Title

Physical Binding Site Modeling for Quantitative Prediction of Biological Activities

Author

Varela, Lawrence Rocco

Publication Date

2013

Physical Binding Site Modeling for Quantitative Prediction of Biological Activities

by

Lawrence Rocco Varela

DISSERTATION

Submitted in partial satisfaction of the requirements for the degreee of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Dedicated to my parents Larry Varela and Angela Fletes.

Acknowledgments

First and foremost I would like to acknowledge my parents, Larry Varela Sr. and Angela Fletes. I thank my father for instilling in me the value of hard work and education, and taking pride in everything I set to accomplish. I thank my mother for her support and encouragement, especially throughout graduate school. I would like to extend a special acknowledgment to my wife, Gabriela Varela, for her moral support throughout graduate school and the next steps in life.

I consider myself extremely fortunate to have been guided by an exceptional line of mentors during my time at UCSF. I thank my qualifying exam committee members: Andrej Sali, Tanja Kortemme, Patsy Babbitt, and Tom Ferrin. I thank the members in our lab for the constructive conversations and learning experience: Ann E. Cleves, Russell Spitzer, and Emmanuel Yera. I would finally like to extend a very special acknowledgment to my research advisor Ajay N. Jain. My research experience and training in the Jain lab has been instrumental in my academic and professional growth, and I am extremely grateful for Ajay's guidance and support throughout my graduate studies.

Statement from Ajay Jain, thesis advisor, on chapter co-authors:

Chapters 3 and 4 were adapted from the following paper: R. Varela, W.P. Walters, B.B. Goldman, and A.N. Jain. Iterative refinement of a binding pocket model: Active computational steering of lead optimization. Journal of Medicinal Chemistry, 55(20):8926-8942, 2012. Rocco designed the primary experiments, implemented the relevant algorithms, performed essentially all computations, and made the majority of the analyses. Contributions came from all authors on control experiments and

some aspects of data analysis.

Chapter 5 was largely adapted from a paper, to be published: R. Varela, R. Spitzer, A.E. Cleves, and A.N. Jain. A Structure-Guided Approach for Protein Pocket Modeling and Affinity Prediction. Rocco designed the primary experiments, implemented the relevant algorithms, performed all computations, and made the majority of the analyses. Contributions came from all authors on control experiments, aspects of automated protein binding site alignment, and some aspects of data analysis.

Abstract

Physical Binding Site Modeling for Quantitative
Prediction of Biological Activities

Lawrence Rocco Varela

Computational approaches for binding affinity prediction are most frequently demonstrated through cross-validation within a series of molecules or through performance shown on a blinded test set. Here, we show how such a system performs in two realistic applications: 1. An iterative, temporal lead optimization exercise, and 2. A hybrid strategy that leverages diversified information as input. In the first evaluation, a series of gyrase inhibitors with known synthetic order formed the set of molecules that could be selected for synthesis. Beginning with a small number of molecules, based only on structures and activities, a model was constructed using the newly developed Surflex-Quantitative Modeling (QMOD) approach. Compound selection was done computationally, each time making five selections based on confident predictions of high activity and five selections based on a quantitative measure of three-dimensional structural novelty. Compound selection was followed by model refinement using the new data. Iterative computational candidate selection produced rapid improvements in selected compound activity, and incorporation of explicitly novel compounds uncovered much more diverse active inhibitors than strategies lacking active novelty selection. For the second evaluation we present a hybrid structure-guided strategy that combines molecular similarity, docking, and multiple-instance learning such that information from protein structures can be used to inform models of structure-activity relationships. The Surflex-QMOD approach has been shown to produce accurate predictions of binding affinity by constructing an interpretable physical model of a binding site with no experimental binding site structural information. Here we introduce a methodological enhancement to integrate protein structure information into the model induction process in order to construct more robust physical models. The structure-guided models accurately predict binding affinities over a broad range of compounds while producing more accurate representations of the protein pockets and ligand binding modes. Structure-guidance for the QMOD method yielded significant performance improvements, especially in cases where predictions were made on ligands very different from those used for model induction.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Information retrieval, processing, and interpretation plays a critical role in the acceleration of biological discoveries. Over the last decade there has been concerted effort towards the expansion of biological data. We have seen the completion of the human genome project.[1,2] The 1000 Genomes Project Consortium has recently completed efforts to provide deep characterization of human genome sequence variation[3] of 1000 anonymously selected individuals from different ethnic groups and geographic regions. An additional 66,145 structures have been deposited in the Protein Data Bank[4] since 2003. The ChEMBL database now provides approximately 1.4M compound records of bioactive drug-like small molecules, covering 9,570 targets and summarizing research findings of over 48K publications.[5] Computational sciences play a necessary and invaluable role in generating and testing hypothesis from data of this magnitude. Hypothesis testing and validation via direct experimentation can be an expensive and time consuming challenge. Computer technology provides an inexpensive means of efficiently reducing efforts for direct experimentation.

