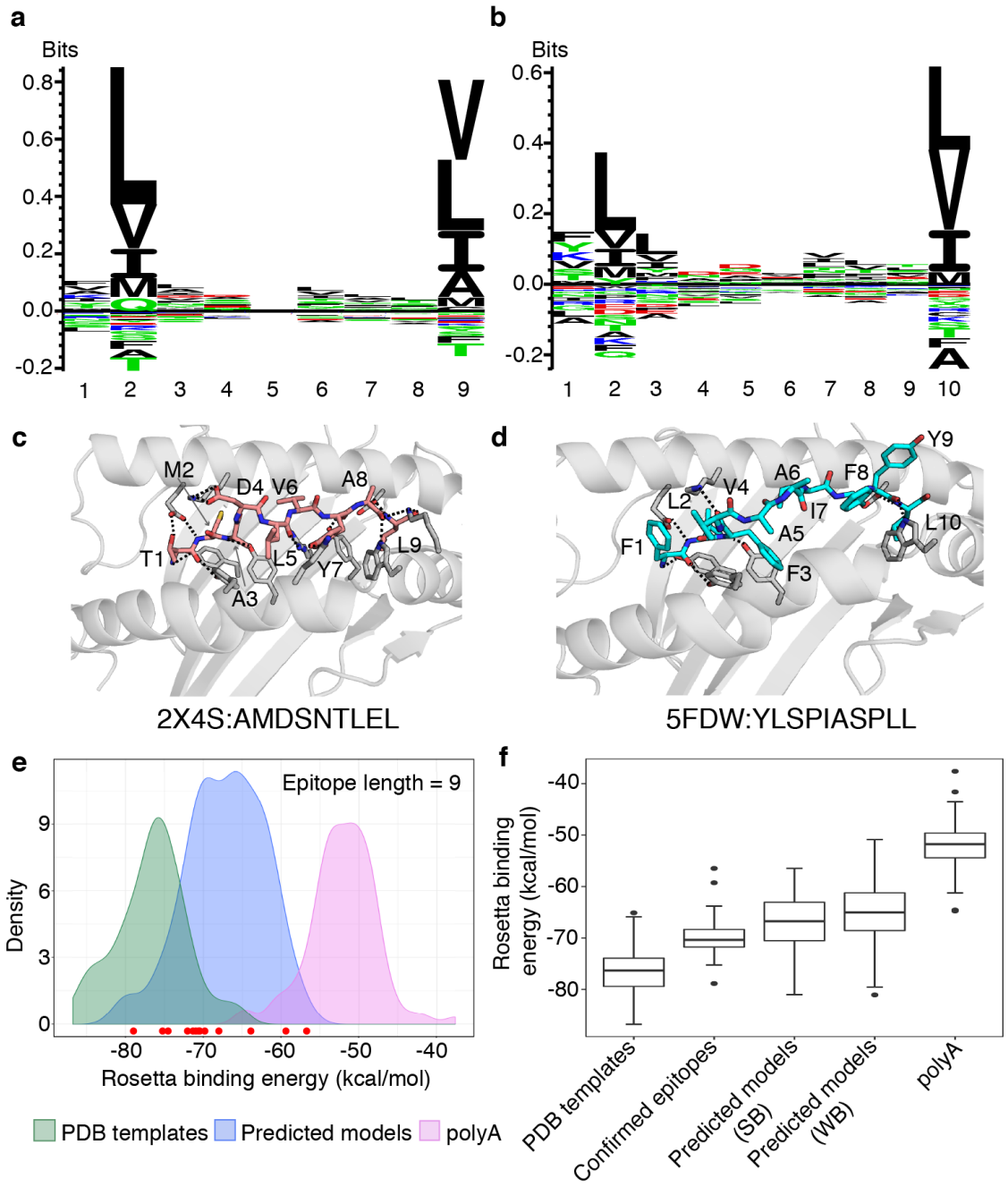

FIGURE 3. Summary of RosettaMHC results for SARS-CoV-2 peptide epitopes

Sequence logos for all HLA-A*02:01-restricted epitopes in the SARS-CoV-2 genome, predicted by NetMHCPan4.0 (35) are shown for epitopes of lengths **(a)** 9 and **(b)** 10. Top-scoring models with epitopes of lengths **(c)** 9 (TMADLVYAL, from RNA polymerase) and **(d)** 10 (FLFVAAIFYL, from nsp3). Dotted lines indicate polar contacts between the peptide and heavy chain residues, with peptide residues labelled. The template PDB IDs and original peptides used for modeling the targets are indicated below each model. **(e)** Density plots showing distribution of average Rosetta binding energies (kcal/mol) for epitopes of length 9. Distributions reflect 93 PDB templates (green), 164 strong binding epitopes (blue), and 93 poly alanine peptides, used as

