



## NRC Publications Archive Archives des publications du CNRC

### **Assessment of solvated interaction energy function for ranking antibody–antigen binding affinities**

Sulea, Traian; Vivcharuk, Victor; Corbeil, Christopher R.; Deprez, Christophe; Purisima, Enrico O.

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

#### **Publisher's version / Version de l'éditeur:**

<https://doi.org/10.1021/acs.jcim.6b00043>

*Journal of Chemical Information and Modeling*, 56, 7, pp. 1292-1303, 2016-07-01

#### **NRC Publications Record / Notice d'Archives des publications de CNRC:**

<https://nrc-publications.canada.ca/eng/view/object/?id=843145f1-2034-449e-bf44-6c38d5296be9>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=843145f1-2034-449e-bf44-6c38d5296be9>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

#### **Questions?** Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

**Vous avez des questions?** Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.



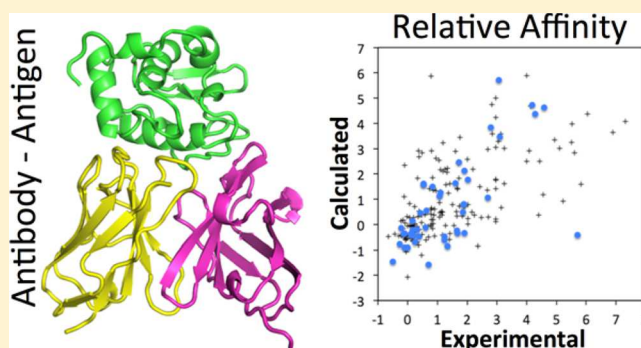
# Assessment of Solvated Interaction Energy Function for Ranking Antibody–Antigen Binding Affinities

Traian Sulea, Victor Vivcharuk,<sup>†</sup> Christopher R. Corbeil, Christophe Deprez, and Enrico O. Purisima\*

Human Health Therapeutics, National Research Council Canada, 6100 Royalmount Avenue, Montreal, QC, Canada H4P 2R2

**S** Supporting Information

**ABSTRACT:** Affinity modulation of antibodies and antibody fragments of therapeutic value is often required in order to improve their clinical efficacies. Virtual affinity maturation has the potential to quickly focus on the critical hotspot residues without the combinatorial explosion problem of conventional display and library approaches. However, this requires a binding affinity scoring function that is capable of ranking single-point mutations of a starting antibody. We focus here on assessing the solvated interaction energy (SIE) function that was originally developed for and is widely applied to scoring of protein–ligand binding affinities. To this end, we assembled a structure–function data set called Single-Point Mutant Antibody Binding (SiPMAB) comprising several antibody–antigen systems suitable for this assessment, i.e., based on high-resolution crystal structures for the parent antibodies and coupled with high-quality binding affinity measurements for sets of single-point antibody mutants in each system. Using this data set, we tested the SIE function with several mutation protocols based on the popular methods SCWRL, Rosetta, and FoldX. We found that the SIE function coupled with a protocol limited to sampling only the mutated side chain can reasonably predict relative binding affinities with a Spearman rank-order correlation coefficient of about 0.6, outperforming more aggressive sampling protocols. Importantly, this performance is maintained for each of the seven system-specific component subsets as well as for other relevant subsets including non-alanine and charge-altering mutations. The transferability and enrichment in affinity-improving mutants can be further enhanced using consensus ranking over multiple methods, including the SIE, Talaris, and FOLDEF energy functions. The knowledge gained from this study can lead to successful prospective applications of virtual affinity maturation.



## INTRODUCTION

The past decade has witnessed an exponential growth of the field of monoclonal antibodies (mAbs) intended for therapeutic use.<sup>1,2</sup> As of 2015, several hundred mAbs were undergoing clinical trials, and about 50 were approved or in review in the European Union or the United States ([https://en.wikipedia.org/wiki/List\\_of\\_therapeutic\\_monoclonal\\_antibodies](https://en.wikipedia.org/wiki/List_of_therapeutic_monoclonal_antibodies)). These molecules are used in humans for treating various diseases and conditions, including not only cancer<sup>3</sup> but also chronic inflammatory diseases, transplantation, and infectious and cardiovascular diseases, among others.<sup>4,5</sup> Recent platforms to improve antibody efficacy and/or reduce side effects include bispecific antibodies and antibody–drug conjugates, both of which have great potential and currently enjoy tremendous interest.<sup>6,7</sup>

Many of these platforms depend on antibodies obtained by animal immunization, e.g., via the hybridoma technique,<sup>8</sup> and often require subsequent optimization.<sup>9,10</sup> Humanization, required for most antibodies from animal sources, typically leads to a reduction in affinity. Other modifications may also be also introduced in order to modulate other properties such as thermal stability, water solubility, pharmacokinetics, etc., which

often results in partial losses of the original antigen-binding affinity. Finally, cross-species specificity is oftentimes required during preclinical studies to enable testing in animal models, which requires restoration or enhancement of antibody affinity for the orthologous antigens found in those species.

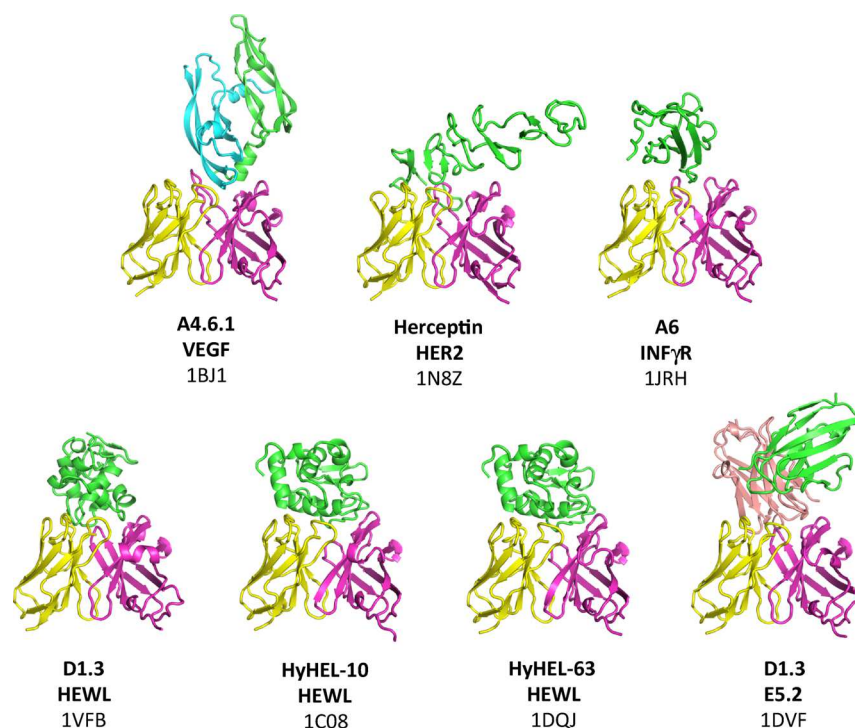
It thus comes as no surprise that several technologies have been devised in order to address antibody affinity-maturation needs. The most widely used are display approaches based on mutant selection on the surface of bacteriophage or yeast<sup>11,12</sup> and screening approaches using synthetic antibody libraries.<sup>13</sup> However, multipoint optimization with these approaches is difficult because of the size of the sequence space that needs to be explored. For example, the theoretical number of all possible four-point mutations in a typical complementarity-determining region (CDR) of an antibody is in the order of  $10^{12}$ , and each additional mutation site incurs another factor of  $10^3$ . This very large space cannot be explored exhaustively, but it can be prioritized. To this end, stepwise virtual affinity maturation has the potential to quickly focus on the critical maturation hotspot

Received: January 29, 2016

Published: July 1, 2016

Table 1. Overview of the SiPMAB Data Set

antibody	antigen	PDB code	resolution (Å)	no. of single-point mutants	binding assay	ref assay	exptl $\Delta\Delta G_{\text{bind}}$ (kcal/mol)		no. of mutants in subsets			
							range	SD	$\Delta Q=0$	$\Delta Q\neq 0$	to-Ala	to-nonAla
A4.6.1	VEGF	1BJ1	2.4	68	SPR, ELISA	28	4.65	1.05	61	7	65	3
D1.3	HEWL	1VFB	1.8	39	SPR, ITC, ELISA	14, 29–31	4.51	1.14	23	16	17	22
HyHEL-10	HEWL	1C08	2.0	33	ITC, spectrophoto	32–35	7.26	1.95	27	6	14	19
herceptin	HER2	1N8Z	2.5	26	ITC	36	6.26	1.73	21	5	21	5
A6	INF $\gamma$ R	1JRH	2.8	19	SPR	37	3.47	0.90	13	6	19	0
D1.3	E5.2	1DVF	1.9	17	SPR	29	3.95	1.24	12	5	16	1
HyHEL-63	HEWL	1DQJ	2.0	10	SPR	38	5.75	1.93	9	1	10	0
totals				212			7.97	1.56	166	46	162	50



**Figure 1.** Overview of parent antibody–antigen complexes for the systems in the SiPMAB data set. Structures are oriented with respect to their antibody Fv fragments, which are rendered in magenta and yellow for the heavy and light chains, respectively. Ribbons rendered in other colors represent the antigen structures. The names of the antibody and antigen and the PDB code of their complex are indicated below each structure.

residues without the sampling limitations of conventional display and library approaches.<sup>14</sup>

In this protocol, given a crystal structure of a parent antibody in complex with its antigen, one begins with an exhaustive computational scan of the antibody's CDR residues in order to predict the most promising single-point mutations that improve the antigen-binding affinity. Typically, up to two dozen single-point mutants are selected for experimental validation, with the confirmed mutations being combined into higher-order mutants. Several cycles of computational predictions interwoven with experimental validation continue up to typically quadruple mutants.

