Superposition of Protein Structures Using Electrostatic Isopotentials

Ziyi Guo
Dept. of Computer Science and Engineering
Lehigh University
Bethlehem, PA, USA
zig312@lehigh.edu

Juliana Hong
Cornell University
Ithaca, NY , USA
julianahhong@gmail.com

Katya Scheinberg
Dept. of Industrial and Systems Engineering
Lehigh University
Bethlehem, PA, USA
katyas@lehigh.edu

Brian Y. Chen*
Dept. of Computer Science and Engineering
Lehigh University
Bethlehem, PA, USA
chen@cse.lehigh.edu

***Abstract*—Algorithms for comparing protein structures are widely used to identify proteins with similar functions and to examine the mechanisms of binding specificity. In order to make accurate comparisons, two structures must first be superposed, so that differences in position and orientation do not create misleading dissimilarities. Most algorithms generate these superpositions by aligning atoms of the peptide backbone. This approach is rapid, but it may not reflect similarities or differences in all mechanisms that proteins use to bind other molecules. Electric fields, for example, play a large role in recognition and their substantial range can interact with other molecules long before backbone contacts occur. To compare proteins based on their electric fields, we have developed the first algorithm designed to superpose protein structures using electric fields alone. Our method works by searching rotational and translational space for a superposition that maximizes the overlapping volume between electrostatic isopotentials. Applying this method to compare the serine protease and enolase superfamilies, our results demonstrate that our electrostatic superposition algorithm can distinguish very similar proteins with different binding preferences.**

I. Introduction

Atom coordinates are widely used to represent the geometry of proteins in structure comparison algorithms. Some algorithms detect large sets of alpha carbons with similar interatomic distances [1]–[5]. These methods can identify proteins with similar folds [6], [7] and find remote evolutionary relationships [8]. A second class of methods finds clusters of similar atoms that can reveal proteins that catalyze the same reactions [8]–[17]. Whether the atoms are detected by graph analysis [6], [18], dynamic programming [2], [19], by proximity to binding cavities [20] or through evolutionary analyses [10], [14], most approaches reported to date use atom coordinates to superpose protein structures before comparison. This approach, with many proven successes, exploits the fact that atoms define protein folds as well as the pattern of steric hindrance that is imposed on potential binders, so they are a logical choice for comparative analysis. Nonetheless, when considering all influences on molecular recognition, longer distance electrostatic effects can have a selective influence on binding partners even before they come into contact with the molecular surface. In such cases, a superposition of electrostatic potentials may reveal information about binding preferences that are not encoded in the geometry of atoms.

The problem we address in this paper concerns the case where positive and negative electrostatic isopotentials from two protein structures have been computed, and it is of interest to geometrically superpose those isopotentials so regions with similar potentials overlap as much as possible. To find the superposition with maximum overlap, we use Derivative Free Optimization (DFO) [21], a mathematical optimization technique. DFO operates by strategically evaluating dozens of individual superpositions, which are not based on atomic alignments, in a search for one with maximal overlapping volume. To evaluate overlapping volume, DFO calls on VASP-E [22]. VASP-E uses techniques from constructive solid geometry (CSG) to measure the volume of regions of intersection between two molecular solids. In each iteration, VASP-E measures the overlapping volume between the two positive isopotentials and separately between the two negative isopotentials, returning the sum of the intersection volumes to DFO.

The superposition technique presented in this paper differs fundamentally from existing superposition algorithms. By ignoring the positions of atoms in both proteins, the superpositions generated exclusively reflect similarities and differences in the electric fields experienced by potential binding partners. Because of the evolutionary pressure on these proteins to reliably interact with their intended binding partners, we hypothesize that similar fields, exhibited by larger intersection volumes, should occur between proteins that prefer to bind the same partners. Likewise, we expect that different fields, as identified by smaller volumes of intersection, should be a marker for proteins that prefer to bind different binding partners. We will evaluate these hypotheses on two families of proteins with diversified binding preferences in our experimental results.

This paper provides the first proof of concept that an

* Corresponding author.

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

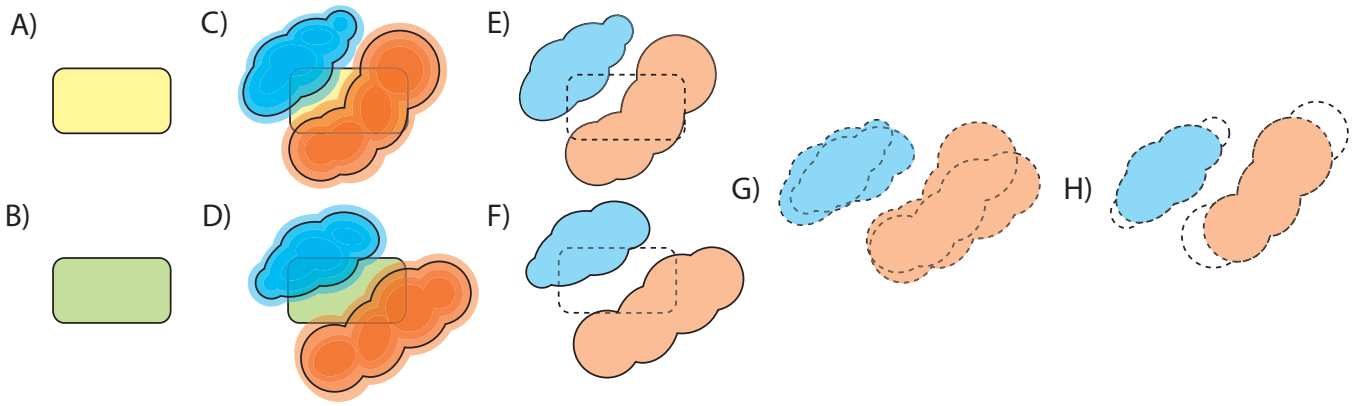

Fig. 1. An overview of our method. A) The protein structure A (rectangle in yellow). B) The protein structure B (rectangle in green). C, D) The positive potential (black outline in blue zone) that is larger than $|k|\ kT/e$ and the negative potential (black outline in red zone) that is smaller than $-|k|\ kT/e$. E, F) The selected isopotentials of structures A and B. G) The comparison of positive or negative isopotentials using VASP-E. H) DFO tests a variety of superpositions in search of one superposition where the isopotentials have the largest overlapping volume.

exclusively electrostatic superposition of protein structures is possible. Superpositions of this kind create a new analytical capability for the study of binding preferences: By superposing protein structures in a way that maximizes electrostatic similarities, we can examine what role the remaining differences play in binding specificity. We will examine these capabilities with applications to the serine proteases and the enolase superfamily.