negative binders (polyA; pink). Confirmed SARS T cell epitopes from the IEDB (37) are indicated by red circles at the bottom of the plot. **(f)** Box plots showing distribution of average binding energies for PDB templates, poly alanine peptides, confirmed epitopes and RosettaMHC models for strong (SB) and weak (WB) binder 9mer epitopes predicted from the SARS-CoV-2 proteome using NetMHCPan4.0 (35).

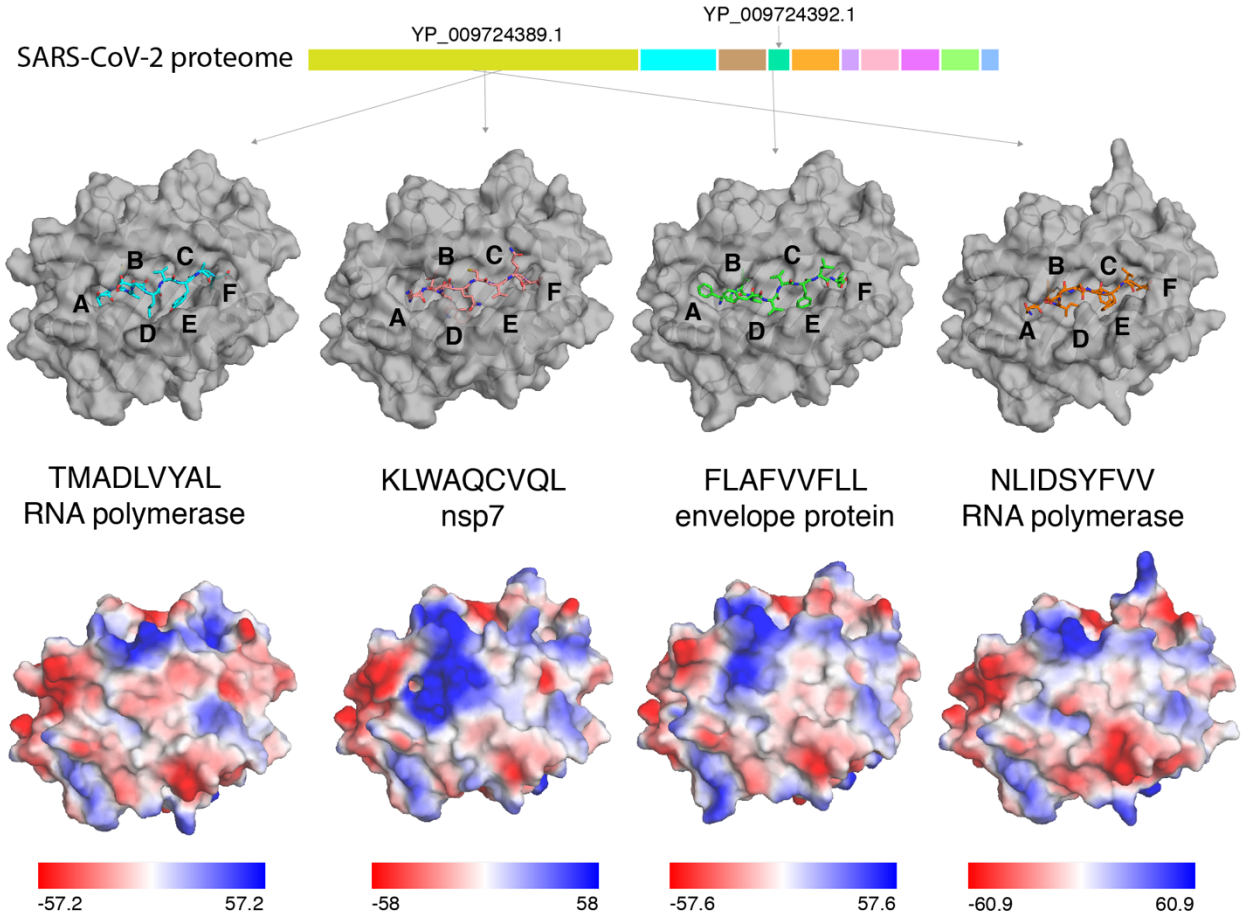

FIGURE 4. Variability in TCR recognition features of HLA-A02 with different high-affinity peptides. Molecular surfaces of SARS-CoV-2/HLA-A*02:01 RosettaMHC models are shown for four top-scoring epitopes (ranked by Rosetta binding energy from left to right) captured in the A, B, C, D, E and F pockets of the MHC-I groove (top panel). The origins of the peptide epitopes in the ~30 kbp SARS-CoV-2 genome are noted. Electrostatic surfaces of the same top scoring models (from top panel) are shown in the bottom panel. Solvent-accessible surface representation with electrostatic potential in the indicated ranges (down to −60 kcal/(mol·*e*) in red and up to +61 kcal/(mol·*e*) in blue) were calculated using the APBS solver (44) in Pymol (45). All calculations were performed at 150 mM ionic strength, 298.15 Kelvin, pH 7.2, protein dielectric 2.0, and solvent dielectric 78.54. Electrostatic potentials are given in units of kT/e. A 1.4 Å solvent (probe) radius and 10.0 points/$Å^2$ density was used to calculate molecular surfaces.

285 **Code and Data availability**
286 An online web-interface for visualization and download of all models is available at:
287 https://rosettamhc.chemistry.ucsc.edu. The RosettaMHC source code is available at
288 https://github.com/snerligit/mhc-pep-threader.Rosetta binding energies for all 718 HLA-
289 A*02:01-restricted peptides in our set are provided in Supplemental Table 1.
290
291 **Acknowledgements**
292 The authors are grateful to Alison Lindberg and ITS/ADC staff at UCSC for assistance in setting
293 up the web server for the database, and Hiram Clawson (UCSC Genome Browser) for providing
294 the SARS-CoV-2 protein sequence data. We thank Andrew McShan, Hailey Wallace (UCSC) for
295 assistance in the development of the RosettaMHC protocol, and David Haussler (UCSC), Michael
296 Betts (University of Pennsylvania) for helpful discussions. This work was supported through