Clearly, the ability to make accurate affinity predictions at the single-point mutation level is key to the success of structure-based stepwise affinity maturation of antibodies, since these mutations are carried over throughout the entire process. Therefore, in order to provide an assessment of the ability to address this critical step, in this study we directed our efforts on two fronts: (1) evaluation of scoring functions for ranking

single-point antibody mutants with companion protocols for mutant structure generation and (2) creation of a suitable structure–affinity data set for carrying out this evaluation.

It is of the utmost importance to have an energy function that can deliver useful predictions leading to affinity-matured antibodies. While much work has been devoted to developing scoring functions for binding affinity predictions,<sup>15,16</sup> typically predictions of protein–protein binding affinities provide a greater challenge than protein–small-molecule binding affinities.<sup>17</sup> This may be due to multiple effects (entropy, strain, solvation, etc.) that may be approximated or even neglected. Another major effect that is typically ignored is protein flexibility. In most mutation studies (including this one), it is assumed that the protein is fairly rigid, and therefore, backbone conformational changes are treated as rare events and ignored. In fact, recent work showed that increasing protein flexibility decreased the accuracy of binding affinity prediction.<sup>18</sup> Even with these deficiencies in current scoring function models, Tidor and Wittrup showed almost a decade ago that it is

Table 2. Protocol Description

protocol			sampling and refinement of mutant structures		affinity ranking		
type	name	ID	bound state	free state	scoring function	Boltzmann ensemble	self-mutation
SIE-SCWRL	SIE-Scwrl <sub>mut</sub>	S1	(1) repack and minimize <b>mutated side chain only</b> ; (2) postminimize around all mutated sites (AMBER)	rigid separation	SIE	no	no
SIE-Rosetta	SIE-Ros <sub>mut</sub>	S2	(1) repack and minimize <b>mutated side chain only</b> ; (2) postminimize around all mutated sites (AMBER)	rigid separation	SIE	no	no
	SIE-Ros <sub>face-sc</sub>	S3	(1) repack and minimize <b>side chains at the Ab–Ag interface</b> ; (2) postminimize around all mutated sites (AMBER)	rigid separation	SIE	yes	yes
	SIE-Ros <sub>CDR-loop</sub>	S4	(1) backrub sample backbone, repack side chains, and minimize <b>CDR loop containing the mutated site</b> ; (2) postminimize around all mutated sites (AMBER)	rigid separation	SIE	yes	yes
Rosetta	Ros <sub>mut</sub>	R1	repack and minimize <b>mutated side chain only</b>	rigid separation	Talaris-interface	no	no
	Ros <sub>face-sc</sub>	R2	repack and minimize <b>side chains at the Ab–Ag interface</b>	repack after separation	Talaris2013	yes	yes
	Ros <sub>CDR-loop</sub>	R3	backrub sample backbone, repack side chains, and minimize the <b>CDR loop containing the mutated site</b>	repack after separation	Talaris2013	yes	yes
FoldX	FoldX <sub>B</sub>	FB	repack <b>side chains around mutated site</b>	rigid separation	FOLDEF-binding	yes	yes
	FoldX <sub>S</sub>	FS	N/A	repack <b>side chains around mutated site</b>	FOLDEF-stability	yes	yes

possible to design mutations leading to affinity-matured antibodies.<sup>14</sup>

Here we focus on the solvated interaction energy (SIE) function that we originally developed for scoring protein–small-molecule-ligand binding affinities.<sup>19,20</sup> The SIE function was extensively tested in several rounds of the CSAR and SAMPL challenges, constantly ranking as one of the best scoring functions.<sup>21–24</sup> Additionally, it has been applied by various laboratories around the world and has been shown to consistently produce reasonable predictions of protein–ligand binding affinities irrespective of the protein target.<sup>25</sup> We also explore several protocols for mutation, given that structural sampling has the potential to greatly affect computational scoring of binding affinities. Alongside SIE we also test other energy functions, such as those provided with the popular FoldX and Rosetta programs, which were originally calibrated on protein stability and protein–protein binding affinity data.<sup>18,26,27</sup>

All of these tests are meaningful only if performed on a suitable data set that is dedicated to the task at hand. To this end, we assembled from the published literature a structure–affinity data set called Single-Point Mutant Antibody Binding (SiPMAB) consisting of 212 single-point mutants of antibodies in their CDRs, curated to use only reliable binding affinity data based on high-quality measurements. A defining characteristic of this unbiased set is its composition based on seven statistically suitable subsets of single-point mutants corresponding to various antibody–antigen systems. These system-based subsets are essential for testing not only the performance within each system but also the transferability across various systems. High-resolution crystal structures of the parent antibodies in complex with their antigens are available in each of these systems. This data set also allows decompositions into various property-based subsets, like non-alanine mutations or charge-altering mutations, which further challenge the ranking performance of the computational protocols being tested.

To increase the usefulness of this data set, not only are the mutations and associated affinities deposited as supplementary data, but the structures generated with the best-performing protocols are made available as well. Our goal in releasing this

database is that the community will use it as a resource for developing the next generation of scoring functions for computer-aided biologics design.

## MATERIALS AND METHODS

**Assembly of the SiPMAB Antibody–Antigen Structure–Affinity Data Set.** The SiPMAB database comprises 212 single-point mutations that belong to seven antibody–antigen systems curated from the available literature.<sup>14,28–38</sup>

The curated binding affinity data are available in Table S1 in the [Supporting Information](#). Overviews of the data set are given in [Table 1](#) and [Figure 1](#). The following criteria were used in order to arrive at this composition of the SiPMAB data set: (i) crystal structure available for the antigen in complex with the parent antibody in each antibody–antigen system; (ii) all single-point mutations located in the antibody CDR loops (Kabat definition); (iii) binding affinity data measured with high-resolution methods, mainly by surface plasmon resonance (SPR) or isothermal titration calorimetry (ITC); (iv) relative binding affinity range and standard deviation (SD) in each antibody–antigen system of at least 3 and 0.6 kcal/mol (1kT), respectively; (v) minimum of 10 single-point mutants in each system.

In a few instances (8% of mutants), several experimental methods were tested on the same mutations, sometimes reported in different publications and from different laboratories. The database records all of these data and classifies them as main data (typically either determined with the higher-accuracy assay or belonging to the largest subset of mutants measured with a given assay) or alternate data. While the absolute binding affinities may vary for the same parent antibody–antigen reference pair, generally there is a good correlation between the relative binding affinities in the main and alternate groups. In cases for which only binding affinity limits are available (5% of mutants), these values were used to calculate relative binding affinities. If crystal structures for mutant–antibody–antigen complexes are available (5% of mutants), their PDB codes are also recorded. Occasionally, when the residue ID numbers of literature and deposited crystal



structures differ, the latter numbers are taken as the primary identifiers and the alternate IDs are recorded in the database.

General structural preparation of the parent structures starting from the crystal structures in the database included removal of water and counterions, deletion of antibody residues after the variable domains of the heavy and light chains, and capping of chain termini as charged states or blocked with neutral groups (acetyl and methylamino) as appropriate. The protonation state at neutral pH was adopted, and polar H atoms at the antibody–antigen interface were visually inspected and manually preoriented to maximize the H-bonding network. (Retrospectively, we found that this manual procedure led to the same results as a much faster automated procedure for systematic scanning and optimization of polar hydrogen positioning that we developed while this work was in progress.<sup>22</sup>) Structural refinement was then carried out by energy minimization using the AMBER force field<sup>39,40</sup> with a distance-dependent dielectric ( $4\epsilon_{ij}$ ) and infinite cutoff for nonbonded interactions, constraining the non-hydrogen atoms at their crystallographic positions with a harmonic force of 5 kcal mol<sup>-1</sup> Å<sup>-2</sup>. Additional refinement was also applied for each sampling–scoring protocol described below.

**Sampling–Scoring Protocols for Mutant Affinity Ranking.** Several mutant structure generation protocols based on the SCWRL and Rosetta methods were employed in order to introduce various degrees of flexibility around the mutated position prior to ranking of mutant binding affinities with the SIE scoring function. For comparative purposes, the Rosetta and FoldX scores (binding affinity and stability) were also obtained using their generated structures. The main characteristics of the sampling–scoring protocols employed in this study are listed in Table 2 and described in detail below. An in-depth description of the terms in the SIE, Rosetta, and FoldX scoring functions can be found in the Appendix in the Supporting Information.