II. Related Work

Comparisons of protein structures are essential to infer protein function. Most methods use points in three dimensions to describe whole protein structures [2], [3], [6]–[8], [19], [23] or functionally related active sites [9], [11], [14], [24]–[27] to minimize Root Mean Squred Distance (RMSD). This capacity enables structure comparison algorithms to efficiently detect similar proteins with maximal geometrical and biochemical similarity. Another kind of comparison method uses molecular surfaces which are the specification of protein structure at finer resolution [28], [29]. However, sometimes proteins perform their function because of the electrostatic interactions between typical protein components before binding partners contact with the molecular surface, and superposition of protein electrostatic potentials may exhibit preferences on protein binding specificity.

Some efforts have been made to analyze molecular electrostatic potentials that reveals protein function. These methods quantify charge distribution over the whole protein structure [30], [31] or localized region such as protein domains [32], active sites [33] protein-protein binding interfaces [34] or structural motifs [35]. Few more methods compared electrostatic potentials of proteins directly by computing a similarity index [36] or constructing tree-based structures [37]. Kinoshita et al. compared the electrostatic potentials on molecular surfaces to infer protein function [38], [39]. However, protein charges are unevenly distributed and the electrostatic potentials of any two points on the protein molecular surface could be essentially different. The comparison of isosurface of protein electrostatic potentials is an alternative direction.

This paper examines an electrostatic comparison of protein structures without depending on the molecular surface or on atomic positions. Rather than using atoms for superpostion, we begin with an arbitrary starting superposition. Using DFO, we search for superpositions that better overlap the electrostatic isopotentials. DFO has been successfully applied to superpose protein binding cavities [40]. This paper evaluates the hypothesis that it can also be used to superpose electrostatic

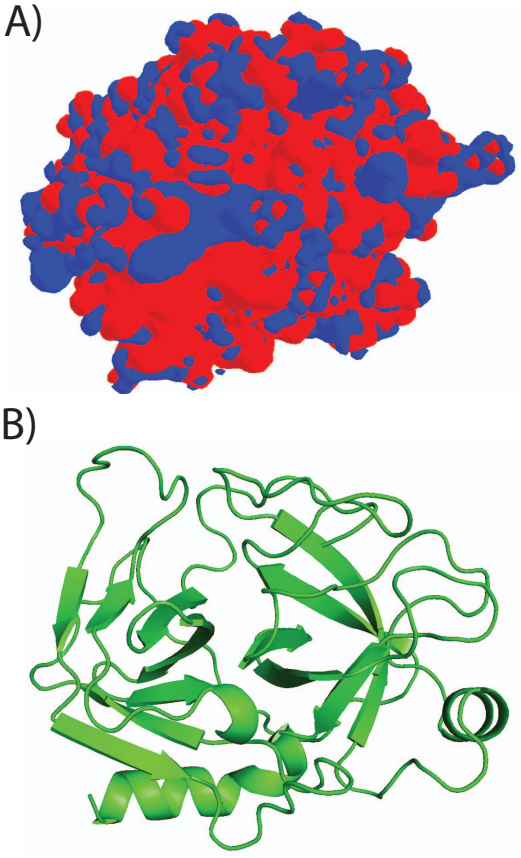

Fig. 2. A) electrostatic isopotential surfaces of the Atlantic salmon trypsin (pdb:$1a0j$). The red surface indicates the negative isopotential generated at $-5.0\ kT/e$ and blue indicates the positive isopotential generated at $5.0\ kT/e$. The surfaces are highly convoluted and pass very closely to each other, but do not come in contact. B) The cartoon visualization of $1a0j$ structure.

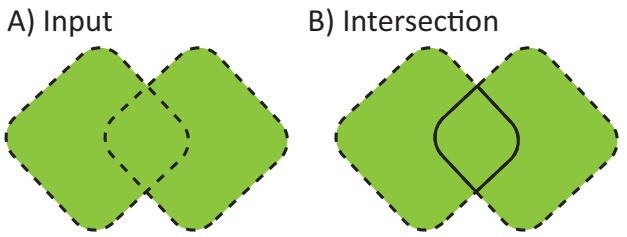

Fig. 3. A diagram of CSG intersection. A) input regions (green) defined by isopotential surfaces (dotted). B) input regions (green) isopotential surfaces (dotted), output intersection region (solid line).

isopotentials.

III. Methods

Beginning with the electrostatic isopotentials of two proteins as input, DFO searches for the superposition that maximizes overlapping volume by systematically testing a range of superpositions. For every intermediate superposition overlapping volume is calculated by VASP-E, which is called as a subalgorithm. To explain these calculations, we first describe how solid representations of electrostatic isopotentials are generated from protein structures. We then explain how we compare isopotentials in a given superposition by computing the sum of overlapping isopotential volumes. Finally, we explain how derivative free optimization searches for an optimal superposition. An overview of our method is illustrated in Figure 1 and is detailed in the following sections.

A. Solid representation of protein electrostatic isopotentials

As input, VASP-E requires the protein structure A, the electrostatic field A_E, the isopotential threshold k kT/e. When k is positive, VASP-E represents regions with electrostatic potential greater than k within a solid region, and when k is negative, regions with potential less than k are represented. This rule prevents the generation of infinitely large isopotential solids, which lead to degenerate comparisons and superpositions (Fig. 1C,1D).

