(Compressed) sensing and sensibility

Vijay S. Pande¹

Departments of Chemistry, Structural Biology, and Computer Science, Stanford University, Stanford, CA 94305

For decades, researchers have built computer models of molecular interactions to predict properties of new molecules (1). These models take the form of potential functions, equations that can be used to predict the molecular energy of interaction. Potential functions have very broad applications. Other than *ab initio* quantum mechanics-based approaches (2), all simulations studies rely on some sort of model for potential functions. Thus, the ability to improve the accuracy of potential functions would have very wide sweeping impact. In PNAS, AlQuraishi and McAdams (3) propose a new approach to this challenging and important problem and demonstrate considerable gains in accuracy with their new method.

Physics vs. Data

In the past, there have traditionally been two approaches to the construction of potentials. One approach used high-level theory (such as quantum mechanics) or other laws of physics to give a physics-based foundation for interaction potentials. For example, one can use the law of Coulomb to describe the nature of electrostatic interaction between charged molecules. A physics-based approach is very appealing because of the rigor and quantitative nature of the physical sciences from which these potentials were derived. However, the challenge in these methods lies in the inevitable approximations needed to yield a computationally tractable computer model (e.g., approximations made to efficiently handle the nature of water) as well as the challenges of obtaining sufficiently accurate parameters for the physical potentials (1).

An alternative approach which has seen much practical value has been to build knowledge-based potentials (4). This approach uses existing experimental data to parameterize potential functions. Often, knowledge of the underlying physics does not come in at all, or comes just in the inspiration for functional forms of the potentials. The benefits of these approaches have been that they can more directly be of practical value, as they typically are less computationally demanding to parameterize and use.

Although appealing and useful, knowledge-based methods have not been a panacea (5). One central concern is that knowledge-based methods are, naturally, limited by the input data (i.e., the

“knowledge”) used to parameterize the model. Moreover, even in cases in which one is in the data-rich regime, these models are still prone to overfitting (6), which leads to limited transferability (i.e., limited regime of applicability). When overfit, a model does reasonably well

AlQuraishi and McAdams demonstrate an impressive advance in predictive capability.

within the regime it was fit (corresponding to situations very similar to the input data), but does very poorly in cases that start to differ. Overfitting is a common challenge in many data-driven areas, as even the recognition that a model can be overfit is at times underappreciated; consequences of this unfamiliarity often take the form of models with many more parameters than can be sufficiently determined by the data.

New Approach

Recently, there has been a revolution in statistical approaches which can help address these issues and take the data-driven approach to new levels. In particular, AlQuraishi and McAdams (3) innovatively combine the compressed sensing statistical approach (7)—best known in applications of electrical engineering and signal processing applications, such as in MRI image processing (8)—with the potential function determination process. The result is a completely new way to look at the potential function determination problem.

At its heart, compressed sensing is a technique for finding sparse solutions to underdetermined linear systems, i.e., solutions that avoid overfitting. This is done by using a regularization scheme, which systematically avoids overfitting by penalizing overly complex models (6). For example, for least-squares fitting of parameters, one does not merely find the parameters that yield the smallest deviation from the experiment in a least-squares manner, but one adds an additional penalty term that down-weights solutions with additional parameters.

Thus, the method of AlQuraishi and McAdams (3) recasts the potential determination problem into an undeter-

mined set of linear equations. By using the regularization schemes intrinsic to compressed sensing methods, their method yields a solution of the equations that avoids overfitting in a systematic means. The result is an optimal use of the data to yield new potential functions. One important flip side to avoiding overfitting is the challenges of underfitting, i.e., building a model that is too simple given the data. What is particularly intriguing about the method of AlQuraishi and McAdams (3) is that they infer not just parameters for a potential function, but the function itself. This is likely key to the performance of this approach, as it is not limited by *a priori* biases to the fundamental nature of the potential functions.

With its roots in statistical rigor and statistical mechanics, this approach is very appealing. However, in the end, the key question in the end is how its performance compares with other competing approaches. AlQuraishi and McAdams (3) demonstrate an impressive advance in predictive capability: when applied to the prediction of the binding specificity of proteins to DNA, they found approximately 90% accuracy, compared with approximately 60% for the best-performing alternative computational methods. It will be exciting to see future applications of this method to other areas.

With all its strengths, it is important to stress that this model does not completely solve the problem of transferability. Although these models are sufficiently regularized to avoid overfitting, they are still limited by the fundamental nature of the data used as input. This differs from a physics-based approach that, in principle, does not suffer from this issue of transferability, if all the relevant degrees of freedom are included in the model (9). This suggests that a fusion of both approaches is particularly appealing, whereby physical properties are used as prior knowledge (i.e., as a prior in a Bayesian formulation) but then the model is derived from existing data. The ability to fuse data-driven and physics-based approaches could push these types of models even further.

Author contributions: V.S.P. wrote the paper.

The author declares no conflict of interest.

See companion article on page 14819.

¹E-mail: pande@stanford.edu.

1. Leach AR (2001) *Molecular Modelling: Principles and Applications* (Prentice Hall, New York), 2nd Ed.
2. Kresse G (1995) Ab-initio molecular-dynamics for liquid-metals. *J Non-Cryst Solids* 193:222–229.
3. AlQuraishi M, McAdams HH (2011) Direct inference of protein–DNA interactions using compressed sensing methods. *Proc Natl Acad Sci USA* 108:14819–14824.
4. Sippl MJ (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5:229–235.
5. Das R (2011) Four small puzzles that Rosetta doesn't solve. *PLoS ONE* 6:e20044.
6. Hastie T, Tibshirani R, Friedman JH (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York), 2nd Ed.
7. Donoho DL (2006) Compressed sensing. *IEEE Trans Inform Theory* 52:1289–1306.
8. Lustig M, Donoho D, Pauly JM (2007) Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn Reson Med* 58:1182–1195.
9. Ponder JW, Case DA (2003) Force fields for protein simulations. *Adv Protein Chem* 66:27–85.