**SIE-SCWRL.** The side-chain repacking program SCWRL version 4.0 (Fox Chase Cancer Center, Philadelphia, PA; <http://dunbrack.fccc.edu/scwrl4>)<sup>41</sup> was used to generate a mutant structure for each complex, and these structures were then scored using the SIE end-point force-field-based method.<sup>19–21,25</sup> In this sampling protocol, termed *Scwrl<sub>mut</sub>*, we adopted the most structurally conservative approach in our study, that is, only the mutated side chain was repacked with SCWRL. Since SIE is a force-field-based method, the resulting SCWRL-generated structures of the mutants were again energy-minimized with the AMBER force field<sup>39,40</sup> to ensure that we are in a local minimum for the SIE scoring function. A set of mobile residues was defined extending 6 Å around the mutated residue, keeping all of the other residues fixed. These refined structures were then used for calculation of the SIE scores in the *SIE-Scwrl<sub>mut</sub>* protocol.

**SIE-Rosetta.** Three protocols employing the Rosetta software version 3.5 (University of Washington, Seattle, WA; <http://www.rosettacommons.org>) were considered. First, the *Ros<sub>mut</sub>* protocol repacks and relaxes only the mutated side chain in a rigid protein environment. Second, the *Ros<sub>iface-sc</sub>* protocol repacks and then relaxes the side chains at the antibody–antigen interface while keeping the backbone rigid. Third, the *Ros<sub>CDR-loop</sub>* protocol performs limited backbone sampling of the CDR loop containing the mutation site using the backrub method,<sup>42</sup> followed by repacking and relaxation of side chains in that loop. The two latter protocols were included to explore the effect of increased protein flexibility during the mutant

structure modeling stage. Before the application of these protocols, the parent complexes were first refined with a Rosetta built-in relaxation routine (Relax) by constraining the backbone and energy-minimizing the side chains freely using the all-atom Talaris2013 scoring function. The XML files of the three structure-generation protocols, *Ros<sub>mut</sub>*, *Ros<sub>iface-sc</sub>*, and *Ros<sub>CDR-loop</sub>*, are provided in the Supporting Information. Conformational ensembles of 10 and 100 replicates were generated for each modeled mutant structure with the *Ros<sub>iface-sc</sub>* and *Ros<sub>CDR-loop</sub>* protocols, respectively, and were also generated in the same way for self-mutations at the mutated positions with these two protocols. The structures resulting from each Rosetta protocol were energy-minimized on the respective sets of mobile residues as described for SIE-SCWRL and then scored using SIE. The final *SIE-Ros<sub>iface-sc</sub>* and *SIE-Ros<sub>CDR-loop</sub>* binding affinity scores were averaged over each resulting conformational ensemble by applying Boltzmann factors derived from the total energies of the complexes obtained with the Talaris2013 function during the Rosetta structural modeling phase of the protocol. No averaging was required in order to calculate the *SIE-Ros<sub>mut</sub>* score, which was based on a unique structure generated by the *Ros<sub>mut</sub>* protocol.

**Rosetta.** For comparative purposes, we also used the structures generated with the three Rosetta protocols described earlier, *Ros<sub>mut</sub>*, *Ros<sub>iface-sc</sub>*, and *Ros<sub>CDR-loop</sub>*, for binding affinity ranking with the Talaris2013 scoring function (eq 2 in the Supporting Information), the default energy function in Rosetta 3.5. As suggested in a virtual alanine scanning mutagenesis protocol that refines only the side chain of the mutated residue in the fixed environment,<sup>18</sup> the *Ros<sub>mut</sub>* protocol employed Talaris-interface, a Talaris energy function with interface weights calibrated specifically for mutations to alanine ([https://guybrush.ucsf.edu/benchmarks/benchmarks/alanine\\_scanning](https://guybrush.ucsf.edu/benchmarks/benchmarks/alanine_scanning)). The other two protocols, *Ros<sub>iface-sc</sub>* and *Ros<sub>CDR-loop</sub>*, used the unscaled Talaris2013 weights for affinity ranking. Binding affinities were calculated as differences in the total energies in the bound and free states (the RosettaScripts XML files are provided in the Supporting Information). It should be noted that the free states were repacked independently before energy calculations, as recommended for the *Ros<sub>iface-sc</sub>* and *Ros<sub>CDR-loop</sub>* protocols with increased protein flexibility but not for the more conservative *Ros<sub>mut</sub>* protocol. Also for the *Ros<sub>iface-sc</sub>* and *Ros<sub>CDR-loop</sub>* protocols, the binding scores were averaged using Boltzmann factors derived from the total energies of the bound states in the conformational ensembles of replicates for mutants and self-mutations, but averaging and self-mutations were not used in the case of *Ros<sub>mut</sub>* (also see Table 2).

**FoldX.** Two types of scores were calculated with FoldX software version 3.0-beta-6 (Center for Genomic Regulation, Barcelona, Spain; <http://foldx.crg.es>),<sup>26</sup> antibody–antigen binding affinity scores and antibody stability scores. The RepairPDB routine was used to refine the structures of the parent antibody–antigen complexes for affinity scoring and also the parent antibody structures extracted from these complexes for stability scoring. Antibody mutants were modeled in the complex state and also in the free state (after antibody extraction from the complex) using the BuildModel routine, with structural flexibility limited to side-chain repacking around the mutated site. The number of replicates was set to 10 for the generation of the mutant structures as well as for self-mutations at each position in order to generate replicates of the parent structures that are compatible with the mutants. The binding

affinity scores, FoldX<sub>B</sub>, for the generated mutant and parent complexes were obtained with the AnalyseComplex routine, whereas the stability scores, FoldX<sub>S</sub>, for the generated mutant and parent antibody structures in the free states were obtained with the Stability routine. Final relative binding and stability scores (FoldX<sub>B</sub> and FoldX<sub>S</sub>, respectively) were calculated as differences between Boltzmann-averaged scores for the mutated structures relative to the corresponding Boltzmann-averaged scores for the parent structures via self-mutations. The Boltzmann factors were derived from the total energies of the complexes or the free antibodies using the Stability routine in both cases. Default settings were used throughout the FoldX routines, except for the ionic strength, which was set to 0.1 M. A strong rotamer penalization due to internal inter-residue clashes was applied in order to account for internal strain and obtain more realistic protein structures (VdWDesign option set to 2). The clash penalty was included in the energy score if it had a positive value.

**Consensus Scoring and Data Analysis.** Prior to being combined into a consensus score, the raw values  $x_i^{\text{SF}_j}$  obtained for mutants  $i$  with individual scoring functions  $\text{SF}_j$  were first normalized into  $Z$  scores:

$$Z_i^{\text{SF}_j} = \frac{x_i^{\text{SF}_j} - x_{\text{median}}^{\text{SF}_j}}{1.4826 \cdot \text{MAD}}$$

where  $x_{\text{median}}^{\text{SF}_j}$  and MAD are the median and the median-absolute-deviation values, respectively. In comparison to the more common  $Z$  scores based on the average and SD values, this median approach is less sensitive to outliers. The consensus score was then calculated as the arithmetic average over the normalized  $Z$  scores obtained with  $n$  scoring functions:

$$\langle Z_i^{\text{SF}_j} \rangle_n = \frac{1}{n} \sum_{j=1}^n Z_i^{\text{SF}_j}$$

In order to reduce the outlier bias, performance is reported using the Spearman rank-order correlation coefficient ( $\rho$ ) between the experimental and calculated data, with a bootstrapped standard deviation over 1000 replicate samples with replacement, as calculated in R.<sup>43</sup> The receiver operating characteristic (ROC) curves and their area under the curve (AUC) calculations for enrichment analysis were also performed in R using 10 000 bootstrapped replicate samples.

## RESULTS AND DISCUSSION

**Single-Point Mutant Antibody Binding (SiPMAB) Structure–Affinity Data Set.** The composition of the SiPMAB database is shown in Table 1 and Figure 1, and details of the curation process are presented in Table S1. Relative experimental and calculated binding affinities are given in Table S2. Overall, the data set contains 212 single-point mutants of antibodies in their CDRs, with curated measured binding affinities relative to parent antibodies having high-resolution crystal structures available in complex with their respective antigens. The measured affinity data span 8 kcal/mol with a standard deviation of 1.6 kcal/mol and were determined mainly by ITC and SPR measurements. It should be noted that in this data set and throughout this study, the terms antibody and binding affinity refer to antibody fragments bearing a single paratope (i.e., Fab or Fv fragments) rather than full-length dimeric antibodies (having two identical paratopes that introduce avidity effects into affinity measurements).

The SiPMAB data set includes seven system-specific subsets, each with at least 10 single-point mutants having measured antigen-binding affinities relative to the parent antibody crystallized in complex with the antigen. These subsets are important in order to test the transferability of the SIE function to different antibody–antigen systems, since good overall performance does not guarantee good rankings within the individual component systems. This follows the philosophy used for the development of the SIE function, which was calibrated on 11 protein targets, each with congeneric sets of small-molecule ligands having measured binding affinities.<sup>19</sup> Hence, in addition to having at least 10 data points in each system, we also imposed the constraint that the measured relative binding affinities in each subset span at least 3 kcal/mol with a standard deviation of at least 1kT (0.6 kcal/mol).

