# Putting the pieces together: *Integrative Modeling Platform* for structure determination of macromolecular assemblies

Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson, and Andrej Sali\*

Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences (QB3), University of California, San Francisco, San Francisco, CA 94158

Keren Lasker is also a graduate student at the Raymond and Beverly Sackler Faculty of Exact Sciences,Blavatnik School of Computer Science, Tel Aviv University, Tel-Aviv, Israel

Bret Peterson is currently at Google, Mountain View, CA

\* Corresponding author: Email: [sali@salilab.org](mailto:sali@salilab.org), Telephone: +415-514-4227, Fax: +415-514-4231

## Introduction

Building models of a biological system that are consistent with the myriad data available is one of the key challenges in biology. In particular, we are interested in the models of structure and dynamics for macromolecular assemblies. These models can give insights into how these systems work, how they evolved, how they can be controlled, how similar functionality can be designed. While modeling can also help with planning future experiments. current publishing norms make it hard to capitalize on published models. Here, we present steps towards a future where a scientist can read a paper, download a script, add new data, and see how the new data improved the current model.

One successful approach, integrative structure modeling, casts the building of such models as a computational optimization problem where information about the assembly is encoded into a scoring function used to evaluate candidate models. Previous applications of integrative modeling include determining the structures of the 26S proteasome from an electron microscopy (EM) map, proteomics data, and comparative protein structure models of components [[1](#_ENREF_1)]; the bacterial type II pilus from sparse NMR data and X-ray crystallography structures of constituent proteins [[2](#_ENREF_2)]; chromatin from 5C data [[3](#_ENREF_3)]; auxilin bound to clathrin from an EM map and comparative models of components [[4](#_ENREF_4)]; the human voltage dependent anion channel from NMR spectroscopy and X-ray crystallography structures of constituent proteins [[5](#_ENREF_5)]; eukaryotic initiation factor 3 from mass spectrometry and proteomics data [[6](#_ENREF_6)]; and the nuclear pore complex (NPC) from the biophysical, proteomics, and EM data [[7](#_ENREF_7)].

The model of the yeast NPC [[8](#_ENREF_8)] illustrates the value of integrative modeling. The sheer size and flexibility of the NPC makes it all but impossible to solve its molecular architecture by conventional atomic resolution techniques, such as X-ray crystallography. However, integrating information from multiple sources, including stoichiometry from protein quantification, protein proximities from subcomplex purification, protein positions from immuno-EM, sedimentation analysis informative about the protein and subcomplex shapes, and the overall NPC shape from EM, resulted in an ensemble of medium resolution models. The models were summarized by a 3D probability map, resembling an EM map and localizing the 456 constituent proteins with an average precision of ~5 nm. This map has revealed fundamental new insights into the function and evolution of the NPC [[7](#_ENREF_7),[9](#_ENREF_9),[10](#_ENREF_10),[11](#_ENREF_11)].

Integrative modeling is a computational encoding of the standard scientific cycle of gathering data, proposing hypotheses, and then gathering more data to test and refine those hypotheses. That is, it proceeds through repeated iterations of the stages of gathering information, choosing how to represent and evaluate models, finding models that score well, and analyzing the models and information (Table 1). The cycle terminates when a convergent ensemble of models is found fitting the current information and the models have been judged to be satisfactory [[12](#_ENREF_12)]. When new information is gathered, whether by other scientists, labs, or techniques, the cycle is resumed.

The integrative approach has a number of advantages over informal or partial consideration of available information (Table 2). Fully realizing these advantages requires encoding modeling efforts as integrative modeling applications that consist of the scripts and the associated information. Adoption of integrative modeling can occur through (i) a tight collaboration between a computational lab and an experimental lab, (ii) adoption by an experimental lab, or (iii) experimentalists modifying existing integrative modeling applications. To facilitate widespread adoption, we have developed the *Integrative Modeling Platform* (IMP).

## Introducing IMP

IMP is a software package that facilitates (i) writing integrative modeling applications; (ii) developing new model representations, scoring functions, sampling schemes, and analysis methods; and (iii) distributing integrative modeling applications (Table 3).