The work presented here describes the research and development of computational methods for addressing central challenges in drug development and design. This work applies machine learning to molecular data to improve methods for quantifying protein-ligand binding affinity. A unique aspect of machine learning in the area of binding affinity prediction is that the precise relationship of a ligand bound to a protein is not known. This uncertainty presents its self as free or partially constrained variables that define the conformation and alignment of a ligand within a protein pocket. Formally, this presents a problem of multiple-instance learning.[6,7] The work presented here has direct applications within drug development, yet the underlying techniques described herein may be useful in any multiple-instance learning problem.

This dissertation begins with an overview of the drug development cycle in Section 1.1. A description of the underlying principles of protein-ligand binding will be presented in Section 1.2, followed with an overview of physics-based methods for scoring protein-ligand interactions (Section 1.3). This chapter closes with a brief review of machine learning (Section 1.4) followed by summarizing comments.

The remainder of this dissertation is organized into chapters describing the general field of binding affinity prediction (Chapter 2) and the contributions of the work presented here. Note that aspects of the underlying theory and algorithms for molecular similarity, docking, and QSAR presented in Chapters 1 and 2 have been excerpted or paraphrased from a manuscript to be published as a book (used with permission).[8] Chapter 3 describes the underlying method of our work in quantitative binding affinity prediction and lays the foundation for the technology used in the subsequent chapters. Chapter 4 describes methodological enhancements and applications in iterative model refinement and guidance in molecular design. Chapter 5 describes further methodological improvements that integrate protein structural information in the model induction process to derive more accurate and widely applicable models. Final remarks and insights of future directions are discussed in the concluding chapter.

1.1 Drug Development Cycle

The financial feasibility of new drug development depends on the expected costs and returns to research and development (R&D). Capitalized preclinical and clinical costs per investigational biopharmaceutical compound is reported as ranging from tens to hundreds of millions of dollars. Table 1.1 shows data collected on capitalized costs of investigational biopharmaceutical compounds.[9] Majority of the expenditures are typically consumed during preclinical research (right row, Table 1.1). When R&D costs are substantial it is critical to examine approaches that could reduce those costs. Innovation throughput and eventual benefit to patients hinges on new drug development productivity, and the improvement thereof.

Table 1.1: Capitalized Preclinical and Clinical Period costs per Investigational Biopharmaceutical Compound.[9]

Testing Phase	**Expected Out-of-Pocket Cost ($)**	**Phase Length (mos.)**	**Monthly Cost ($)**	**Start of Phase to Approval (mos.)**	**End of Phase to Approval (mos.)**	**Expected Capitalized Cost ($)**
Preclinical	59.88	52.0	1.15	149.7	97.7	185.62
Phase I	32.28	19.5	1.66	97.7	78.2	71.78
Phase II	31.55	29.3	1.08	78.2	48.9	56.32
Phase III	45.26	32.9	1.38	48.9	16.0	60.98
Total						374.70

The drug discovery life cycle is commonly considered to be a linear process. New targets (typically proteins) are identified through knowledge of a particular biological process associated with particular disease. Drug-like compounds in chemical libraries are typically tested in high-throughput screens (HTS) for their ability to bind to or modulate the target of interest. Compounds revealing acceptable levels of activity are selected, and subsequently optimized through further testing screens to produce "leads" that have the required pharmacokinetic properties. Leads revealing the required efficacy in *in vivo* disease models are further optimized into clinical drug candidates, followed by subsequent testing in human clinical trials.[10]