In this study we also wanted to examine the behavior of sampling–scoring protocols against subsets that are most relevant for affinity maturation. Hence, a subset of 50 single-point mutations to residues other than alanine (the “to-nonAla” subset) was created as opposed to mutations to alanine (the “to-Ala” subset). Arguably, predicting structures and affinity changes upon mutations to alanine is easier than for other mutations because of the reduction in size and lack of side-chain conformational flexibility. Additionally, there is a growing consensus that predictions of mutation effects that alter electrostatic interactions are more challenging, possibly because of the noise associated with canceling large opposing electrostatic terms in the adopted molecular mechanics plus solvation formalism.<sup>44</sup> At the same time, computational predictions of enhanced electrostatic interactions were found to lead to validated affinity maturation of antibodies, as opposed to designs via enhanced nonpolar contacts.<sup>14</sup> For these reasons, we examined a subset of 46 single-point mutations that change the net charge of the antibody (the “ $\Delta Q_{\text{mut}} \neq 0$ ” subset) vis-à-vis the subset of mutants that maintain the charge of the parent antibody (the “ $\Delta Q_{\text{mut}} = 0$ ” subset).

The defining characteristics of the SiPMAB data set highlight its unbiased nature. In addition, it does become apparent that the SiPMAB data set serves a particular purpose. The aim here is to assess the suitability of the SIE scoring function, with a companion mutation modeling protocol, for single-point exhaustive screening of the antibody CDR. As mentioned in the Introduction, this is the first and most critical stage of a stepwise virtual affinity maturation pipeline. Hence, SiPMAB differs considerably from other previously published protein–protein structure–affinity data sets. For example, one of the more recent sets used in blind testing of methods for affinity ranking is based on low-resolution experimental data drawn from mutation frequencies.<sup>45</sup> Another set, which has high-resolution affinity data for broad protein–protein interaction systems, lacks subsets of multiple variants (mutants) per system.<sup>17</sup> An earlier set that was used for calibrating a Rosetta energy function basically contains only alanine-scanning mutagenesis data for various protein–protein interfaces.<sup>27</sup> The recent AB-Bind data set, which contains about 1100 mutational data points across 32 complexes, serves a broader purpose and includes non-CDR interfaces and a large proportion of non-SPR/ITC affinity measurements.<sup>46</sup> The closest in spirit is the data set of 173 single-point mutants belonging to seven systems previously employed in enrichment analyses for evaluating the predictive performances of several protocols for virtual affinity maturation. While three of those system-based subsets are similar to those in the SiPMAB

Table 3. Ranking the Performance of Various Individual Sampling–Scoring Protocols and Consensus Scoring<sup>a</sup>

Protocol Name	SIE-Scwrl <sub>mut</sub>	SIE-Ros <sub>mut</sub>	SIE-Ros <sub>iface-sc</sub>	SIE-Ros <sub>CDR-loop</sub>	Ros <sub>mut</sub>	Ros <sub>iface-sc</sub>	Ros <sub>CDR-loop</sub>	FoldX <sub>B</sub>	FoldX <sub>S</sub>	Cons1	Cons2	Cons3	Cons4
Protocol ID	S1	S2	S3	S4	R1	R2	R3	FB	FS	S1+R1	S1+R1+FB	S1+R1+FS	S1+R1+FB+FS
All	0.58	0.58	0.49	0.55	0.60	0.31	0.35	0.49	0.26	0.61	0.63	0.67	0.66
1BJ1	0.52	0.36	0.27	0.48	0.33	0.24	0.22	0.25	0.36	0.53	0.49	0.71	0.61
1VFB	0.74	0.64	0.60	0.65	0.64	0.46	0.54	0.69	0.32	0.74	0.76	0.70	0.72
1C08	0.53	0.50	0.59	0.53	0.51	0.27	0.50	0.38	0.29	0.59	0.51	0.57	0.56
1N8Z	0.53	0.54	0.59	0.38	0.48	0.06	0.41	0.41	0.37	0.53	0.47	0.60	0.58
1JRH	0.54	0.45	0.21	0.28	0.58	0.02	0.20	0.22	0.09	0.48	0.49	0.60	0.51
1DVF	0.47	0.64	0.45	0.59	0.64	0.43	0.37	0.69	0.66	0.62	0.67	0.72	0.74
1DQJ	0.45	0.38	0.12	0.21	0.50	0.78	0.44	0.62	0.24	0.40	0.47	0.76	0.70
to-Ala	0.57	0.62	0.49	0.57	0.65	0.30	0.36	0.50	0.28	0.64	0.66	0.71	0.70
to-nonAla	0.57	0.46	0.54	0.52	0.48	0.34	0.45	0.44	0.26	0.57	0.58	0.57	0.57
$\Delta Q_{mut}=0$	0.58	0.61	0.48	0.59	0.60	0.30	0.31	0.49	0.25	0.60	0.60	0.68	0.68
$\Delta Q_{mut}\neq 0$	0.60	0.56	0.58	0.49	0.63	0.32	0.53	0.51	0.25	0.71	0.67	0.65	0.62
average	0.56	0.53	0.45	0.49	0.55	0.32	0.39	0.47	0.30	0.58	0.58	0.66	0.64
SD	0.07	0.09	0.16	0.13	0.09	0.18	0.11	0.14	0.13	0.09	0.09	0.06	0.07

<sup>a</sup>Values represent Spearman rank-order correlation coefficients,  $\rho$ . Color shading is proportional to the  $\rho$  value. Standard errors from 1000 bootstrapped samples are given in Table S3.

database, the other four subsets are not suitable for the present affinity ranking study. Namely, three subsets have very narrow ranges of measured binding affinities, one subset contains only five mutants, one subset has low-resolution phage-ELISA-based affinity data (with a significant fraction not measured accurately), and another subset is not an antibody–antigen pair.<sup>47</sup> We also believe that including only single-point mutation data at this stage is the most informative way of assessing sampling–scoring protocols for antibody affinity scoring and a stepping stone in moving forward toward multiple-point mutations and de novo antibody designs.

**SIE Scoring Function.** We tested four protocols to model antibody single-point mutations, each followed by SIE affinity ranking. Among these, the SIE-Scwrl<sub>mut</sub> and SIE-Ros<sub>mut</sub> protocols, which repack and relax the mutated side chains and thus are the most structurally conservative ones (see Table 2), perform best on the overall set. The overall performance of the SIE scoring function on this data set of 212 antigen–antibody mutant complexes generated with the SIE-Scwrl<sub>mut</sub> and SIE-Ros<sub>mut</sub> protocols (Spearman rank-order correlation coefficient,  $\rho$ , of 0.58) is comparable with its performance tested on the CSAR-2010 benchmark data set of 343 protein–ligand complexes ( $\rho = 0.62$ ).<sup>21</sup> As can be seen in Table 3, SIE-Scwrl<sub>mut</sub> clearly has the most consistent performance across all of the individual subsets, with no  $\rho$  value below 0.45, whereas SIE-Ros<sub>mut</sub> leads to  $\rho$  values of 0.36 and 0.38 for two of the systems.

Introducing more protein flexibility appears to somewhat degrade the performance and transferability, as evidenced by the results obtained with the SIE-Ros<sub>iface-sc</sub> and SIE-Ros<sub>CDR-loop</sub> protocols (Table 3). From a comparison of these two protocols, local side-chain and backbone sampling of the entire CDR loop containing the mutation site with SIE-Ros<sub>CDR-loop</sub> seems superior in terms of both overall ranking performance and transferability across different systems in comparison to SIE-Ros<sub>iface-sc</sub> that repacks only side chains but over the entire antibody–antigen interface.

Besides examining the overall performance and its consistency across the various system-specific subsets, we were also interested in the performance on the to-nonAla and  $\Delta Q_{mut} \neq 0$  subsets presented earlier, which impose a more stringent test on the sampling–scoring method. It was reassuring to see that the consistent performance of the SIE-

Scwrl<sub>mut</sub> protocol was extended to these challenging subsets as well (Table 3), considering the limited protein flexibility allowed by this protocol. The evenness of the SIE-Scwrl<sub>mut</sub> protocol across the four subsets, to-Ala, to-nonAla,  $\Delta Q_{mut} = 0$ , and  $\Delta Q_{mut} \neq 0$ , is impressive and distinguishes it further from its closest competitor, SIE-Ros<sub>mut</sub>, which shows degraded ranking performance especially for the very important to-nonAla subset. Increasing the conformational sampling flexibility with SIE-Ros<sub>iface-sc</sub> and SIE-Ros<sub>CDR-loop</sub> does improve the performance on mutants to non-alanine residues relative to the more rigid SIE-Ros<sub>mut</sub> approach, but they do not surpass the performance of the SIE-Scwrl<sub>mut</sub> protocol. The performance of the SIE-Scwrl<sub>mut</sub> protocol on the challenging  $\Delta Q_{mut} \neq 0$  subset is also superior or competitive to the other SIE-based protocols tested.

There are several plausible sources for the leading ranking performance of the SIE-Scwrl<sub>mut</sub> protocol. First, the preminimization of the parent structure with the AMBER force field, on which the SIE function is rooted, sets SIE-Scwrl<sub>mut</sub> at an advantage relative to the other SIE protocols based on a Rosetta-compatible parent structure, even if mobile residues are postminimized with AMBER for these Rosetta-based protocols. Second, conformational sampling may differ between SCWRL and Rosetta even for the most similar conservative protocols, SIE-Scwrl<sub>mut</sub> and SIE-Ros<sub>mut</sub>, in which only the mutated side chain is sampled. The degradation of performance with increased flexibility is not a new finding and has been noted before in the context of Rosetta scoring of binding affinities<sup>18</sup> and SIE scoring of protein–ligand binding affinities, where the recommended protocol is to use strong constraints during the energy minimization of the protein–ligand interface.<sup>25</sup> Apparently, the inaccuracies inherent in any classical mechanics force field or empirical energy function are exacerbated when the number of degrees of freedom in the system increases, leading to unreliable structural and energetic predictions. This may lead to poorer performance especially in the case of SIE-Ros<sub>iface-sc</sub> where a large number of side chains are allowed to repack during mutation. An attempt to minimize these effects by performing Boltzmann averaging of scores over conformational ensembles could not offset the noise introduced by the increase in flexibility.