In IMP, models are encoded as collections of particles, each representing a piece of the system. Depending on the data available, particles can be used to create atomic, coarse-grained, or hierarchical representations. It is straightforward to represent a protein at any resolution, from fully flexible atomic models (*ie,* one particle per atom), to rigid bodies, to coarse-grained models consisting of only one or a few particles for the whole protein. Different parts of the model can be represented differently, as dictated by the available information. Each particle has associated attributes, such as coordinates, radius, atom type, rigid body composition, residue information, and mass. If the numerous attributes already in IMP are not sufficient, new attributes can be created and used similarly to the predefined ones. For example, for coarse-grained small angle X-ray scattering (SAXS) scoring, a scattering factor attribute could be associated with the particles representing amino acid residues.

A given model is evaluated by a scoring function composed of terms called restraints, each of which measures how well a model agrees with the information from which the restraint was derived. The precision and accuracy of the resulting model increases with the amount and quality of data that is encoded in the restraints. IMP’s ever-growing set of scoring function types includes ones for SAXS profiles [[13](#_ENREF_13)], proteomics data [[12](#_ENREF_12)], EM images and density maps [[14](#_ENREF_14),[15](#_ENREF_15)], NMR spectroscopy [[2](#_ENREF_2)], the CHARMM force-field [[16](#_ENREF_16)], alignment with related structures [[17](#_ENREF_17)], and a variety of statistical potentials [[18](#_ENREF_18)]. Multiple research groups are implementing additional types of scoring, including those encoding various mass spectrometry measurements and the energy function for atomic structure prediction and docking from the Rosetta program [[19](#_ENREF_19)].

For experimental data, the scoring is generally implemented using a “forward model” [[20](#_ENREF_20)], which simulates the measurements based on the model being assessed and then compares the simulated measurements to the actual measurements. For example, to evaluate the fit to an EM density map, a restraint uses the coordinates, radii, and masses of a set of particles representing the assembly to simulate its density map and then evaluates the cross-correlation with the experimental map.

As with most computational structure efforts, the main computational cost in integrative modeling is sampling models that satisfy the restraints (good-scoring models). IMP provides a wide variety of tools for building these sampling protocols, including optimization algorithms such as Monte Carlo [[21](#_ENREF_21)] and conjugate gradients [[22](#_ENREF_22)], the simplex optimizer from Gnu Scientific Library (GSL; [[23](#_ENREF_23)]), simulation schemes such as molecular dynamics and Brownian dynamics [[24](#_ENREF_24)] (either using our molecular and Brownian dynamics codes or external ones such as OpenMM [[25](#_ENREF_25)]), and the Bullet rigid body dynamics engine (http://www.bulletphysics.com), as well as full sampling schemes such as the Gibbs sampler [[20](#_ENREF_20)] and replica exchange [[26](#_ENREF_26)]. In addition, we are developing a divide-and-conquer sampler called DOMINO [[27](#_ENREF_27)]. DOMINO divides the particles based into loosely coupled subsets, solves the subsets independently, and merges the solutions to enumerate all models with good scores over a supplied discrete sampling space. IMP is structured so that applications can be run in a distributed manner (*eg,* on a cluster of computing nodes).

Finally, IMP provides a variety of tools for comparing, clustering, and analyzing models. These tools can be used to check for quality-of-fit, the existence of multiple states of the system [[3](#_ENREF_3)], and inconsistent information. Models can be clustered based on root-mean-square deviation (RMSD), placement score [[14](#_ENREF_14)], and various other metrics. Supported clustering algorithms include k-means, centrality betweenness clustering [[28](#_ENREF_28)], and simple binning. The resulting clusters and the constituent models as well as restraints can be exported to Chimera [[29](#_ENREF_29)] and Pymol [[30](#_ENREF_30)] for visual inspection and further analysis.

IMP has been used to produce a number of models; for example, a eukaryotic ribosome [[31](#_ENREF_31)], a mammalian ribosome [[32](#_ENREF_32)], an RyR channel [[33](#_ENREF_33)], the 26S proteasome [[1](#_ENREF_1)], the Hsp90 chaperonin [[34](#_ENREF_34)], the TRiC/CCT chaperonin [[35](#_ENREF_35)], the actin-scruin complex [[36](#_ENREF_36)], chromatin [[3](#_ENREF_3)], and the NPC [[7](#_ENREF_7)].