To generate the electrostatic field, we first remove all hydrogens and then protonate the structure using the reduce tool of the MolProbity package [41] and the protonated structure is given as input to DelPhi [42] to compute a numerical solution to the Poisson-Boltzmann Equation (PBE). DelPhi approximates the electrostatic field A_E within a bounding box that covers the protein structure. As output, VASP-E generates electrostatic isopotentials as solid 3D objects.

The resulting isopotential solids can have highly convoluted shape. While isopotentials at different thresholds never overlap, they can be formed in close proximity to each other, as can be seen in Figure 2.

B. Electrostatic isopotential comparison

As input, VASP-E takes the desired resolution r, and solid representations of two electrostatic isopotentials A and B. VASP-E is a lattice based approximation algorithm that evaluates isopotential similarity by measuring the volume of intersection.

The intersection is computed using a method described earlier [43] but paraphrased here. First, a lattice of cubes is formed to completely surround both input isopotentials. For the corner of every cube, VASP-E determines whether the corner is inside or outside both input isopotentials. If a corner is inside both input isopotentials, it must be inside the output intersection region, but if it is outside any isopotential, then it must be outside the output intersection region. After identifying which corners are inside and outside the output region, the set of cubes that must contain the surface of the intersection region are found. The surface approximating the intersection region is approximated inside each cube, based on the points at which the cube intersects each input region. The volume of the intersection region is computed using the Surveyor's Formula [44].

When comparing isopotentials from related proteins, we say that larger intersection volume indicate greater similarity, and smaller intersection volumes indicate less similarity.

C. Derivative free optimization

Derivative free optimization methods focus on the unconstrained optimization problem. In such cases, the first derivative of the objective function is not available and cannot be approximated by traditional methods because the objective function is computationally costly. The optimal superposition of two isopotentials by maximizing intersection volume is one such problem since the objective function does not have a closed-form expression. Since we use VASP-E to approximate the volume of intersection, it becomes our objective function.

DFO represents the relative positions of isopotentials in a seven dimensional vector $\mathbf{x} = [A_x, A_y, A_z, \theta, T_x, T_y, T_z]$ where $[A_x, A_y, A_z]$ specifies the axis of rotation, θ specifies the rotation angle and $[T_x, T_y, T_z]$ are the translation vector. Our optimization problem can be formulated by the following equation:

$$min\{f(\mathbf{x}) = f^{+}(\mathbf{x}) + f^{-}(\mathbf{x}), \mathbf{x} \in R^n\} \quad (1)$$

where $f^{+}(\mathbf{x})$ denotes the overlapping volume between two positive isopotentials and $f^{-}(\mathbf{x})$ denotes the overlapping volume between two negative isopotentials. We adopt a trust-region based DFO method described in [21]. The trust region is the region around the current search point where model function that sufficiently approximates the objective function value is constructed. At each iteration, the model function is then minimized in the trust region to search for the best position for the next search point. More details about trust-region methods can be found in [45]. In practice, DFO tends to converge to a local optimizer and find a reasonable intersection volume, but the global optimal solution cannot be guaranteed. Hence, the choice of starting point may produce different final results. In this work, we translate centroids of input isopotentials to the origin before DFO as a trivial optimization to prevent unnecessary searching in translational space.

D. Data set

We demonstrate our method on two protein superfamilies, the serine proteases and the enolase superfamily. In the serine protease superfamily, we selected the trypsin, chymotrypsin

Serine Protease Superfamily:
Chymotrypsins: 1eq9, 8gch
Elastases: 1b0e, 1elt
Trypsins: 1a0j, 1aks, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91
Enolase Superfamily:
Enolases: 1e9i, 1iyx, 1te6, 1pdy, 2pa6, 3otr
Mandelate Racemase: 1mdr, 2ox4
Muconate Lactonizing Enzyme: 2pgw, 2zad

Fig. 4. **PDB codes used in the data set.**

and elastase subfamilies and in the enolase superfamily we selected the enolase, mandelate racemase and muconate lactonizing enzyme subfamilies. Each subfamily contained at least two nonredundant protein structures. Figure 4 lists all protein structures by their Protein Data Bank [46] (PDB) code and they are classified by similar ligand binding preferences.

The serine proteases are a family of enzymes that cleave peptide bonds in proteins where serine functions as the nucleophilic binding residue. Adjacent amino acids are recognized at specificity subsites numbered $S_4, S_3, S_2, S_1, S_1', S_2', S_3', S_4'$. Each subsite preferentially binds one amino acid before or after the $S_1 - S_1'$ hydrolyzed bond. In this work, we focus on the S_1 subsite of serine proteases, which reveals three different binding specificity: positively charged amino acid (lysine or arginine) [47] in trypsin, large and hydrophobic amino acid (tyrosine, phenylalanine or tryptophan) [48] in chymotrypsin and small hydrophobics (alanine, glycine or valine) [49].

The enolase superfamily catalyzes reactions by the abstraction of a proton from a carbon adjacent to a carboxylic acid with the help of a divalent metal ion [50]. These reactions occur near the C-terminal ends of beta sheets in the conserved TIM-barrel structures. In this work, we focus on three specific catalysts: the enolase subfamily, which converts 2-phosphoglycerate (2-PG) to phosphoenolpyruvate (PEP) [51], the mandelate racemases, which convert between (S)-mandelate and (R)-mandelate [52], and the muconate-lactonizing enzymes, which catalyze the conversion of lignin-derived aromatics, catechol and protocatechuate, to citric acid cycle intermediates [50].

Protein Selection. Protein structures in serine proteases and enolase superfamily were selected on 6.21.2011 with 676 serine proteases and 66 enolases. Protein structures with mutations and disordered regions and enolases with closed or partially closed structures were removed. Next, of member of any pair of structures with more than 90% sequence identity was removed, with preference for keeping structures with publication descriptions. 14 serine proteases and 10 enolases remained. Non protein atoms (ions, water etc.) and hydrogens were removed for uniformity.

We compared our superpositions against Ska, an algorithm for whole structure superpostion using alpha carbons [5]. To generate these alignments, we aligned all serine proteases onto $8gch$ and all enolases onto $1mdr$. These two structures were chose because of the existence of bound ligand, indicating a functional conformation. All structures in the same superfamily exhibit identical protein folds so that the choice of structural alignment algorithm creates little variation [43].