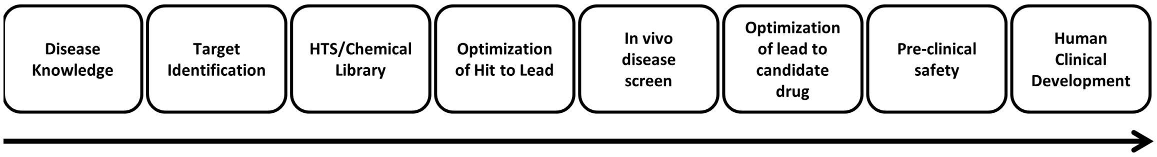


Figure 1.1: The "standard model" of the drug discovery life cycle is considered to be a linear process. Target indentification is followed by high-throughput screens (HTS) of drug-like chemical libraries. Compounds with acceptable levels of activity are subsequently optimized through testing in further screens to reveal leads that have the required pharmacokinetic properties. Leads showing acceptable efficacy in *in vivo* disease models are further optimized into clinical drug candidates. Successful candidates are then tested in human clinical trials.[10]

Early phases of drug discovery proceeds largely by trial and error. Typically, several thousands of compounds are synthesized for each candidate that finally becomes a drug. Each synthesis costs, on average, thousands of dollars and requires a few days to a few weeks of effort. This contributes towards the aforementioned tremendous expense and low productivity of drug discovery. Accurately predicting the biological activity of molecules and understanding the basis of those predictions would make the process more productive. The underlying goal is often to modulate the function of a therapeutic target, typically a protein such as an enzyme or receptor, in attempt to induce or prevent its signaling function of a particular disease of interest. The general strategy is to find a ligand molecule that will bind to the protein's active site. Factors that bind molecules together include van der Waals interactions, the hydrophobic effect, and electrostatics.

1.2 Protein-Ligand Binding

Successful binding activity prediction requires an understanding of the macromolecular interactions that drive protein structure and function, and intermolecular interactions that contribute towards protein-ligand binding. Proteins are linear polymers composed of amino acids joined by peptide bonds.[11] The amino acid sequence, often referred to as the primary structure, gives rise to the secondary structure. Secondary structures are typically represented by local conformational arrangements into alpha helices, beta sheets, or hairpin loops. Composition of multiple secondary structure elements produces the tertiary structure, giving rise to the overall 3D structure of the protein. Multiple polypeptide chains can further come together to organize assemblies of larger functional units. Ligands, as we refer to in this work, are molecules that bind or interact with another through non-covalent forces. In the work presented here, ligands are typically small molecules with molecular weight < 600 Da and the binding counterpart is the protein. The nature of interaction between a ligand and its receptor depends on the delicate balance of physical and chemical forces between them, and the forces between each of the molecules with the solvent environment.

Figure 1.2 illustrates the physical situation to be modeled for docking a ligand to a protein binding pocket. The problem is well-formed in the sense of thermodynamics. The goal is to estimate the difference in free energy between the unbound state at left and the bound state at right. In the case of a "good" ligand, the energy of the bound state is significantly less than that of the unbound state. The following equations are central and give three definitions of the free energy of binding:

Equation 1.1 defines ΔG_{bind} as the difference in free energy between the bound and unbound states of the system. Equation 1.2 relates ΔG_{bind} to changes in enthalpy and entropy. Enthalpic changes arise from the van der Waals and electrostatic interactions made between protein and ligand atoms, replacing those lost with solvent. Entropic