**Other Scoring Functions.** Although the primary focus here is to test the performance of the SIE protein–ligand



binding affinity scoring function in ranking antibody–antigen interactions, we were also interested in positioning the performance of SIE relative to those of other popular functions for scoring protein–protein binding, such as Rosetta's Talaris energy function and the FoldX energy function FOLDEF.

It has been suggested that the weights of the Talaris functional form can be fitted specifically for alanine scanning (mutations to Ala only) and that a protocol based on protein flexibility limited to the mutated site affords the best performance.<sup>18</sup> Hence, we used here the Talaris-interface weights in the Ros<sub>mut</sub> protocol, which repacks and relaxes only the mutated side chain (non-Ala side chains present in the to-nonAla subset of this study). We found that Ros<sub>mut</sub> affords a highly competitive performance with SIE-Ros<sub>mut</sub> and SIE-Scwrl<sub>mut</sub> for affinity ranking on the overall set. It also shows good transferability across various system-specific subsets, somewhat less than SIE-Scwrl<sub>mut</sub> but slightly improved versus SIE-Ros<sub>mut</sub> (see Table 3). The latter behavior is understandable given the fact that the Rosetta-generated structure is more compatible with Rosetta scoring using the Talaris-interface function than with the SIE function. Importantly, the Ros<sub>mut</sub> protocol does best what it has been trained for, namely, scoring mutations to alanine. Its ranking performance on the to-Ala subset ( $\rho = 0.65$ ) is best across all of the tested individual protocols (excluding the consensus ones). However, this is in striking discrepancy to its performance on the to-nonAla subset ( $\rho = 0.48$ ), which is similar to the performance of SIE-Ros<sub>mut</sub> ( $\rho = 0.46$ ) but inferior to that of SIE-Scwrl<sub>mut</sub> ( $\rho = 0.57$ ). The Ros<sub>mut</sub> protocol, however, shows the best performance on the other challenging subset,  $\Delta Q_{mut} \neq 0$ , with  $\rho = 0.63$  slightly surpassing that of SIE-Scwrl<sub>mut</sub> ( $\rho = 0.60$ ).

As expected, the performance of the structurally conservative protocol Ros<sub>mut</sub> degraded with increased protein flexibility, as evidenced by the data obtained with the Ros<sub>iface-sc</sub> and Ros<sub>CDR-loop</sub> protocols described earlier (Tables 2 and 3). This parallels the trend seen with the protocols involving SIE-based ranking, although the degradation in performance among the Rosetta protocols involving Talaris-based ranking is more pronounced. This is in part due to the fact that the flexible protocols Ros<sub>iface-sc</sub> and Ros<sub>CDR-loop</sub> were matched with Rosetta's default Talaris2013 energy function instead of Talaris-interface weights (recommended for rigid protocols and alanine-scanning mutagenesis), which is appropriately employed in Ros<sub>mut</sub>.

One can conclude on the basis of the results so far that the SIE-Scwrl<sub>mut</sub> protocol is competitive with the standard Ros<sub>mut</sub> protocol for virtual alanine-scanning mutagenesis but outperforms it for ranking mutations to residues other than alanine. The similar ranking performances obtained with the Ros<sub>mut</sub> and SIE-Ros<sub>mut</sub> protocols (especially the split into to-Ala and to-nonAla subsets), which are based on similar structures but different scoring functions, may indicate that the choice of scoring function (e.g., SIE vs Talaris-interface) is not that critical. This is more complicated, however, since the distribution of calculated scores with the two functions shows a much larger spread (Figure S1) than what may be anticipated by comparing the rank-order correlation coefficients to experiment (Table 3). Two points can be made here. First, unique combinations of sampling protocols and scoring functions can provide improved ranking performance and system transferability, and SIE-Scwrl<sub>mut</sub> appears to be one such winning combination. Second, it is quite impressive that a scoring function like SIE with only four energy terms and a

single global weighting factor can match and even outperform a much more complex functional form like Talaris, which has 16 energy terms, each weighted independently. It is possible that the accuracy of the continuum solvation model used by SIE is responsible for its performance.

In the FoldX<sub>B</sub> protocol used here for binding affinity ranking, we found that its FOLDEF scoring function including repulsive clash terms could rank antigen–antibody binding affinities with an overall  $\rho$  of 0.49, which is inferior to the more conservative SIE-Scwrl<sub>mut</sub> and Ros<sub>mut</sub> protocols (Table 3). However, there were a few particular system-specific subsets (1DVF, 1DQJ) for which FOLDEF was superior to both SIE-Scwrl<sub>mut</sub> and Ros<sub>mut</sub>. In terms of sampling, FoldX repacks side chains around the mutation site, and thus, it is most similar to the flexible Rosetta protocols, particularly Ros<sub>iface-sc</sub>, although the extent of the sampled set of residues as well as the sampling implementation differ. Together with differences between the various energy terms of the FOLDEF and SIE functions, it is perhaps not that surprising that although the overall performance of FoldX<sub>B</sub> and SIE-Ros<sub>iface-sc</sub> are similar (each affording  $\rho = 0.49$  with respect to experiment), quite a different behavior is seen for several system-specific subsets (FoldX<sub>B</sub> is better on 1VFB, 1DVF, and 1DQJ, and SIE-Ros<sub>iface-sc</sub> is better on 1C08 and 1N8Z).

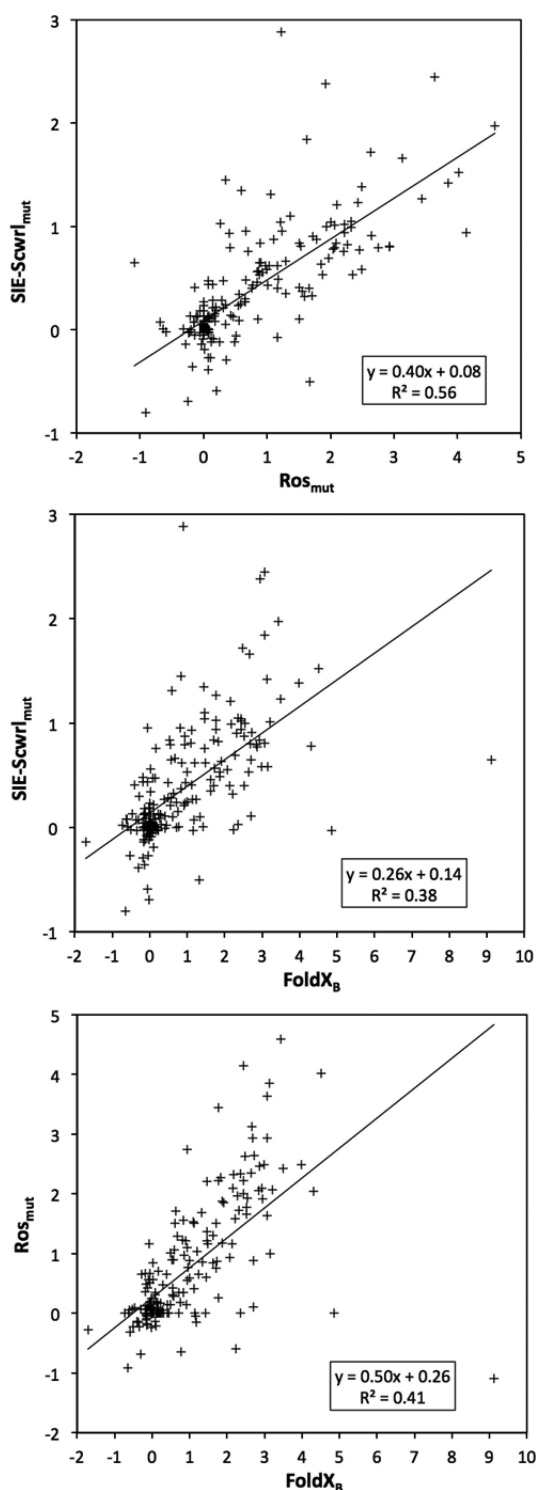
Finally, it was interesting to note that the stabilities of antibody mutants relative to the parent antibodies seem to rank-correlate, albeit weakly, with the measured antigen binding affinities (Table 3, FoldX<sub>S</sub> protocol). Although antibody side chains were repacked in the free state prior to stability calculations, the FoldX<sub>S</sub> stability scores reflect the folding process from the unfolded state all the way to the bound state for the backbone conformation, which may not be the lowest-energy state of an unbound antibody mutant. Hence, the FoldX<sub>S</sub> scores include some of the conformational penalty of transitioning between the unbound and bound states, which is a strain energy cost directly related to binding affinity.