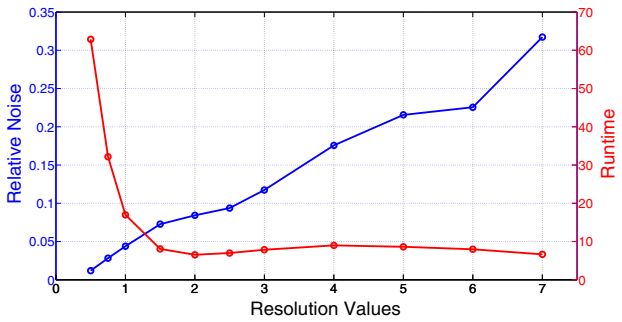

Fig. 5. VASP-E trade-off plot between relative noise (blue plot) and runtime (red plot) on a range of resolutions. The approximated noise-free resolution r^* is set to 0.25.

Superpositions generated with Ska can also be used as a starting point for the DFO superposition search. We refer to these superpositions as warm starts.

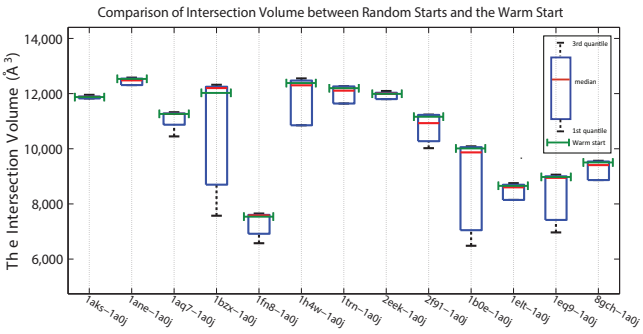

Fig. 6. The overlapping volume comparison of electrostatics isopotentials superposition from pdb $1a0j$ between random starts and warm starts. The isopotentials were generated at $10.0\ kT/e$. The boxplot illustrates the intersection volumes from LHS random starts with the mean value highlighted in red line where the warm start volumes are shown in green line.

IV. Experimental Results

In this section, we first calibrate the resolution to be used by VASP-E. Then we demonstrate the performance of warm start superpositions. Next, we perform superposition experiments on the serine proteases and the enolase superfamily, to examine whether our superposition technique can distinguish proteins with similar and different binding preferences. Finally, we compare our electrostatic superposition method to atom-based superpositions, to evaluate the difference between our superpositions and atom-based superpositions.

A. Calibrating VASP-E

As resolution r decreases, VASP-E approximates the electrostatic similarity using finer cubes, thus leading to more accurate evaluation but taking substantially longer time. To compute accuracy/time trade-offs, given the the resolution r for any input $\mathbf{x}$, the relative noise δ_r is defined as [53]:

$$\delta_r = \frac{f^*(\mathbf{x}) - f_r(\mathbf{x})}{f^*(\mathbf{x})} \tag{2}$$

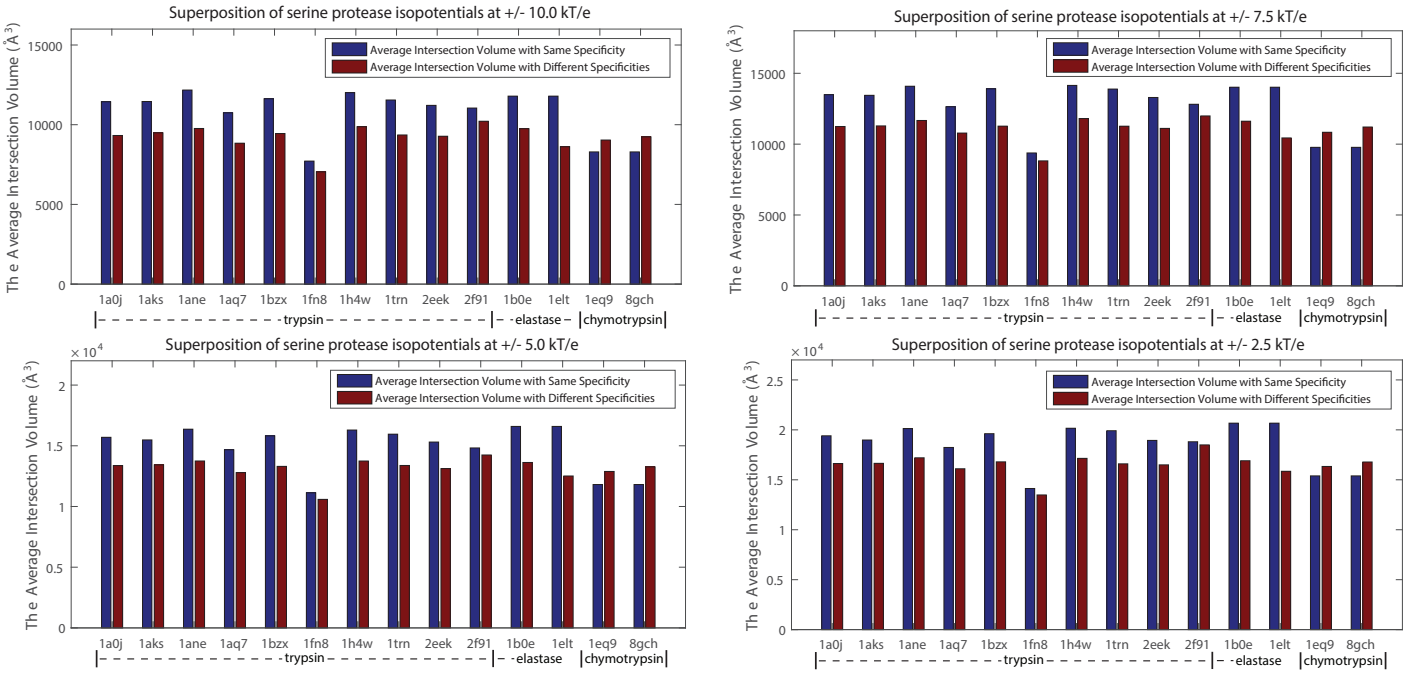

Fig. 7. The overlapping volume from simultaneous alignment of both positive and negative electrostatic isopotentials reveal protein-ligand binding specificities on serine proteases. Blue bars indicate the average overlapping volume between proteins with identical binding specificity while red bars indicate the average overlapping volume between proteins with different binding specificities. The calculation comes from isopotentials generated at 10.0 (top left), 7.5 (top right), 5.0 (bottom left) and 2.5 (bottom right) kT/e.

where $f^*(\mathbf{x})$ indicates the noise-free true function value and $f_r(\mathbf{x})$ indicates the function value computed by VASP-E with resolution r. Let r^* be the smallest resolution that is computational available, the true function value can be approximted by $f^*(\mathbf{x}) = f_{r^*}(\mathbf{x})$.