More information about IMP can be found at [http://integrativemodeling.org](http://www.google.com/url?q=http%3A%2F%2Fwww.integrativemodeling.org&sa=D&sntz=1&usg=AFQjCNFL5A79aCzDOE3jQCNoOmO7JmJzTQ)/. The web site provides a technical introduction, a tutorial, as well as a variety of examples to help users get started. In addition, it contains nightly tests, user and developer email lists, a wiki, and a bug tracker.

## Towards open structure modeling

Publication of macromolecular structures has evolved from printed words and pictures to include deposition of coordinates in the Protein Data Bank [[37](#_ENREF_37)], and more recently deposition of raw input data such as X-ray scattering factors [[37](#_ENREF_37)], NMR restraints [[38](#_ENREF_38)], and EM particle images [[39](#_ENREF_39)]. However, the conversion of the raw data to the final structures is often only briefly described and all too rarely available in a directly usable form [[40](#_ENREF_40),[41](#_ENREF_41),[42](#_ENREF_42)], making reproduction and use of the published results laborious or even impossible.

If published papers included integrative modeling applications, a wide variety of researchers would benefit. Particularly, experimental labs, which are unlikely otherwise to go through the effort of modeling systems themselves, would be able to use the state-of-the-art model in experiment planning by simulating how much benefit would be achieved from new data. It would also be easy to see how much each new measurement contributes to the current model as well as whether or not it is consistent with it. Other computational groups could more easily experiment with new scoring, sampling, and analysis methods, without having to reimplement the existing methods from scratch. The common abstraction would make it easier to mix-and-match parts of other modeling packages [[16](#_ENREF_16),[17](#_ENREF_17),[19](#_ENREF_19),[20](#_ENREF_20),[43](#_ENREF_43),[44](#_ENREF_44),[45](#_ENREF_45),[46](#_ENREF_46),[47](#_ENREF_47),[48](#_ENREF_48),[49](#_ENREF_49),[50](#_ENREF_50),[51](#_ENREF_51)] to improve the applications of integrative modeling. Finally, the authors themselves would maximize the impact of their work, increasing the odds that their results are incorporated into future modeling.

## Acknowledgements

We would like to thank Frank Alber and Friedrich Förster for their contributions to the development of the integrative modeling paradigm; Ben Schwarz and Yannick Spill for contributing to the IMP code and design; Riccardo Pellarin, Massimiliano Bonomi and Ben Schwarz for insightful comments on early iterations of the paper; and Jeremy Phillips, Frank Alber, Friedrich Förster, Marc Marti-Renom and Davide Bau for being early users of the IMP library. The work was funded by NIH grants R01 GM083960, PN2 EY016525, and U54 RR022220.

## References

1. Förster F, Lasker K, Nickell S, Sali A, Baumeister W (2010) Toward an integrated structural model of the 26S proteasome. Molecular & cellular proteomics : MCP 9: 1666-1677.

2. Simon B, Madl T, Mackereth CD, Nilges M, Sattler M (2010) An efficient protocol for NMR-spectroscopy-based structure determination of protein complexes in solution. Angewandte Chemie (International ed in English) 49: 1967-1970.

3. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, et al. (2010) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nature structural & molecular biology.

4. Fotin A, Cheng Y, Grigorieff N, Walz T, Harrison SC, et al. (2004) Structure of an auxilin-bound clathrin coat and its implications for the mechanism of uncoating. Nature 432: 649-653.

5. Bayrhuber M, Meins T, Habeck M, Becker S, Giller K, et al. (2008) Structure of the human voltage-dependent anion channel. Proceedings of the National Academy of Sciences of the United States of America 105: 15370-15375.

6. Zhou M, Robinson CV (2010) When proteomics meets structural biology. Trends in Biochemical Sciences 35: 522-529.

7. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, et al. (2007) The molecular architecture of the nuclear pore complex. Nature 450: 695-701.

8. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, et al. (2007) Determining the architectures of macromolecular assemblies. Nature 450: 683-694.

9. Wente SR, Rout MP (2010) The nuclear pore complex and nuclear transport. Cold Spring Harbor perspectives in biology 2: a000562.

10. Devos D, Dokudovskaya S, Williams R, Alber F, Eswar N, et al. (2006) Simple fold composition and modular architecture of the nuclear pore complex. Proceedings of the National Academy of Sciences of the United States of America 103: 2172-2177.

11. DeGrasse JA, DuBois KN, Devos D, Siegel TN, Sali A, et al. (2009) Evidence for a shared nuclear pore complex architecture that is conserved from the last common eukaryotic ancestor. Molecular & cellular proteomics : MCP 8: 2119-2130.

12. Alber F, Förster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. Annual review of biochemistry 77: 443-477.

13. Schneidman-Duhovny D, Hammel M, Sali A (2011) Macromolecular docking restrained by a small angle X-ray scattering profile. Journal of Structural Biology 173: 461-471.

14. Lasker K, Sali A, Wolfson HJ (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. Proteins 78: 3205-3211.

15. Lasker K, Phillips JL, Russel D, Velazquez-Muriel J, Schneidman-Duhovny D, et al. (2010) Integrative Structure Modeling of Macromolecular Assemblies from Proteomics Data. Molecular & cellular proteomics : MCP 9: 1689-1702.

16. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, et al. (2009) CHARMM: the biomolecular simulation program. Journal of computational chemistry 30: 1545-1614.

17. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology 234: 779-815.

18. Shen M-Y, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein science : a publication of the Protein Society 15: 2507-2524.

19. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, et al. (2011) Rosetta3 an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology 487: 545-574.

20. Rieping W, Nilges M, Habeck M (2008) ISD: a software package for Bayesian NMR structure calculation. Bioinformatics (Oxford, England) 24: 1104-1105.

21. N. Metropolis AWR, M.N. Rosenbluth, A.H. Teller, E. Teller (1953) Equation of state calculation by fast computing machines. Journal of chemical physics 21: 1087-1092.

22. W H Press SAT, W T Vetterling, B P Flannery (1997) Numerical Recipes in C: Cambridge University Press.

23. Gough B, editor (2006) GNU Scientific Library Reference Manual: Network Theory Ltd.

24. Schlick T (2002) Molecular Modeling and Simulation: Springer.

25. Eastman P, Pande V (2010) OpenMM: A Hardware-Independent Framework for Molecular Simulations. Computing in Science & Engineering 12: 34-39.

26. Zhou R (2007) Replica exchange molecular dynamics method for protein folding simulation. Methods in molecular biology 350: 205-223.

27. Lasker K, Topf M, Sali A, Wolfson HJ (2009) Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. Journal of Molecular Biology 388: 180-194.

28. Freeman L (1977) A set of measures of centrality based on betweenness. Sociometry.

29. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry 25: 1605-1612.

30. DeLano W (2002) The PyMOL molecular graphics system. citeulikeorg.

31. Taylor DJ, Devkota B, Huang AD, Topf M, Narayanan E, et al. (2009) Comprehensive molecular structure of the eukaryotic ribosome. Structure 17: 1591-1604.

32. Chandramouli P, Topf M, Ménétret J-F, Eswar N, Cannone JJ, et al. (2008) Structure of the mammalian 80S ribosome at 8.7 A resolution. Structure 16: 535-548.

33. Serysheva II, Ludtke SJ, Baker ML, Cong Y, Topf M, et al. (2008) Subnanometer-resolution electron cryomicroscopy-based domain models for the cytoplasmic region of skeletal muscle RyR channel. Proceedings of the National Academy of Sciences of the United States of America 105: 9610-9615.

34. Krukenberg KA, Förster F, Rice LM, Sali A, Agard DA (2008) Multiple conformations of E. coli Hsp90 in solution: insights into the conformational dynamics of Hsp90. Structure 16: 755-765.

35. Booth CR, Meyer AS, Cong Y, Topf M, Sali A, et al. (2008) Mechanism of lid closure in the eukaryotic chaperonin TRiC/CCT. Nature structural & molecular biology 15: 746-753.