**Consensus Scoring.** Previous experience in the field of protein structure and binding affinity predictions indicates that prediction accuracy can be improved by applying a consensus approach over several methods.<sup>48–51</sup> Hence, here we sought to apply the same strategy for ranking antigen–antibody mutant binding affinities, since reasonable ranking performance was afforded by several protocols and scoring functions with subtle differences on several subsets. Because individual scoring functions can lead to different score magnitudes, scales, and dynamic ranges, relative ranks are often used instead of absolute binding scores. Since the rank alone lacks information about the data spread, we instead used Z scores for the normalization of raw scores.<sup>52</sup> In addition, our Z scores are based on the median value instead of the average value for each scoring function, which reduces the sensitivity of Z scores to outliers.

In Table 3 we show the performance of four possible consensus functions, termed Cons1 through Cons4, which are arithmetic averages of the Z scores obtained with various individual scoring protocols. The strategy was to use the best-performing protocols based on the SIE, Talaris, and FOLDEF energy functions, that is, SIE-Scwrl<sub>mut</sub>, Ros<sub>mut</sub>, and FoldX<sub>B</sub>. As shown in Figure 2, there is a reasonable correlation between the raw scores obtained with these protocols, which is required for building successful consensus functions,<sup>53</sup> but there is also sufficient variation that further warrants exploration of the consensus approach.

The Cons1 function includes the Z scores from SIE-Scwrl<sub>mut</sub> and Ros<sub>mut</sub>, the leading protocols based on the SIE and Talaris





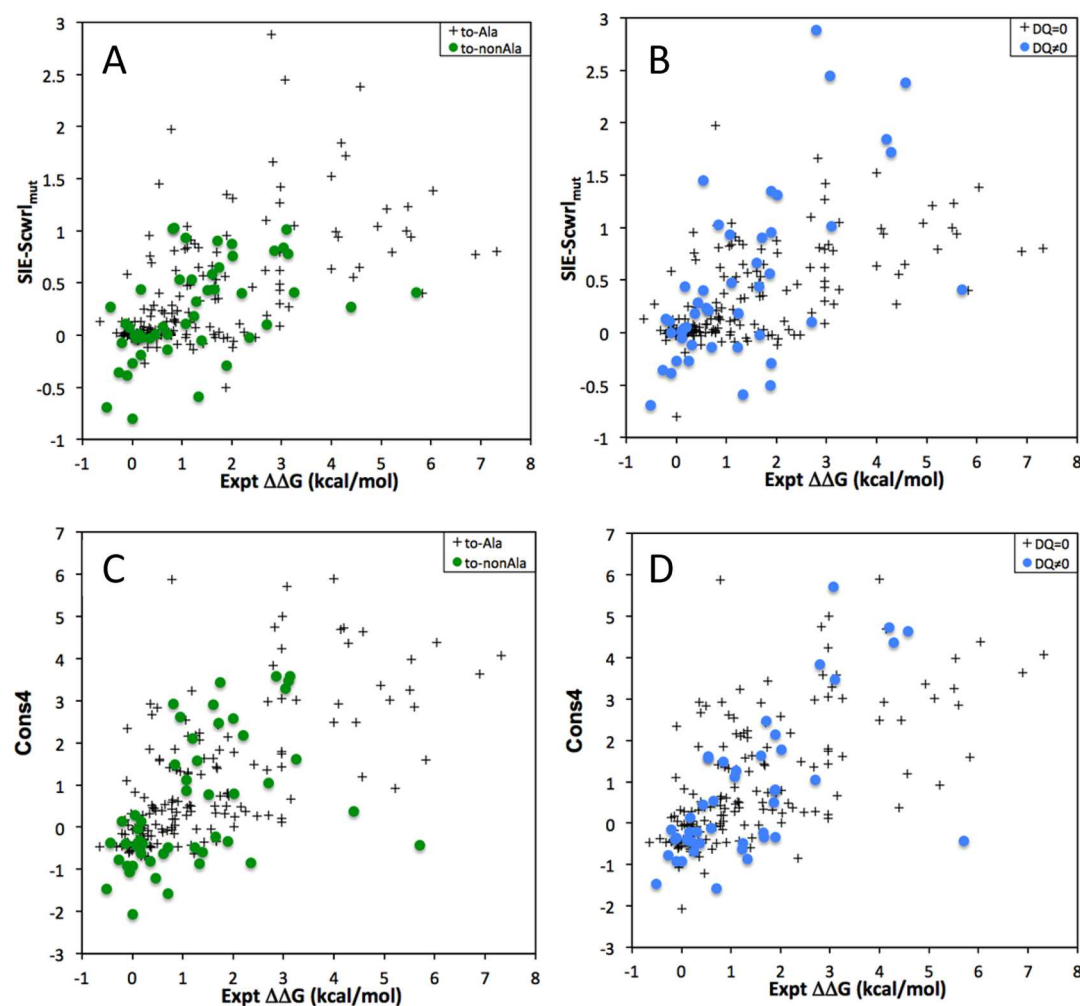
**Figure 2.** Correlations between the raw scores obtained with the three scoring functions analyzed in this study. The protocols that performed best in ranking experimental relative binding affinities are used: SIE-Scwrl<sub>mut</sub> for the SIE function, Ros<sub>mut</sub> for the Talaris function, and FoldX<sub>B</sub> for the FOLDEF function. The least-squares correlation line is indicated and the associated linear equation and squared correlation coefficient are provided for each pair.

scoring functions. Although Cons1 does not significantly improve the overall performance, it combines the best behaviors of its component functions, namely, it retains the best performance obtained with SIE-Scwrl<sub>mut</sub> on the important

to-nonAla subset without compromising the excellent performance of Ros<sub>mut</sub> in ranking mutations to alanine. The result is a consistently good performance across all four property-based subsets (to-Ala, to-nonAla,  $\Delta Q_{\text{mut}} = 0$  and  $\Delta Q_{\text{mut}} \neq 0$ ). Cons2 is a three-function Z-score-based consensus that builds upon Cons1 and introduces the FOLDEF scoring function via the FoldX<sub>B</sub> protocol. The overall performance is again only marginally increased, but there is some improvement relative to Cons1 on the system-specific subsets where FoldX<sub>B</sub> performed best (1DVF and 1DQJ) while maintaining the consistent performance on the four property-specific subsets.

Finally, we added to Cons1 and Cons2 the Z scores from antibody stability calculations with the FoldX<sub>S</sub> protocol. This was based on the observation that although by itself the FoldX<sub>S</sub> antibody stability score correlates only weakly with antibody–antigen binding affinity, it has the surprising ability to consistently improve the affinity ranking performance of the other individual protocols (see Figure S2). The resulting Cons3 and Cons4 consensus functions outperform the best individual methods SIE-Scwrl<sub>mut</sub> and Ros<sub>mut</sub> as well as the consensus methods Cons1 and Cons2 overall and on the vast majority of system-based and property-based subsets. Upon inclusion of FoldX<sub>S</sub> into the consensus, there is a tendency to improve the performance for the to-Ala subset while maintaining the performance on the to-nonAla subset and also to improve the performance on the  $\Delta Q_{\text{mut}} = 0$  subset and decrease it on the  $\Delta Q_{\text{mut}} \neq 0$  subset (still above the level obtained with individual protocols). We note that the FOLDEF energy function was trained on folding free energy data for mutations to alanine.<sup>26</sup> Although the improvements are marginal in some of the cases (also see the bootstrapping standard errors in Table S3), they are very consistent across the various subsets, as evidenced by the increase in the average and the decrease in the standard deviation of  $\rho$  values across all 12 sets (Table 3). This is a notable achievement for the consensus approach, particularly when considering that no training of weights was applied to the component Z scores from individual protocols. A visual comparison of the ranking ability of Cons4 versus SIE-Scwrl<sub>mut</sub> is presented in the scatter plots shown in Figure 3.

A legitimate question concerns the extent to which the predictive power is affected by the limitations of the physical model rather than the scoring function. In an attempt to separate the effects of structural predictions from scoring, we analyzed 11 mutants from the SiPMAB database for which crystal structures are available and which belong to two systems (six mutants from 1C08<sup>32,33</sup> and five mutants from 1VFB<sup>30</sup>). We found that the modeled mutant structures produced by the better-performing protocols for each of the three scoring functions (SIE-Scwrl<sub>mut</sub>, Ros<sub>mut</sub> and FoldX<sub>S</sub>) are very similar to each other and to their crystal structures, indicating that the structural predictions are reasonable. However, there appears to be a marked improvement in the relative ranking of these mutants when a single scoring function (e.g., SIE) is replaced by consensus scoring (Figure S3). This points toward a greater role of the scoring component than the structural modeling component in the performance of the model. Outlier analysis lends further support to this hypothesis by focusing on the subset of mutants that incur charge changes (three Asn-to-Asp mutations from the 1C08 system), which are misranked by SIE (Figure S3A), with two of them appearing as moderate outliers even after consensus scoring (Figure S3B). However, their structural predictions are the closest to their corresponding crystal structures among the analyzed set of 11 crystallized



**Figure 3.** Comparison of the consensus approach vs the best single-method. Shown are scatter plots between experimental relative binding affinities and raw scores obtained with (A, B) the SIE-Scwrl<sub>mut</sub> protocol and (C, D) the Cons4 consensus approach for (A, C) to-Ala/to-nonAla subset decomposition and (B, D)  $\Delta Q_{\text{mut}} = 0/\Delta Q_{\text{mut}} \neq 0$  subset decomposition.