Figure 5 plots the relative noise estimation relative to runtime with respect to resolution between electrostatic isopotentials at 10.0 kT/e between $1a0j$ and $1b0e$. $1a0j$ and $1b0e$ are selected as examples, and are representative of other superpositions. It was observed that the runtime increases superlinearly and relative noise decreases as r decreases. This is expected because the number of lattice cubes used in VASP-E grows at a superlinear rate as r decreases. Similar patterns can also be found in other pairs and other isopotential thresholds. In our experiments, the smaller relative noise level is sufficient to obtain solution of acceptable accuracy, and the resolution r was set to 1.0 on our large-scale experiments.

B. Atomic superposition as warm starts

Optimization algorithms cannot guarantee that they identify a global optimum. However, in practice, we can evaluate how frequently DFO arrives at a logical optimum that is comparable to what can be found from existing methods. To perform this comparison, for every pair of superposed proteins, we generated 10 random starting superpositions, and searched for the optimal superposition from each. We compared these random start superpositions to the optimal superposition found when starting from backbone alignment generated by Ska: a warm start. Random starting vectors were generated by Latin Hypercube Sampling (LHS) [54], a statistical technique for calculating a distribution of initial starting vectors from a multidimensional distribution.

Figure 6 compares the final intersection volume from random starting vectors and from warm starting vectors between Atlantic Salmon Trypsin (pdb: $1a0j$) and other serine proteases at 10.0 kT/e. In general, the final intersection volume from random starts exhibited a dense concentration above the median and a few trailing superpositions well below the median. In almost all cases, final intersection volume from warm starts were between the median and the 75th percentile of the final intersection volume from random starts (Figure 6).

This behavior illustrates that while some random starting points yielded highly suboptimal superpositions, many random starting points provide superpositions that are comparable to that of warm started superpositions. Restarting DFO two or three times can thereby guarantee a high quality superposition.

C. Electrostatic isopotential superposition reveals binding specificities

To demonstrate that the superposition of electrostatic isopotentials reflects ligand binding specificities, we computed the average overlapping volume between one protein and the others with the same binding specificity and also with different binding specificities. Figures 7 and 8 report the average final intersection volume between serine proteases and enolases computed at isopotentials generated at 2.5, 5.0, 7.5 and 10.0 kT/e. Since we are computing overlapping volumes from positive and negative isopotentials, these tests optimized the superposition of the $+2.5$ kT/e and the -2.5 kT/e isopotentials together, then the $+5.0$ kT/e and the -5.0 kT/e isopotentials together, and so on.

At all four thresholds, the trypsin, elastase, enolase, and mandelate racemase subfamilies exhibited greater similarity to