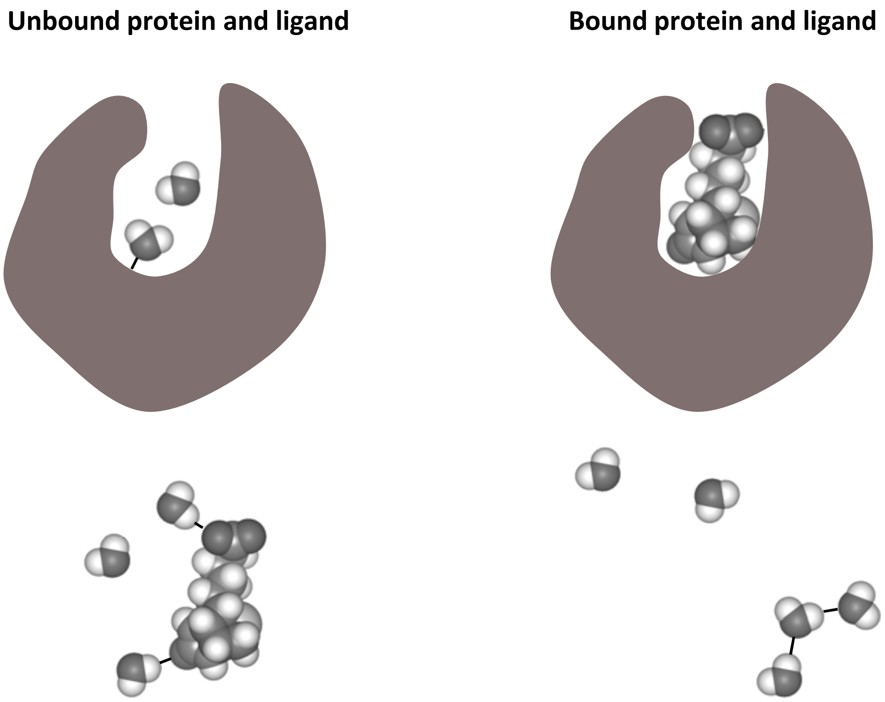


Figure 1.2: Solvated protein and ligand moving from unbound to bound states.

change encompasses the degrees of freedom (translational, rotational, vibrational) lost to the protein and ligand due to binding. Equation 1.3 defines the relationship between ΔG_{bind} and the dissociation constant between the ligand and protein, which can be measured experimentally. Scoring functions for molecular docking generally return values directly related to ΔG_{bind}, whether in kilojoules (kJ), kilocalories (kcal), units of pKd (-log(Kd)), or in arbitrary units suggested to be related to energy.

$$\Delta G_{bind} = \Delta G_{complex} - (\Delta G_{ligand} + \Delta G_{receptor}) \tag{1.1}$$

$$\Delta G_{bind} = \Delta H - T\Delta S \tag{1.2}$$

$$\Delta G_{bind} = -RT\ln(K_d) \tag{1.3}$$

$$A = U - TS \tag{1.4}$$

$$H = U + PV \tag{1.5}$$

$$G = H - TS \tag{1.6}$$

Statistical mechanics offers a formal means to treat systems such as those encountered in non-covalent binding. Equation 1.4 defines the Helmholtz free energy of a system (constant temperature and volume), denoted A. Equation 1.6 defines the Gibbs free energy (denoted G) of a system (constant temperature and pressure), and H denotes the *enthalpy*. The Gibbs formulation is more common in chemistry (see Eqs. 1.1–1.3 from above). For non-covalent protein-ligand binding, there is typically no PV "work" involved, so the formulations are essentially interchangeable.

The Gibbs free energy of binding is given by Eq. 1.7 (see[12] for additional details on derivation).

$$\Delta G_{bind} = -RT \ln \left(\frac{1}{8\pi^2} \frac{\int e^{-(U(r_{PL})+W(r_{PL}))/RT} dr_{PL}}{\left(\int e^{-(U(r_P)+W(r_P))/RT} dr_P\right) \left(\int e^{-(U(r_L)+W(r_L))/RT} dr_L\right)} \right) \tag{1.7}$$

In Eq. 1.7, U denotes energies dependent on the internal coordinates of the species in question, and W denotes solvation energies. The integrals are intended to cover *all* configurations of internal coordinates (r_P, r_L, and r_{PL}), but those with high energy can be neglected. So, with respect to the protein-ligand bound component in the numerator, all reasonably low-energy bound states must be identified and sampled, with energies computed in the presence of solvent. For the unbound components in the denominator, enumeration of energetically accessible solvated states of protein and ligand are also required. Adequate sampling in order to estimate (ΔG_{bind}) based on these theoretical considerations is not feasible for molecular docking applications (even with the advent of inexpensive computational resources). However, all methods for molecular docking rely upon scoring functions that make various approximations to Eq. 1.7, which itself can be seen to embed all of the aspects of protein-ligand binding illustrated in Figure 1.2. For example, configurations in which the bound complex shows favorable energetics relative to unbound states will promote binding. Configurational entropy will disfavor binding to the extent that there are a large number of equally stable configurations in the unbound state but only a limited range in the bound state. Solvation effects favor binding if, for example, solvation energies are high for unbound protein and ligand, which will tend to occur when both the ligand and protein binding cavity are significantly hydrophobic.