297 Grants from NIAID (5R01AI143997) and NIGMS (5R35GM125034).
298
299 **Disclosures**
300 The authors have no financial conflicts of interest.
301
302 **References**
303 1. Liu, C., Q. Zhou, Y. Li, L. V. Garner, S. P. Watkins, L. J. Carter, J. Smoot, A. C. Gregg, A. D.
304 Daniels, S. Jervey, and D. Albaiu. 2020. Research and Development on Therapeutic Agents and
305 Vaccines for COVID-19 and Related Human Coronavirus Diseases. *ACS Cent. Sci.* .
306 2. Kumaki, Y., J. Ennis, R. Rahbar, J. D. Turner, M. K. Wandersee, A. J. Smith, K. W. Bailey, Z.
307 G. Vest, J. R. Madsen, J. K.-K. Li, and D. L. Barnard. 2011. Single-dose intranasal
308 administration with mDEF201 (adenovirus vectored mouse interferon-alpha) confers protection
309 from mortality in a lethal SARS-CoV BALB/c mouse model. *Antiviral Res.* 89: 75–82.
310 3. Kishimoto, T. 2006. Interleukin-6: discovery of a pleiotropic cytokine. *Arthritis Res. Ther.* 8:
311 S2.
312 4. Rock, K. L., E. Reits, and J. Neefjes. 2016. Present Yourself! By MHC Class I and MHC
313 Class II Molecules. *Trends Immunol.* 37: 724–737.
314 5. Kaufman, J. 2018. Unfinished Business: Evolution of the MHC and the Adaptive Immune
315 System of Jawed Vertebrates. *Annu. Rev. Immunol.* 36: 383–409.
316 6. Thevarajan, I., T. H. O. Nguyen, M. Koutsakos, J. Druce, L. Caly, C. E. van de Sandt, X. Jia,
317 S. Nicholson, M. Catton, B. Cowie, S. Y. C. Tong, S. R. Lewin, and K. Kedzierska. 2020.
318 Breadth of concomitant immune responses prior to patient recovery: a case report of non-severe
319 COVID-19. *Nat. Med.* 1–3.
320 7. Goronzy, J. J., F. Fang, M. M. Cavanagh, Q. Qi, and C. M. Weyand. 2015. Naïve T cell
321 maintenance and function in human aging. *J. Immunol. Baltim. Md 1950* 194: 4073–4080.
322 8. Oyarzun, P., and B. Kobe. 2015. Computer-aided design of T-cell epitope-based vaccines:
323 addressing population coverage. *Int. J. Immunogenet.* 42: 313–321.
324 9. Birnbaum, M. E., J. L. Mendoza, D. K. Sethi, S. Dong, J. Glanville, J. Dobbins, E. Ozkan, M.
325 M. Davis, K. W. Wucherpfennig, and K. C. Garcia. 2014. Deconstructing the peptide-MHC
326 specificity of T cell recognition. *Cell* 157: 1073–1087.
327 10. Wooldridge, L., J. Ekeruche-Makinde, H. A. van den Berg, A. Skowera, J. J. Miles, M. P.
328 Tan, G. Dolton, M. Clement, S. Llewellyn-Lacey, D. A. Price, M. Peakman, and A. K. Sewell.