36. Cong Y, Topf M, Sali A, Matsudaira P, Dougherty M, et al. (2008) Crystallographic conformers of actin in a biologically active bundle of filaments. Journal of Molecular Biology 375: 331-336.

37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic acids research 28: 235-242.

38. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, et al. (2008) BioMagResBank. Nucleic acids research 36: D402-408.

39. Lawson CL, Baker ML, Best C, Bi C, Dougherty M, et al. (2011) EMDataBank.org: unified data resource for CryoEM. Nucleic acids research 39: D456-464.

40. Mesirov JP (2010) Computer science. Accessible reproducible research. Science 327: 415-416.

41. Barnes N (2010) Publish your computer code: it is good enough. Nature 467: 753.

42. Merali Z (2010) Computational science: ...Error. Nature 467: 775-777.

43. Ludtke SJ (2010) 3-D structures of macromolecules using single-particle analysis in EMAN. Methods in molecular biology 673: 157-173.

44. Shaikh TR, Gao H, Baxter WT, Asturias FJ, Boisset N, et al. (2008) SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nature protocols 3: 1941-1974.

45. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. Journal of magnetic resonance 160: 65-73.

46. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta crystallographica Section D, Biological crystallography 66: 213-221.

47. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. Journal of computational chemistry 26: 1668-1688.

48. Sharma S, Ding F, Nie H, Watson D, Unnithan A, et al. (2006) iFold: a platform for interactive folding simulations of proteins. Bioinformatics 22: 2693-2694.

49. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation 4: 435-447.

50. Peter E (2010) OpenMM: A Hardware-Independent Framework for Molecular Simulations. In: Vijay P, editor. pp. 34-39.

51. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with NAMD. Journal of computational chemistry 26: 1781-1802.

52. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, et al. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic acids research 39: D465-474.

53. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic acids research 39: D698-704.

54. Ma W, Trusina A, El-Samad H, Lim WA, Tang C (2009) Defining network topologies that can achieve biochemical adaptation. Cell 138: 760-773.

55. Schneidman-Duhovny D, Hammel M, Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic acids research 38: W540-544.

**Figure 1. Integrative structure modeling of the human RNA Polymerase II [**[**15**](#_ENREF_15)**]**. The first round of modeling was performed using only the EM density map of the assembly from EMDB [[39](#_ENREF_39)] and subunit comparative models from ModBase [[52](#_ENREF_52)] based on the crystallographic structures of the yeast RNAPII proteins. The data was found to be insufficient to uniquely resolve the structure. To overcome this challenge, protein interaction networks extracted from BioGrid [[53](#_ENREF_53)] were added. The addition of this data resulted in a single structure. The scripts are available as part of IMP.

**Table 1: The four stages of the integrative modeling cycle.**

|  |  |
| --- | --- |
| *gathering information* | This information consists of data from wet lab experiments such as those listed above, as well as statistical tendencies such as atomic statistical potentials, physical laws such as molecular mechanics force field, and anything else that can be converted into a score on features of a structural model. |
| *choosing how to represent and evaluate models* | The resolution of the representation depends on the quantity and resolution of the available information, and should be commensurate with the resolution of the final models; different parts of a model may be represented at different resolutions and a part of the model may be represented at several different resolutions simultaneously. The scoring function evaluates whether or not a given model is consistent with the input information, taking into account the uncertainty in the information. |
| *finding models that score well* | The search is performed using any of a variety of sampling and optimization schemes (*eg*, Monte Carlo method). There may be many models that score well if the data is incomplete or none if the data is inconsistent due to errors or unconsidered states of the assembly. |
| *analyzing models and information* | The ensemble of good-scoring models need to be clustered and analyzed to ascertain their precision and accuracy, and to check for inconsistent information. Analysis can also suggest what are likely to be the most informative experiments to perform in the next iteration. |

Integrative modeling iterates through these cycles until a satisfactory model is built. Many iterations of the cycle may be required before a satisfactory model is found due to the need to gather more data as well as resolve errors and inconsistent data.