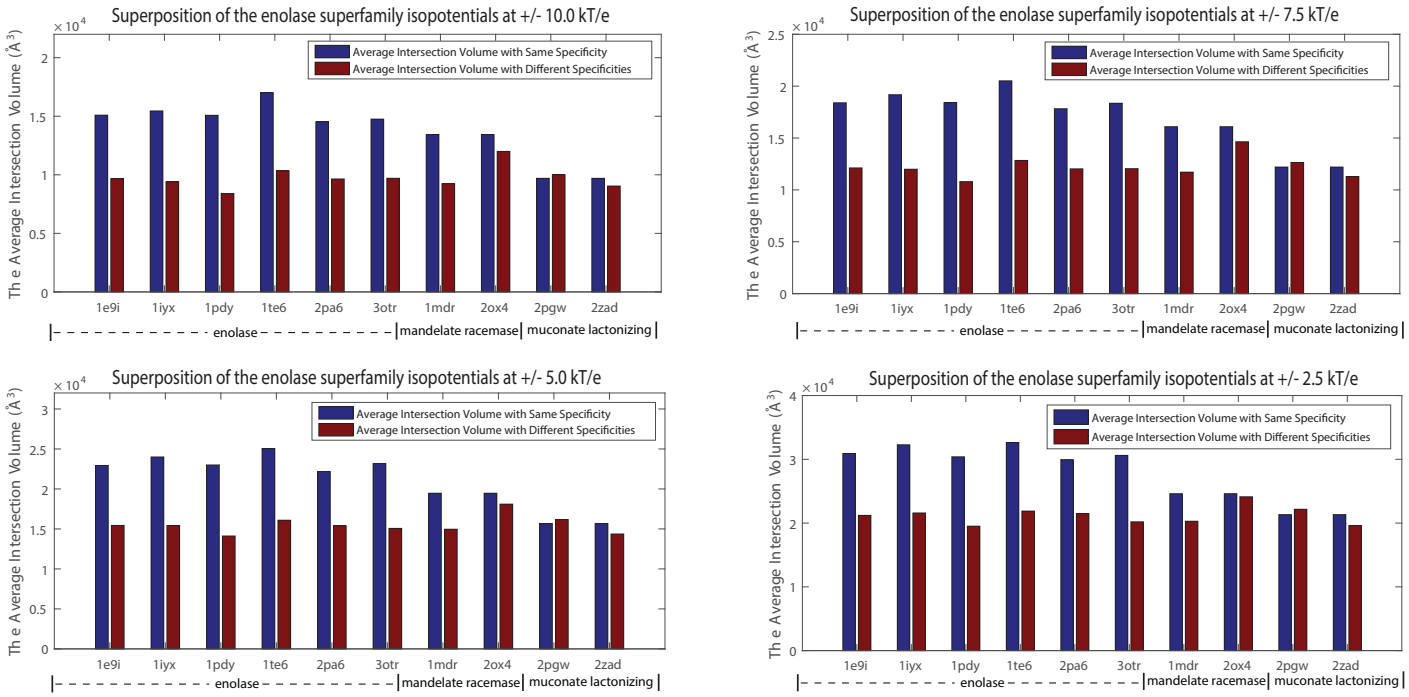

Fig. 8. The overlapping volume from simultaneous alignment of both positive and negative electrostatic isopotentials reveal protein-ligand binding specificities on the enolase superfamily. Blue bars indicate the average overlapping volume between proteins with identical binding specificity while red bars indicate the average overlapping volume between proteins with different binding specificities. The calculation comes from isopotentials generated at 10.0 (top left), 7.5 (top right), 5.0 (bottom left) and 2.5 (bottom right) kT/e.

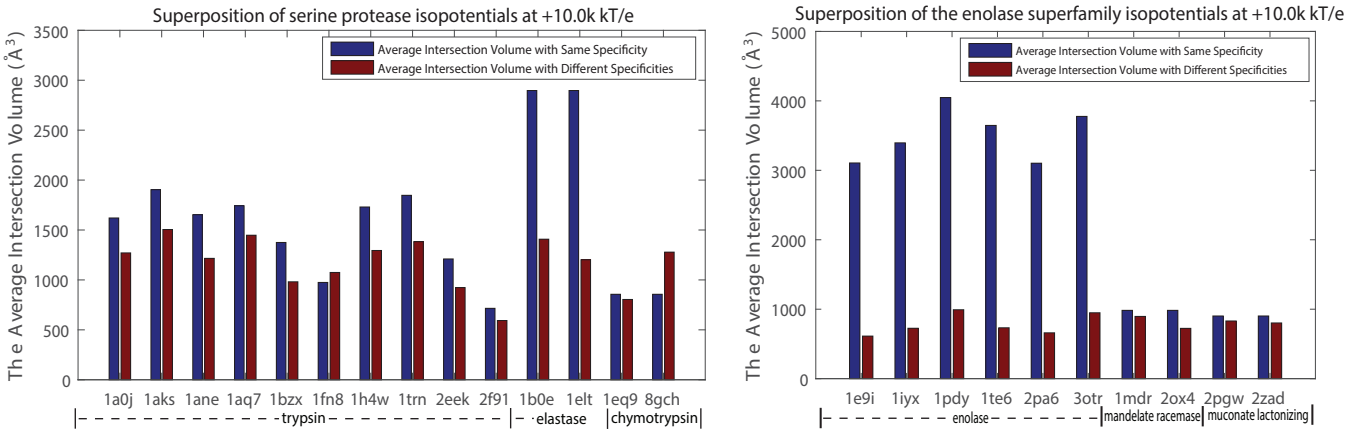

Fig. 9. The single charge isopotential alignment could enhance the binding specificity comparisons. Blue bars indicate the average overlapping volume between proteins with identical binding specificity while red bars indicate the average overlapping volume between proteins with different binding specificities. The calculation comes from positive isopotential alignment generated at 10.0 kT/e on serine proteases (left) and the enolase superfamily (right).

proteins with similar binding preferences than to proteins with different binding preferences, respectively. The chymotrypsin and muconate lactonizing enzyme proteins exhibited similar or slightly greater electrostatic similarity to proteins with different binding preferences, indicating that electrostatic superposition occasionally does not distinguish proteins with different binding preferences.

As a test, we also performed electrostatic superpositions using only a single electrostatic isopotential, rather than symmetric positive and negative isopotentials. Figure 9 illustrates the superposition of only the positive isopotential at 10.0 kT/e on two superfamilies. On serine proteases, two elastases ($1b0e$ and $1elt$) exhibited significantly greater similarity than when comparing both positive and negative isopotentials. A similar effect was observed on the enolase subfamily. Also, electrostatic similarity was slightly enhanced between chymotrypsins and muconate lactonizing enzymes. Repeating this computational at different isopotential thresholds revealed a similar pattern of enhanced similarities.

D. Structural deviations from isopotential superposition

Random starting vectors often yielded final superpositions that exhibit overlapping volumes that are comparable to warm

start superpositions. However, a large overlapping volume is no guarantee that the superposition actually reflects basic functional similarities, such as the superposition of similar binding sites.

We verified our superpositions against superpositions generated by ska. We began with superposition vectors generated by DFO from random starting points. Next, we rotated and translated the atoms of the protein according to the superposition vector as if they, and not the isopotentials, were superposed. Finally, we computed the all atom RMSD between the atoms of this rotated and translated structure and the protein when it was aligned using Ska. Note that we are not finding a new superposition by minimizing RMSD, we simply measure the root mean squared deviation between the electrostatic and ska-based superpositions.

Figure 10 illustrates a histogram of RMSDs generated in this manner between all serine proteases aligned on $10.0\ kT/e$ isopotentials at 10 random starting positions. It is apparent that almost all RMSDs exhibit very small values ($< 1.0\ \mathring{A}$). This result indicates that even when starting at random starting vectors, DFO frequently converges on a superposition that closely resembles the ska alignment. A few very large RMSD values ($> 9.0\ \mathring{A}$) also exist. Almost all of these results were generated by suboptimal electrostatic isopotential alignment and resulted in low volumes of isopotential intersection.

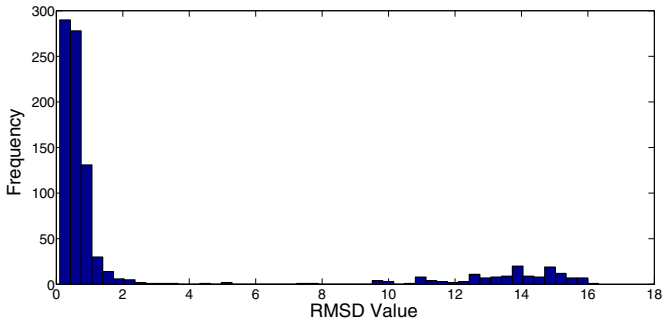

Fig. 10. The RMSD histogram plot on all pairs of serine proteases using electrostatic isopotential superposition at the threshold of $10.0\ kT/e$.