329 2012. A single autoimmune T cell receptor recognizes more than a million different peptides. *J.*
330 *Biol. Chem.* 287: 1168–1177.
331 11. Peters, B., M. Nielsen, and A. Sette. 2020. T Cell Epitope Predictions. *Annu. Rev. Immunol.* .
332 12. Vita, R., S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler,
333 A. Sette, and B. Peters. 2019. The Immune Epitope Database (IEDB): 2018 update. *Nucleic*
334 *Acids Res.* 47: D339–D343.
335 13. Antunes, D. A., J. R. Abella, D. Devaurs, M. M. Rigo, and L. E. Kavraki. 2018. Structure-
336 based Methods for Binding Mode and Binding Affinity Prediction for Peptide-MHC Complexes.
337 *Curr. Top. Med. Chem.* 18: 2239–2255.
338 14. Yanover, C., and P. Bradley. 2011. Large-scale characterization of peptide-MHC binding
339 landscapes with structural simulations. *Proc. Natl. Acad. Sci.* 108: 6981–6986.
340 15. King, C., E. N. Garza, R. Mazor, J. L. Linehan, I. Pastan, M. Pepper, and D. Baker. 2014.
341 Removing T-cell epitopes with computational protein design. *Proc. Natl. Acad. Sci. U. S. A.* 111:
342 8577–8582.
343 16. Liu, T., X. Pan, L. Chao, W. Tan, S. Qu, L. Yang, B. Wang, and H. Mei. 2014. Subangstrom
344 accuracy in pHLA-I modeling by Rosetta FlexPepDock refinement protocol. *J. Chem. Inf.*
345 *Model.* 54: 2233–2242.
346 17. Kyeong, H.-H., Y. Choi, and H.-S. Kim. 2018. GradDock: rapid simulation and tailored
347 ranking functions for peptide-MHC Class I docking. *Bioinformatics* 34: 469–476.
348 18. Rubenstein, A. B., M. A. Pethe, and S. D. Khare. 2017. MFPred: Rapid and accurate
349 prediction of protein-peptide recognition multispecificity using self-consistent mean field theory.
350 *PLoS Comput. Biol.* 13: e1005614.
351 19. Song, Y., F. DiMaio, R. Y.-R. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, and D.
352 Baker. 2013. High-Resolution Comparative Modeling with RosettaCM. *Structure* 21: 1735–
353 1742.
354 20. Frontiers | A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope
355 Conformations | Immunology. .
356 21. Robinson, J., L. A. Guethlein, N. Cereb, S. Y. Yang, P. J. Norman, S. G. E. Marsh, and P.
357 Parham. 2017. Distinguishing functional polymorphism from random variation in the sequences
358 of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.* 13.
359 22. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N.
360 Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28: 235–242.
361 23. Bentzen, A. K., A. M. Marquard, R. Lyngaa, S. K. Saini, S. Ramskov, M. Donia, L. Such, A.
362 J. S. Furness, N. McGranahan, R. Rosenthal, P. T. Straten, Z. Szallasi, I. M. Svane, C. Swanton,
363 S. A. Quezada, S. N. Jakobsen, A. C. Eklund, and S. R. Hadrup. 2016. Large-scale detection of
364 antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat.*
365 *Biotechnol.* 34: 1037–1045.
366 24. Saini, S. K., T. Tamhane, R. Anjanappa, A. Saikia, S. Ramskov, M. Donia, I. M. Svane, S. N.
367 Jakobsen, M. Garcia-Alai, M. Zacharias, R. Meijers, S. Springer, and S. R. Hadrup. 2019. Empty
368 peptide-receptive MHC class I molecules for efficient detection of antigen-specific T cells. *Sci.*
369 *Immunol.* 4.
370 25. Overall, S. A., J. S. Toor, S. Hao, M. Yarmarkovich, S. M. O'Rourke, G. I. Morozov, S.
371 Nguyen, A. S. Japp, N. Gonzalez, D. Moschidi, M. R. Betts, J. M. Maris, P. Smibert, and N. G.
372 Sgourakis. High Throughput pMHC-I Tetramer Library Production Using Chaperone-Mediated
373 Peptide Exchange. *Nat. Commun. Press* .