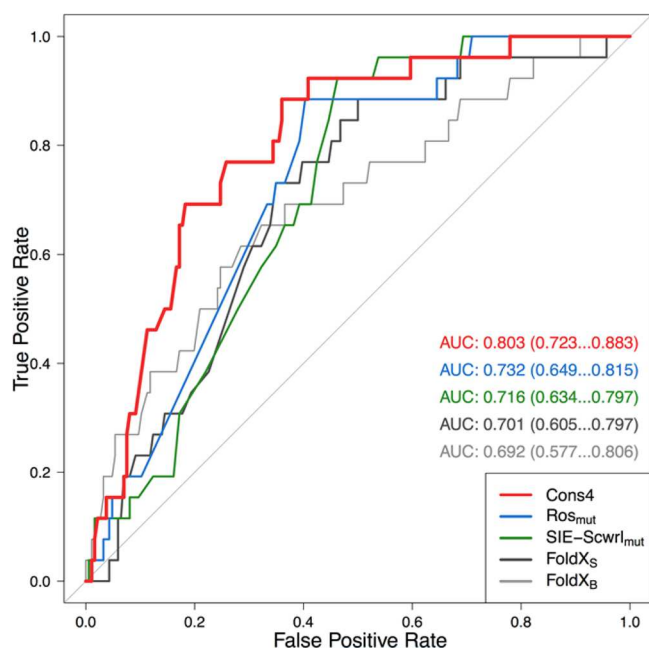
mutants (Figure S3C). Since their mutated sites engage in H-bonding and salt-bridge interactions with the antigen, including bridging-water-mediated contacts,<sup>33</sup> further refinements of the scoring functions may be needed to capture these electrostatic interactions with greater accuracy. Other avenues for further improving scoring functions for protein–protein complexes are also feasible. For example, currently the correlation slope between the SIE scores and experimental affinities is about 0.5 for antibody–antigen complexes (Figure 3), whereas it is close to 1 for small-molecule–ligand–protein complexes.<sup>21,25</sup> Plausible factors for this discrepancy are vibrational and configurational entropies and nonpolar solvation (e.g., different molecular surface curvatures of proteins vs small molecules).

As mentioned at the beginning of the paper, a highly relevant practical application of *in silico* mutagenesis of antibody–antigen interfaces is to guide the affinity maturation of antibodies. With that practical task in view, it is instructive to find out the extent to which the various protocols explored in this study are able to identify mutations that actually improve binding. Available data sets are typically dominated by destabilizing mutations, and the SiPMAB database is no exception, with 26 affinity-improving mutants out of a total of 212 single-point mutants. We analyzed various protocols studied in this paper in terms of their prediction performances

toward enrichments in affinity-improving mutants. These data are shown in Figure 4 as ROC curves and the corresponding AUC values. We see that individual scoring functions with the best-performing protocols for affinity ranking also have reasonable enrichment capabilities, with AUC values of around 0.7 ( $\text{AUC}_{\text{random}} = 0.5$ ,  $\text{AUC}_{\text{ideal}} = 1$ ). The advantage of the consensus approach is also evidenced in this type of analysis, as it gives an AUC of about 0.8, thus outperforming the individual protocols. A relatively robust early enrichment was also obtained with the consensus protocol, that is, several affinity-improving mutants are top-ranked by consensus scoring.

## CONCLUSIONS

We have assessed the suitability of the solvated interaction energy (SIE) function, originally developed for small-molecule protein–ligand binding affinities, for use in antibody–antigen systems. Using a database of 212 single-point mutant antibodies (SiPMAB), we have found that the SIE function together with a very local side-chain rotamer sampling protocol can reasonably predict relative binding affinities with a Spearman rank-order correlation coefficient of about 0.6 across seven antibody–antigen systems. This did not require any reparametrization of the SIE function. We also examined the use of consensus ranking by adding the Rosetta Talaris and FoldX FOLDEF



**Figure 4.** Enrichment analysis of the SiMPAB database using selected protocols applied in this study. ROC curves were calculated using a cutoff of 0.00 kcal/mol in relative binding free energy for classification of mutants as binding-improving (26) and binding-destabilizing (186). Median AUC values and 95% confidence intervals from bootstrapping tests are listed.

scoring functions to the SIE function in a Z-score-normalized way. Consensus ranking only marginally improves the overall Spearman correlation coefficient but does improve the transferability across the different systems and enrichment predictions of mutants with improved binding. These results give us confidence in applying the SIE function to guide the design of antibodies to improve or modulate their binding affinities for their target antigens.

## ■ ASSOCIATED CONTENT

### ■ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00043.

Curated experimental binding affinities in the SiPMAB database (Table S1), calculated binding affinity scores with various protocols (Table S2), and standard errors for Spearman- $\rho$  coefficients from bootstrapping analysis (Table S3) (XLSX)

Appendix describing the various terms in the SIE, FoldX, and Rosetta scoring functions; comparison of raw scores calculated with SIE and Talaris functions on similar Rosetta-generated structures (Figure S1), effect of the FoldX stability score on the performance of individual binding affinity ranking protocols (Figure S2), and outlier analysis of the 11 mutants with available crystal structures (Figure S3); text providing the three Rosetta XML protocols employed in this study (PDF)

Compressed version of the mutant structures generated for 1BJ1 (ZIP)

Compressed version of the mutant structures generated for 1C08 (ZIP)

Compressed version of the mutant structures generated for 1DQJ (ZIP)

Compressed version of the mutant structures generated for 1DVF (ZIP)

Compressed version of the mutant structures generated for 1JRH (ZIP)

Compressed version of the mutant structures generated for 1N8Z (ZIP)

Compressed version of the mutant structures generated for 1VFB (ZIP)

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [Enrico.Purisima@nrc-cnrc.gc.ca](mailto:Enrico.Purisima@nrc-cnrc.gc.ca).

### Present Address

<sup>†</sup>V.V.: Weill Cornell Medical College, Department of Physiology and Biophysics, 1300 York Avenue, New York, NY 10065.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank Hervé Hogues for useful discussions. This is NRC Canada publication number NRC-HHT\_53317. This work was funded in part by a CQDM Explore Grant.

## ■ REFERENCES

- Reichert, J. M. Metrics for Antibody Therapeutics Development. *mAbs* **2010**, *2*, 695–700.
- Reichert, J. M. Antibodies to Watch in 2015. *mAbs* **2015**, *7*, 1–8.
- Reichert, J. M.; Dhimolea, E. The Future of Antibodies as Cancer Drugs. *Drug Discovery Today* **2012**, *17*, 954–963.
- Reichert, J. M.; Rosensweig, C. J.; Faden, L. B.; Dewitz, M. C. Monoclonal Antibody Successes in the Clinic. *Nat. Biotechnol.* **2005**, *23*, 1073–1078.
- Nelson, A. L.; Dhimolea, E.; Reichert, J. M. Development Trends for Human Monoclonal Antibody Therapeutics. *Nat. Rev. Drug Discovery* **2010**, *9*, 767–774.
- Teicher, B. A.; Chari, R. V. Antibody Conjugate Therapeutics: Challenges and Potential. *Clin. Cancer Res.* **2011**, *17*, 6389–6397.
- Byrne, H.; Conroy, P. J.; Whistock, J. C.; O’Kennedy, R. J. A Tale of Two Specificities: Bispecific Antibodies for Therapeutic and Diagnostic Applications. *Trends Biotechnol.* **2013**, *31*, 621–632.
- Little, M.; Kipriyanov, S. M.; Le Gall, F.; Moldenhauer, G. Of Mice and Men: Hybridoma and Recombinant Antibodies. *Immunol. Today* **2000**, *21*, 364–370.
- Gilliland, G. L.; Luo, J.; Vafa, O.; Almagro, J. C. Leveraging SBDD in Protein Therapeutic Development: Antibody Engineering. *Methods Mol. Biol.* **2012**, *841*, 321–349.
- Strohl, W. R.; Strohl, L. M. *Therapeutic Antibody Engineering: Current and Future Advances Driving the Strongest Growth Area in the Pharmaceutical Industry*; Woodhead Publishing: Cambridge, U.K., 2012; pp 111–129, 225–249, 299–328, 377–404.
- Hammers, C. M.; Stanley, J. R. Antibody Phage Display: Technique and Applications. *J. Invest. Dermatol.* **2014**, *134*, e17.
- Boder, E. T.; Raeeszadeh-Sarmazdeh, M.; Price, J. V. Engineering Antibodies by Yeast Display. *Arch. Biochem. Biophys.* **2012**, *526*, 99–106.
- Zhou, H.; Zhang, Y. L.; Lu, G.; Ji, H.; Rodi, C. P. Recombinant Antibody Libraries and Selection Technologies. *New Biotechnol.* **2011**, *28*, 448–452.
- Lippow, S. M.; Wittrup, K. D.; Tidor, B. Computational Design of Antibody-Affinity Improvement Beyond in Vivo Maturation. *Nat. Biotechnol.* **2007**, *25*, 1171–1176.
- Gilson, M. K.; Zhou, H. X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.