**Table 2. Advantages of the integrative structure modeling approach.**

|  |  |
| --- | --- |
| *using new information* | Integrative modeling makes it easy to take advantage of new information and new types of information, resulting in a low barrier for using incremental information that is generally not applied to structure characterization (*eg*, proteomics data). Even when a single type is relatively uninformative, multiple types can give a surprisingly complete picture of an assembly [[12](#_ENREF_12),[15](#_ENREF_15)]. |
| *maximizing accuracy, precision and completeness* | The integrative models by construction fit multiple types of information, and can thus be more accurate, precise, and complete than models based on the individual sources. |
| *understanding the models* | By exhaustively sampling the space of models fitting the information, integrative modeling can find all models fitting the information, not only one. A full sampling of the models of structure can improve the understanding of the function [[54](#_ENREF_54)]. |
| *assessing models* | Since the data is encoded in scoring functions and the full set of models can be found, integrative modeling facilitates assessing the input information and output models in terms of precision and accuracy. |
| *planning experiments* | Integrative modeling provides feedback to guide future experiments, by computationally testing the impact of hypothetical datasets. As a result, experiments can be chosen to best improve our knowledge of the assembly. |
| *understanding data errors* | Data errors present a challenge for all methods of model building. Integrative modeling can detect inconsistent data as no models will exist that fit all the data. In addition, integrative modeling facilitates the application of more sophisticated methods for error estimation, such as Inferential Structure Determination [[20](#_ENREF_20)]. |

**Table 3. Key aspects of the Integrative Modeling Platform.**

|  |  |
| --- | --- |
| *provides a high level interface against which to write integrative modeling applications* | A high level interface in Python and C++ limits the amount of code that needs to be written, debugged, maintained, and documented [[42](#_ENREF_42)] when modeling a particular biological system. Thus, a third party can more easily add new data, tweak the representation, and improve the sampling scheme. |
| *allows easy representation of molecules at a variety of resolutions* | When modeling large assemblies, we often do not have enough information to warrant a fully atomic or even a residue level model, so being able to handle coarse-grained as well as atomic models is essential. |
| *allows almost any type of information to be used in modeling* | We need to use all available information to maximize accuracy, precision, completness, and efficiency of modeling efforts. |
| *is easily extendable* | As a single research group cannot write the code to support all information types, any developer must be able to easily add and distribute support for new information types (*eg*, by implementing a single C++ or Python class). IMP is structured as a collection of modules, each of which groups together functionality based on, for example, a particular type of information and identity of the authors; these modules encompass source code, documentation, parameter files, authorship, *etc*. Since modules are self-contained, they can be developed and distributed separately from the main IMP code, while still benefiting from the IMP infrastructure. As a result, it is easy to build on other developers’ code and methods. |
| *is easy to obtain and run* | IMP is freely downloadable, modifiable, and distributable by others. It supports all the common operating systems (Mac, Windows, and various Linux distributions) and provides a complete, self-contained environment that is designed to make it easy to find an appropriate historical version when reproducing old results. IMP is distributed under an open source license (LGPL) so that users can build it on unsupported platforms. |
| *provides higher level tools for solving common modeling problems* | These tools include Restrainer for converting proteomics data into spatial restraints on the configuration of multisubunit assemblies [[8](#_ENREF_8)]; MultiFit for assembling multiple subunits based on an EM density map, proteomics data, and molecular docking [[14](#_ENREF_14)]; and FoXS for computing a SAXS profile of a given structure [[55](#_ENREF_55)]. MultiFit and FoXS are also distributed with Chimera [[29](#_ENREF_29)], which provides a visualization of the IMP inputs and outputs. |
| *reduces the maintenance burden on integrative modeling application developers* | By insulating application developers from the details of the platform on which the application is being run, the platform makes scripts easier to maintain and run. |
| *supports the full pipeline from data to publishable results* | The encoding of the modeling process makes it easier to understand what was done and experiment with minor changes to any stage of the process. When coupled with standard version control software such as Subversion (<http://subversion.apache.org>) or Git (http://git-scm.com), it also provides a log of all dead ends encountered along the way. |

IMP was designed to fill a number of needs that are unmet by the other available software tools.