V. Conclusion

We have presented a computational method that adapts DFO and VASP-E to find superpositions of electrostatic isopotentials by maximizing their overlapping volume. Totally different from existing tools, our method does not use the positions of atoms to generate superpositions. Instead, it superposes isopotential surfaces the mathematical optimization method, DFO.

We tested our method on sequentially nonredundant subsets of two protein superfamilies: the serine proteases and the enolase superfamily. Our experiments showed that superposed isopotentials of proteins with identical binding preferences almost always exhibited larger intersection volume than superposed isopotentials from proteins with different binding preferences. This result indicates that the volumetric similarity between electrostatic potentials could be effective marker to infer protein binding partners.

Our method has great potential for applications to the comparison of electric fields. Representing electric fields as geometric entities, our method can identify local regions with similar potentials that are directly relevant to substrate binding. By maintaining algorithmic independence from any atomic structure, our method avoids biases that may be unavoidable for atom-based methods. Such independence, when considering the effect of electric fields on partner molecules, could yield useful insights into molecular design.

Acknowledgement

This work was supported in part by National Science Foundation Grant 1320137 to Brian Chen and Katya Scheinberg.

References

[1] R. Nussinov and H. J. Wolfson, "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques," *Proceedings of the National Academy of Sciences*, vol. 88, no. 23, pp. 10 495–10 499, 1991.

[2] C. A. Orengo and W. R. Taylor, "Ssap: sequential structure alignment program for protein structure comparison," *Computer methods for macromolecular sequence analysis*, 1996.

[3] D. Petrey and B. Honig, "Grasp2: visualization, surface properties, and electrostatics of macromolecular structures and sequences," *Methods in enzymology*, vol. 374, pp. 492–509, 2003.

[4] I. N. Shindyalov and P. E. Bourne, "An alternative view of protein fold space," *Proteins: Structure, Function, and Bioinformatics*, vol. 38, no. 3, pp. 247–260, 2000.

[5] A.-S. Yang and B. Honig, "An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance." *J Mol Biol*, vol. 301, no. 3, pp. 665–78, Aug. 2000.

[6] J.-F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison," *Current opinion in structural biology*, vol. 6, no. 3, pp. 377–385, 1996.

[7] L. Holm and C. Sander, "Mapping the protein universe," *Science*, vol. 273, no. 5275, pp. 595–602, 1996.

[8] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments," *Proceedings of the National Academy of sciences*, vol. 105, no. 14, pp. 5441–5446, 2008.

[9] J. A. Barker and J. M. Thornton, "An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis," *Bioinformatics*, vol. 19, no. 13, pp. 1644–1649, 2003.

[10] B. Y. Chen, D. H. Bryant, V. Y. Fofanov, D. M. Kristensen, A. E. Cruess, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction," *Journal of bioinformatics and computational biology*, vol. 5, no. 02a, pp. 353–382, 2007.

[11] R. B. Russell, "Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution," *Journal of molecular biology*, vol. 279, no. 5, pp. 1211–1227, 1998.

[12] B. J. Polacco and P. C. Babbitt, "Automated discovery of 3d motifs for protein function annotation," *Bioinformatics*, vol. 22, no. 6, pp. 723–730, 2006.

[13] T. A. Binkowski and A. Joachimiak, "Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites," *BMC structural biology*, vol. 8, no. 1, p. 45, 2008.

[14] B. Y. Chen, V. Y. Fofanov, D. H. Bryant, B. D. Dodson, D. M. Kristensen, A. M. Lisewski, M. Kimmel, O. Lichtarge, and L. E. Kavraki, "The mash pipeline for protein function prediction and an algorithm for the geometric refinement of 3d motifs," *Journal of Computational Biology*, vol. 14, no. 6, pp. 791–816, 2007.

[15] K. P. Peters, J. Fauck, and C. Frömmel, "The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria," *Journal of molecular biology*, vol. 256, no. 1, pp. 201–213, 1996.

[16] S. Schmitt, D. Kuhn, and G. Klebe, "A new method to detect related function among proteins independent of sequence and fold homology," *Journal of molecular biology*, vol. 323, no. 2, pp. 387–406, 2002.

[17] Y. Y. Tseng, J. Dundas, and J. Liang, "Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns," *Journal of molecular biology*, vol. 387, no. 2, pp. 451–464, 2009.

[18] P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice, and P. Willett, "A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures," *Journal of molecular biology*, vol. 243, no. 2, pp. 327–344, 1994.

[19] A.-S. Yang and B. Honig, "An integrated approach to the analysis and modeling of protein sequences and structures. i. protein structural alignment and a quantitative measure for protein structural distance," *Journal of molecular biology*, vol. 301, no. 3, pp. 665–678, 2000.

[20] T. A. Binkowski, A. Joachimiak, and J. Liang, "Protein surface analysis for function annotation in high-throughput structural genomics pipeline," *Protein Science*, vol. 14, no. 12, pp. 2972–2981, 2005.

[21] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to derivative-free optimization*. Siam, 2009, vol. 8.

[22] B. Y. Chen, "Vasp-e: Specificity annotation with a volumetric analysis of electrostatic isopotentials," *PLoS Comput Biol*, vol. 10, no. 8, 08 2014.

[23] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path." *Protein engineering*, vol. 11, no. 9, pp. 739–747, 1998.

[24] B. Chen, "Algorithms for structural comparison and statistical analysis of 3d protein motifs by chen, vy fofanov, dm kristensen, m. kimmel, o. lichtarge, and le kavraki pacific symposium on biocomputing 10: 334-345 (2005)," in *Pacific Symposium on Biocomputing*, vol. 10. Citeseer, 2005, pp. 334–345.