- (16) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.
- (17) Kasttritis, P. L.; Bonvin, A. M. J. J. Are Scoring Functions in Protein-Protein Docking Ready to Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *J. Proteome Res.* **2010**, *9*, 2216–2225.
- (18) O Conchuir, S.; Barlow, K. A.; Pache, R. A.; Ollikainen, N.; Kundert, K.; O'Meara, M. J.; Smith, C. A.; Kortemme, T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLoS One* **2015**, *10*, e0130433.
- (19) Naim, M.; Bhat, S.; Rankin, K. N.; Dennis, S.; Chowdhury, S. F.; Siddiqi, I.; Drabik, P.; Sulea, T.; Bayly, C. I.; Jakalian, A.; Purisima, E. O. Solvated Interaction Energy (SIE) for Scoring Protein-Ligand Binding Affinities. 1. Exploring the Parameter Space. *J. Chem. Inf. Model.* **2007**, *47*, 122–133.
- (20) Cui, Q.; Sulea, T.; Schrag, J. D.; Munger, C.; Hung, M. N.; Naim, M.; Cygler, M.; Purisima, E. O. Molecular Dynamics-Solvated Interaction Energy Studies of Protein-Protein Interactions: The Mpl-P14 Scaffolding Complex. *J. Mol. Biol.* **2008**, *379*, 787–802.
- (21) Sulea, T.; Cui, Q.; Purisima, E. O. Solvated Interaction Energy (SIE) for Scoring Protein-Ligand Binding Affinities. 2. Benchmark in the CSAR-2010 Scoring Exercise. *J. Chem. Inf. Model.* **2011**, *51*, 2066–2081.
- (22) Hogues, H.; Sulea, T.; Purisima, E. O. Evaluation of the Wilma-SIE Virtual Screening Method in Community Structure-Activity Resource 2013 and 2014 Blind Challenges. *J. Chem. Inf. Model.* **2016**, *56*, 955–964.
- (23) Sulea, T.; Hogues, H.; Purisima, E. O. Exhaustive Search and Solvated Interaction Energy (SIE) for Virtual Screening and Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 617–633.
- (24) Hogues, H.; Sulea, T.; Purisima, E. O. Exhaustive Docking and Solvated Interaction Energy Scoring: Lessons Learned from the SAMPL4 Challenge. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 417–427.
- (25) Sulea, T.; Purisima, E. O. The Solvated Interaction Energy Method for Scoring Binding Affinities. *Methods Mol. Biol.* **2012**, *819*, 295–303.
- (26) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *J. Mol. Biol.* **2002**, *320*, 369–387.
- (27) Kortemme, T.; Baker, D. A Simple Physical Model for Binding Energy Hot Spots in Protein-Protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 14116–14121.
- (28) Muller, Y. A.; Chen, Y.; Christinger, H. W.; Li, B.; Cunningham, B. C.; Lowman, H. B.; de Vos, A. M. VEGF and the Fab Fragment of a Humanized Neutralizing Antibody: Crystal Structure of the Complex at 2.4 Å Resolution and Mutational Analysis of the Interface. *Structure* **1998**, *6*, 1153–1167.
- (29) Dall'Acqua, W.; Goldman, E. R.; Eisenstein, E.; Mariuzza, R. A. A Mutational Analysis of the Binding of Two Different Proteins to the Same Antibody. *Biochemistry* **1996**, *35*, 9667–9676.
- (30) Sundberg, E. J.; Urrutia, M.; Braden, B. C.; Isern, J.; Tsuchiya, D.; Fields, B. A.; Malchiodi, E. L.; Tormo, J.; Schwarz, F. P.; Mariuzza, R. A. Estimation of the Hydrophobic Effect in an Antigen-Antibody Protein-Protein Interface. *Biochemistry* **2000**, *39*, 15375–15387.
- (31) Dall'Acqua, W.; Goldman, E. R.; Lin, W.; Teng, C.; Tsuchiya, D.; Li, H.; Ysern, X.; Braden, B. C.; Li, Y.; Smith-Gill, S. J.; Mariuzza, R. A. A Mutational Analysis of Binding Interactions in an Antigen-Antibody Protein-Protein Complex. *Biochemistry* **1998**, *37*, 7981–7991.
- (32) Yokota, A.; Tsumoto, K.; Shiroishi, M.; Kondo, H.; Kumagai, I. The Role of Hydrogen Bonding Via Interfacial Water Molecules in Antigen-Antibody Complexation. The HyHEL-10-HEL Interaction. *J. Biol. Chem.* **2003**, *278*, 5410–5418.
- (33) Yokota, A.; Tsumoto, K.; Shiroishi, M.; Nakanishi, T.; Kondo, H.; Kumagai, I. Contribution of Asparagine Residues to the Stabilization of a Proteinaceous Antigen-Antibody Complex, HyHEL-10-Hen Egg White Lysozyme. *J. Biol. Chem.* **2010**, *285*, 7686–7696.
- (34) Pons, J.; Rajpal, A.; Kirsch, J. F. Energetic Analysis of an Antigen/Antibody Interface: Alanine Scanning Mutagenesis and Double Mutant Cycles on the HyHEL-10/Lysozyme Interaction. *Protein Sci.* **1999**, *8*, 958–968.
- (35) Tsumoto, K.; Ogasahara, K.; Ueda, Y.; Watanabe, K.; Yutani, K.; Kumagai, I. Role of Tyr Residues in the Contact Region of Anti-Lysozyme Monoclonal Antibody HyHEL10 for Antigen Binding. *J. Biol. Chem.* **1995**, *270*, 18551–18557.
- (36) Kelley, R. F.; O'Connell, M. P. Thermodynamic Analysis of an Antibody Functional Epitope. *Biochemistry* **1993**, *32*, 6828–6835.
- (37) Lang, S.; Xu, J.; Stuart, F.; Thomas, R. M.; Vrijbloed, J. W.; Robinson, J. A. Analysis of Antibody A6 Binding to the Extracellular Interferon Gamma Receptor Alpha-Chain by Alanine-Scanning Mutagenesis and Random Mutagenesis with Phage Display. *Biochemistry* **2000**, *39*, 15674–15685.
- (38) Li, Y.; Urrutia, M.; Smith-Gill, S. J.; Mariuzza, R. A. Dissection of Binding Interactions in the Complex between the Anti-Lysozyme Antibody HyHEL-63 and Its Antigen. *Biochemistry* **2003**, *42*, 11–22.
- (39) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 712–725.
- (40) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (41) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr. Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins: Struct., Funct., Genet.* **2009**, *77*, 778–795.
- (42) Smith, C. A.; Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *J. Mol. Biol.* **2008**, *380*, 742–756.
- (43) R Development Core Team. *R: A Language and Environment for Statistical Computing*; The R Foundation for Statistical Computing: Vienna, Austria, 2011.
- (44) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- (45) Moretti, R.; Fleishman, S. J.; Agius, R.; Torchala, M.; Bates, P. A.; Kasttritis, P. L.; Rodrigues, J. P.; Trellet, M.; Bonvin, A. M.; Cui, M.; Rومان, M.; Gillis, D.; Dehouck, Y.; Moal, I.; Romero-Durana, M.; Perez-Cano, L.; Pallara, C.; Jimenez, B.; Fernandez-Recio, J.; Flores, S.; Pacella, M.; Praneeth Kilambi, K.; Gray, J. J.; Popov, P.; Grudin, S.; Esquivel-Rodriguez, J.; Kihara, D.; Zhao, N.; Korkin, D.; Zhu, X.; Demerdash, O. N.; Mitchell, J. C.; Kanamori, E.; Tsuchiya, Y.; Nakamura, H.; Lee, H.; Park, H.; Seok, C.; Sarmiento, J.; Liang, S.; Teraguchi, S.; Standley, D. M.; Shimoyama, H.; Terashi, G.; Takeda-Shitaka, M.; Iwade, M.; Umeyama, H.; Beglov, D.; Hall, D. R.; Kozakov, D.; Vajda, S.; Pierce, B. G.; Hwang, H.; Vreven, T.; Weng, Z.; Huang, Y.; Li, H.; Yang, X.; Ji, X.; Liu, S.; Xiao, Y.; Zacharias, M.; Qin, S.; Zhou, H. X.; Huang, S. Y.; Zou, X.; Velankar, S.; Janin, J.; Wodak, S. J.; Baker, D. Community-Wide Evaluation of Methods for Predicting the Effect of Mutations on Protein-Protein Interactions. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 1980–1987.
- (46) Sirin, S.; Apgar, J. R.; Bennett, E. M.; Keating, A. E. AB-Bind: Antibody Binding Mutational Database for Computational Affinity Predictions. *Protein Sci.* **2016**, *25*, 393–409.
- (47) Oberlin, M.; Kroemer, R.; Mikol, V.; Minoux, H.; Tastan, E.; Baurin, N. Engineering Protein Therapeutics: Predictive Performances of a Structure-Based Virtual Affinity Maturation Protocol. *J. Chem. Inf. Model.* **2012**, *52*, 2204–2214.

- (48) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein-Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.
- (49) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.
- (50) Ginalski, K.; Elofsson, A.; Fischer, D.; Rychlewski, L. 3D-Jury: A Simple Approach to Improve Protein Structure Predictions. *Bioinformatics* **2003**, *19*, 1015–1018.
- (51) Saini, H. K.; Fischer, D. Meta-DP: Domain Prediction Meta-Server. *Bioinformatics* **2005**, *21*, 2917–2920.
- (52) Sastry, G. M.; Inakollu, V. S.; Sherman, W. Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking. *J. Chem. Inf. Model.* **2013**, *53*, 1531–1542.
- (53) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.