One method of predicting these energies using first principles is described in the following section. The empirical approach adopted in this work is foreshadowed in Section 1.4 and fully introduced in Section of the following chapter.

1.3 Physics-Based Methods

Approaches grounded in physics are represented by molecular mechanics treatments to the scoring problem and are exemplified by the pioneering work of Blaney and Kuntz with the DOCK program, further refinement of DOCK by groups such as that of Shoichet, and Abagyan's ICM method.[13–15] The molecular mechanics approach consists of terms for covalent forces (bond lengths, bond angles, and torsions), but for the purposes of modeling protein-ligand binding, we will discuss the non-covalent forces. Due to the difficulty in modeling the entropic aspects of ΔG_{bind}, these methods frequently ignore the entropic components or treat them in a reduced fashion. The key equation in the classical treatment involves pairwise atomic contacts between the ligand and protein, consisting of a Lennard Jones potential and a Coulombic term.

$$\Delta G_{bind} \approx \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right) \tag{1.8}$$

The A and B parameters are defined for each pair of different atom type combinations, R is the distance between atomic centers, q is the partial charge on each atom, and ϵ is the dielectric constant. Precise choice of each of these parameters can have very substantial effects on the behavior and performance of the scoring function. Owing to the difficulty of explicit solvent representation in the context of highthroughput computation, continuum solvent modeling is common. This has resulted in a good deal of effort developing an appropriate treatment of the dielectric constant. The presence of water as a solvent yields a "screening" effect on the effects of charged interactions at some given distance. A typical value for ϵ in solvent is 80.0, whereas more typical values in proximity to a protein are closer to 2.0 (which is between ϵ of 1 in a vacuum and 4 in many non-polar solvents). Figure 1.3 illustrates these potential functions. Recalling Eq. 1.7, we see that this formulation ignores all terms *except* for the direct protein-ligand enthalpic component at the extremum with respect to ligand pose, essentially only $U(r_{PL})$ for the *inter-atomic* energy terms. All aspects of solvation and entropy are ignored, and conformational strain is typically heuristically limited by exploring ligand conformations through torsional variation that avoids internal clashes.

1.4 Machine Learning

Physics-based approaches for potency prediction are applicable in cases where a reliable, high-resolution structure of the protein target is available. While there have been some encouraging reports of success,[16] the problem remains unsolved, with prediction methods suffering from a lack of accuracy and high computational cost.[17–19]

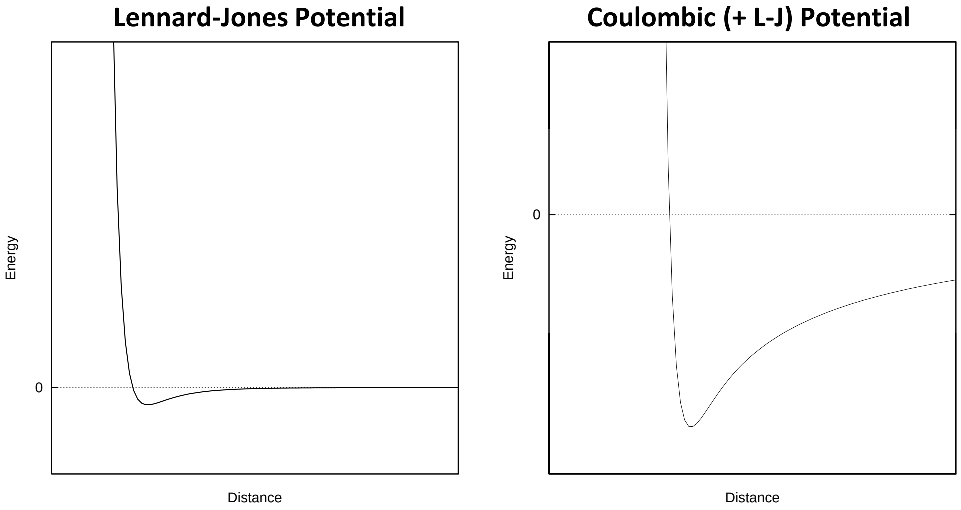


Figure 1.3: Classical potential functions for physics-based scoring functions.