26. Ishizuka, J., K. Grebe, E. Shenderov, B. Peters, Q. Chen, Y. Peng, L. Wang, T. Dong, V. Pasquetto, C. Oseroff, J. Sidney, H. Hickman, V. Cerundolo, A. Sette, J. R. Bennink, A. McMichael, and J. W. Yewdell. 2009. Quantitating T Cell Cross-Reactivity for Unrelated Peptide Antigens. *J. Immunol.* 183: 4337–4345.
27. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* 89: 10915–10919.
28. Chaudhury, S., S. Lyskov, and J. J. Gray. 2010. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* 26: 689–691.
29. Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7: 539.
30. Tyka, M. D., D. A. Keedy, I. André, F. Dimaio, Y. Song, D. C. Richardson, J. S. Richardson, and D. Baker. 2011. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* 405: 607–618.
31. Wu, F., S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579: 265–269.
32. Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and and D. Haussler. 2002. The Human Genome Browser at UCSC. *Genome Res.* 12: 996–1006.
33. Trolle, T., C. P. McMurtrey, J. Sidney, W. Bardet, S. C. Osborn, T. Kaever, A. Sette, W. H. Hildebrand, M. Nielsen, and B. Peters. 2016. The length distribution of class I restricted T cell epitopes is determined by both peptide supply and MHC allele specific binding preference. *J. Immunol. Baltim. Md 1950* 196: 1480–1487.
34. Mishto, M., and J. Liepe. 2017. Post-Translational Peptide Splicing and T Cell Responses. *Trends Immunol.* 38: 904–915.
35. Jurtz, V., S. Paul, M. Andreatta, P. Marcatili, B. Peters, and M. Nielsen. 2017. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol. Baltim. Md 1950* 199: 3360–3368.
36. Alford, R. F., A. Leaver-Fay, J. R. Jeliazkov, M. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. DunbrackJr., R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray. 2017. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* 13: 3031–3048.
37. Grifoni, A., J. Sidney, Y. Zhang, R. H. Scheuermann, B. Peters, and A. Sette. 2020. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* .
38. Rossjohn, J., S. Gras, J. J. Miles, S. J. Turner, D. I. Godfrey, and J. McCluskey. 2015. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annu. Rev. Immunol.* 33: 169–200.
39. Rudolph, M. G., R. L. Stanfield, and I. A. Wilson. 2006. How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* 24: 419–466.
40. Borrman, T., J. Cimons, M. Cosiano, M. Purcaro, B. G. Pierce, B. M. Baker, and Z. Weng. 2017. ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins Struct. Funct. Bioinforma.* 85: 908–916.

419 41. Riley, T. P., G. L. J. Keller, A. R. Smith, L. M. Davancaze, A. G. Arbuiso, J. R. Devlin, and
420 B. M. Baker. 2019. Structure Based Prediction of Neoantigen Immunogenicity. *Front. Immunol.*
421 10.
422 42. Mallon, D. H., C. Kling, M. Robb, E. Ellinghaus, J. A. Bradley, C. J. Taylor, D. Kabelitz,
423 and V. Kosmoliaptsis. 2018. Predicting Humoral Alloimmunity from Differences in Donor and
424 Recipient HLA Surface Electrostatic Potential. *J. Immunol.* 201: 3780–3792.
425 43. Baker, N. A., D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. 2001. Electrostatics of
426 nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* 98:
427 10037–10041.
428 44. Jurrus, E., D. Engel, K. Star, K. Monson, J. Brandi, L. E. Felberg, D. H. Brookes, L. Wilson,
429 J. Chen, K. Liles, M. Chun, P. Li, D. W. Gohara, T. Dolinsky, R. Konecny, D. R. Koes, J. E.
430 Nielsen, T. Head-Gordon, W. Geng, R. Krasny, G.-W. Wei, M. J. Holst, J. A. McCammon, and
431 N. A. Baker. 2018. Improvements to the APBS biomolecular solvation software suite. *Protein*
432 *Sci.* 27: 112–128.
433 45. *The PyMOL Molecular Graphics System*,. Schrödinger, LLC.
434