[25] M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, "Recognition of binding patterns common to a set of protein structures," in *Research in Computational Molecular Biology*. Springer, 2005, pp. 440–455.

[26] D. H. Bryant, M. Moll, B. Y. Chen, V. Y. Fofanov, and L. E. Kavraki, "Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction," *BMC bioinformatics*, vol. 11, no. 1, p. 242, 2010.

[27] Z. Guo and B. Y. Chen, "Variational bayesian clustering on protein cavity conformations for detecting influential amino acids," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 703–710.

[28] K. Kinoshita, J. Furui, and H. Nakamura, "Identification of protein functions from a molecular surface database, ef-site," *Journal of structural and functional genomics*, vol. 2, no. 1, pp. 9–22, 2002.

[29] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov, "Molecular shape comparisons in searches for active sites and functional similarity." *Protein Engineering*, vol. 11, no. 4, pp. 263–277, 1998.

[30] M. K. Gilson, K. A. Sharp, and B. H. Honig, "Calculating the electrostatic potential of molecules in solution: method and error assessment," *Journal of computational chemistry*, vol. 9, no. 4, pp. 327–335, 1988.

[31] D. R. Livesay, P. Jambeck, A. Rojnuckarin, and S. Subramaniam, "Conservation of electrostatic properties within enzyme families and superfamilies," *Biochemistry*, vol. 42, no. 12, pp. 3464–3473, 2003.

[32] D. Murray and B. Honig, "Electrostatic control of the membrane targeting of c2 domains," *Molecular cell*, vol. 9, no. 1, pp. 145–154, 2002.

[33] M. K. Gilson and B. H. Honig, "Calculation of electrostatic potentials in an enzyme active site," *Nature*, vol. 330, no. 6143, pp. 84–86, 1987.

[34] A. J. McCoy, V. C. Epa, and P. M. Colman, "Electrostatic complementarity at protein/protein interfaces," *Journal of molecular biology*, vol. 268, no. 2, pp. 570–584, 1997.

[35] S. A. Botti, C. E. Felder, J. L. Sussman, and I. Silman, "Electrotactins: a class of adhesion proteins with conserved electrostatic and structural motifs," *Protein engineering*, vol. 11, no. 6, pp. 415–420, 1998.

[36] A. M. Richard, "Quantitative comparison of molecular electrostatic potentials for structure-activity studies," *Journal of computational chemistry*, vol. 12, no. 8, pp. 959–969, 1991.

[37] X. Zhang, C. L. Bajaj, B. Kwon, T. J. Dolinsky, J. E. Nielsen, and N. A. Baker, "Application of new multiresolution methods for the comparison of biomolecular electrostatic properties in the absence of global structural similarity," *Multiscale Modeling & Simulation*, vol. 5, no. 4, pp. 1196–1213, 2006.

[38] K. Kinoshita and H. Nakamura, "Identification of protein biochemical functions by similarity search using the molecular surface database ef-site," *Protein Science*, vol. 12, no. 8, pp. 1589–1595, 2003.

[39] K. Kinoshita, Y. Murakami, and H. Nakamura, "ef-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W398–W402, 2007.

[40] R. Chen, K. Scheinberg, and B. Y. Chen, "Aligning ligand binding cavities by optimizing superposed volume," in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1–5.

[41] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "Molprobity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 1, pp. 12–21, 2009.

[42] W. Rocchia, E. Alexov, and B. Honig, "Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions," *The Journal of Physical Chemistry B*, vol. 105, no. 28, pp. 6507–6514, 2001.

[43] B. Y. Chen and B. Honig, "VASP: A volumetric analysis of surface properties yields insights into protein-ligand binding specificity," *PLoS computational biology*, vol. 6, no. 8, p. e1000881, 2010.

[44] J. Schaer and M. Stone, "Face traverses and a volume algorithm for polyhedra," in *New Results and New Trends in Computer Science*. Springer, 1991, pp. 290–297.

[45] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods*. Siam, 2000, vol. 1.

[46] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[47] K. Morihara and H. Tsuzuki, "Comparison of the specificities of various serine proteinases from microorganisms," *Archives of biochemistry and biophysics*, vol. 129, no. 2, pp. 620–634, 1969.

[48] L. Graf, A. Jancso, L. Szilágyi, G. Hegyi, K. Pintér, G. Náray-Szabó, J. Hepp, K. Medzihradszky, and W. J. Rutter, "Electrostatic complementarity within the substrate-binding pocket of trypsin," *Proceedings of the National Academy of Sciences*, vol. 85, no. 14, pp. 4961–4965, 1988.

[49] G. I. Berglund, A. O. Smalas, H. Outzen, and N. P. Willassen, "Purification and characterization of pancreatic elastase from north atlantic salmon (salmo salar)," *Molecular marine biology and biotechnology*, vol. 7, no. 2, pp. 105–114, 1998.

[50] P. C. Babbitt, M. S. Hasson, J. E. Wedekind, D. R. Palmer, W. C. Barrett, G. H. Reed, I. Rayment, D. Ringe, G. L. Kenyon, and J. A. Gerlt, "The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids," *Biochemistry*, vol. 35, no. 51, pp. 16 489–16 501, 1996.

[51] K. Kühnel and B. F. Luisi, "Crystal structure of the escherichia coli rna degradosome component enolase," *Journal of molecular biology*, vol. 313, no. 3, pp. 583–592, 2001.

[52] S. L. Schafer, W. C. Barrett, A. T. Kallarakal, B. Mitra, J. W. Kozarich, J. A. Gerlt, J. G. Clifton, G. A. Petsko, and G. L. Kenyon, "Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the d270n mutant," *Biochemistry*, vol. 35, no. 18, pp. 5662–5669, 1996.

[53] R. Chen, Z. Guo, B. Y. Chen, and K. Scheinberg, "Methodologies and software for derivative-free optimization," in *Optimization Methods in Engineering*. Society for Industrial and Applied Mathematics, submitted.

[54] M. D. McKay, R. J. Beckman, and W. J. Conover, "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.