In addition, for large classes of therapeutically relevant targets, high-resolution protein structures are only rarely available (e.g., ligand-gated ion channels, membrane transporters, and membrane spanning G-protein-coupled receptors). For these reasons, constructing predictive activity models based purely on ligand structure activity data has been long-studied in computer-aided drug discovery. In this form, it is a classic machine-learning problem, that of model induction from training data, and it is not amenable to a direct physics-based approach. This section discusses the applicability of machine learning in computer-aided drug discovery. An in-depth discussion on ligand-based approaches to predicting binding affinity will presented in the following chapter (2).

Machine learning is a branch of computer science focused on developing and studying systems that can automatically learn and improve predictive performance through experience. Experience may come in many forms, but is typically encapsulated by data relevant to the task at hand. The typical machine learning problem involves capturing complex relationships between relevant descriptors and observed outcomes.

The hope is that what is learned from a finite training set can then be applied generally to the remaining space. If each example in the training set has a known descriptor or associated target value, this becomes a supervised learning problem. The goal is to create a rational model specialized to that domain that is capable of classifying or predicting future outcomes.

This is a straight forward task when the association between descriptors and outcomes is fairly linear. In drug development, however, the parameters influencing biological activity tend to be numerous and interact in a non-linear fashion (consider the Lennard Jones potential Eq. 1.8). Additional difficulty arises when there is not enough experimental data, also known as the "curse of dimensionality.[20] If M points are necessary to reasonably define a single dimension, then M^N points are required for N dimensions. Thus, if it takes 10 examples to approximate the relationship of a single parameter, then 17 parameters would require 10^{17} examples. Clearly, encoding knowledge in a higher dimensional space requires astute model selection to succeed. Several such models are embedded in the scoring functions of molecular dockers and modeling systems. These are covered in depth in Chapter 2.

Consider now the scenario in Figure 1.4 where our learner has only partial or incomplete knowledge about each training example. Instead of each example being represented by a single feature vector, each might be represented by a set of potential feature vectors of which only one may be responsible for the observed result. The ambiguous nature of training input arises in the domain of activity prediction for drugs.[6, 21] In this example, the object is a ligand and the observed result is the binding affinity of that ligand with the target. The multiple instances are the various conformations (dictated by rotatable bonds, alignments relative to protein) the ligand can adopt within the binding pocket. At ordinary temperatures, the molecule conformation is constantly changing. Only a few of these can provide the ideal interactions necessary with the protein to produce the required binding affinity. Each conformation has a potential energy that is related to its intra- and inter-atomic interactions. The probability that the molecule exists in any particular conformation is exponentially dependent on the potential energy of the conformation according to the Boltzmann distribution. The most probable conformations are lower in energy, and thus more likely to be the correct binding pose.[22] Identifying and retaining low energy poses out of the space of infinite poses as we simultaneously optimize the protein-ligand interaction (via spatial adjustments of a physical pocket model) is a multiple instance problem. An efficient method of handling this learning complexity is described in Chapter 3 and Chapters 4 and 5 as we describe its applications to complex challenges.

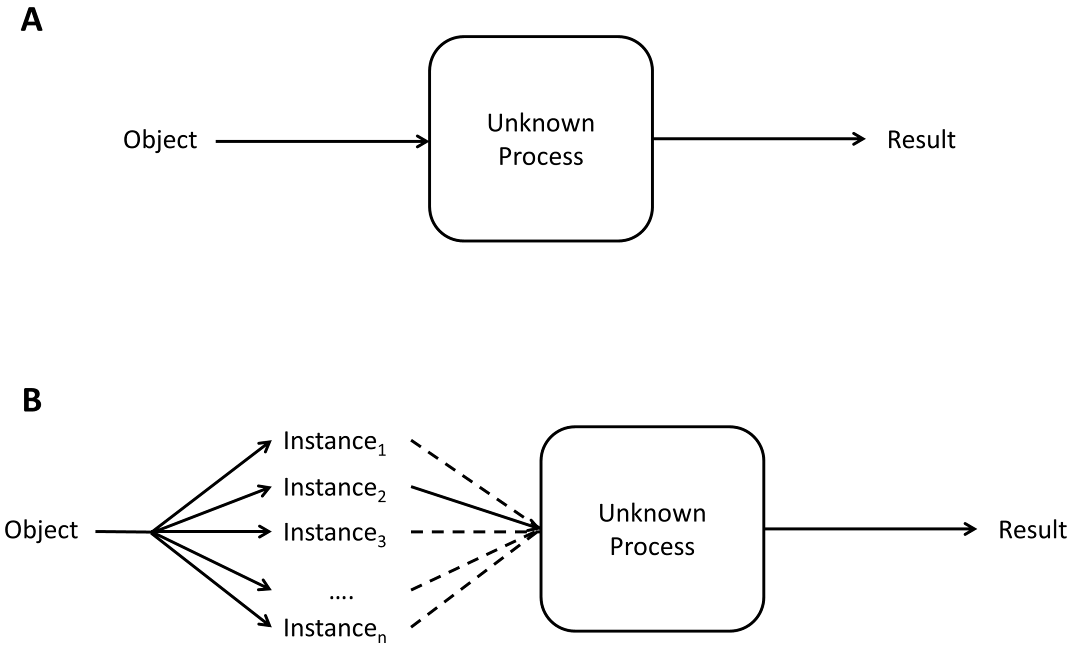


Figure 1.4: Supervised learning: A) Usual scenario, B) Multiple instance scenario.[21]

1.5 Conclusion

This chapter provided context for our work in protein-ligand affinity modeling. The need for increased efficiency in the drug development cycle is motivated in Section 1.1. One step in that cycle where computational methods can be applied to great effect is in early compound design/lead optimization (Figure 1.1) where accurate modeling of the driving forces of protein-ligand binding (Section 1.2) become a critical area in applied research. Here, the game is to accurately predict protein-ligand binding affinity with a high level of accuracy that can distinguish levels of potency within various orders of magnitude. Section 1.3 introduced a physics-based approach towards this challenge, highlighting the complexity and limitations on common systems. Section 1.4 introduced machine learning and its applicability in modeling protein-ligand binding, foreshadowing the heart of the work discussed in the remainder of this dissertation. The next chapter will provide more depth in the field of quantitative binding affinity prediction.

Chapter 2

Quantitative Binding Affinity Prediction

2.1 Introduction

Small molecule drug discovery nearly always involves *specific* non-covalent modulation of enzymes, receptors, transporters, and ligand-gated ion channels. Toward that end, the *a priori* expectation is that if a molecule exhibits its effects against a desired target at very low concentrations (i.e. that the molecule is *potent*), it will tend to exert clinically relevant modulation at that target and few others. This drives a central need to optimize potency against a desired molecular target.

A medicinal chemistry lead optimization project will typically require design and synthesis of several hundred to a few thousand small molecules. The design process seeks to increase or maintain potency (affinity for the desired biological target) while simultaneously addressing aspects of selectivity, solubility, absorption, distribution, metabolism, excretion, and toxicity due to undesirable off-target effects. Potency is just one part of the puzzle, but it is a necessary part, and since it can generally be measured quickly with in vitro assays, it is the subject of direct optimization. Strategies to address other properties require making alterations in either the substituents or basic scaffolding of a chemical series while trying to maintain gains in potency, which may be hard-won. The central importance of the affinity of a ligand for its desired protein target has driven the field of quantitative structure activity relationship (QSAR) modeling. It has also driven biophysical characterization of protein ligand binding (typically X-ray crystallography) coupled with physics-oriented computational means for estimating binding affinities. This chapter will begin with a discussion of the underlying technology required for exploring molecular binding behavior. This includes a quantifiable description for molecular similarity and approaches for conformation and alignment optimization. An in-depth discussion of the empirical scoring methodology will follow. This chapter will conclude with a discussion of quantitative structure-activity relationships, an introduction to our contribution in this field, and closing remarks.

2.2 Molecular Similarity

One of the earliest descriptions of the use of molecular shape in relation to the biological activity of small molecules was provided by Hopfinger.[23] The conceptualization of shape comparison was based on volume overlap of molecules that were modeled as collections of spheres. The concept of spherical volume overlap is the foundational concept of a family of molecular similarity approaches, best exemplified in current practice by the ROCS approach.[24] A separate line of thought characterizes molecular similarity by *surface* overlap, and one of the earliest descriptions of this concept was introduced by Masek et al.[25] In that approach, molecules were characterized as having "skins" of a particular thickness, and the volume of the surface was described by the difference between a collection of spheres with standard atomic radii and one of radii made larger by the skin thickness. Similarity was measured based on the shared skin overlap between two molecules, offering some advantages over volumetric approaches, for example, when comparing molecules of very different overall sizes.

The notion of molecular surface comparison is best exemplified by the Surflex-Sim approach,[26] which owes its provenance to the Compass 3D-QSAR approach.[6]

The fundamental underpinning of widely used 3D molecular similarity approaches makes use of the approximation that a molecule in a particular pose can be thought of as a collection of spheres with radii that depend on atomic type and also may have different chemical properties such as charge or polarity. The fundamental distinctions among metrics revolve around a basic choice between volume comparison or surface comparison. The work presented here is founded on surface-based shape comparisons.

2.2.1 Shape Similarity

A surface-based molecular representation provides two appealing aspects. First, interactions between small molecules and proteins occur between surfaces, and there is a direct relationship between binding free energy and encapsulated hydrophobic surface area of a ligand. Second, as pointed out by Masek et al. with their molecular skins approach,[25] comparison of molecules with different sizes based on shared volume maximization can produce odd results (i.e. embedding a small molecule in the middle of a larger one). While conceptually attractive, the molecular skins approach was computationally demanding. A different approach to capturing molecular surfaces was proposed during the development of the Compass 3D QSAR technique.[6] A collection of observation points was used from which to measure the minimum distance to a molecule's surface, and this distance was compared to a *learned* ideal distance. This basic concept was quickly generalized to define a similarity measure that used a Gaussian reward function.[27] Similarity functions of this type correspond very closely to a *surface* density function formulation of molecular shape, as follows.

$$M_i(r_i) = e^{(-(r_i-\mu_i)^2)/\gamma} \tag{2.1}$$

$$E_k^P(r_k) = e^{(-(r_k - d_k)^2)/\gamma} \tag{2.2}$$

$$R(r) = \left(\sum_i M_i\right)\left(\sum_k E_k^P\right) \tag{2.3}$$

The M_i of Eq. 2.1 are Gaussians with peaks at the atomic surface (set by the atomic radii, denoted μ_i). By itself, the sum over the M_i produces *internal* molecular surfaces in addition to external ones. The E_k^P of Eq. 2.2 define Gaussians on local radial coordinates around each observer point from set P, with peaks at the *molecular* surface (set by the minimum distances from the observers to the molecule, denoted d_k). When γ is chosen carefully, the integral of the product of two molecules' surface density functions R (defined in Eq. 2.3) is very closely approximated by the morphological similarity function used by Surflex-Sim.[26]

Figure 2.1 depicts the volumetric and surface density functions for benzamidine.

The molecule benzamidine was placed in a coordinate frame such that the XY plane bisected the aromatic ring. The surface density function R is depicted, again with the significant density shown with red shading. The green curves indicate the relative value of the density functions along the X axis, penetrating two hydrogen atoms and three aromatic carbons. The surface density function leaves the interior of the molecule with extremely low values, creating a peaked zone that also shows smoothing at saddle points.

In the surface density approach (Eq. 2.3), the density function is the product of two sums, one "lighting up" the surfaces of each atom of a molecule and the other lighting up the surfaces of spheres packed around the molecule. The product produces a function that has significantly non-zero values only at points close to the overall molecular surface, as shown in Figure 2.1. Consider two molecules A and B and one

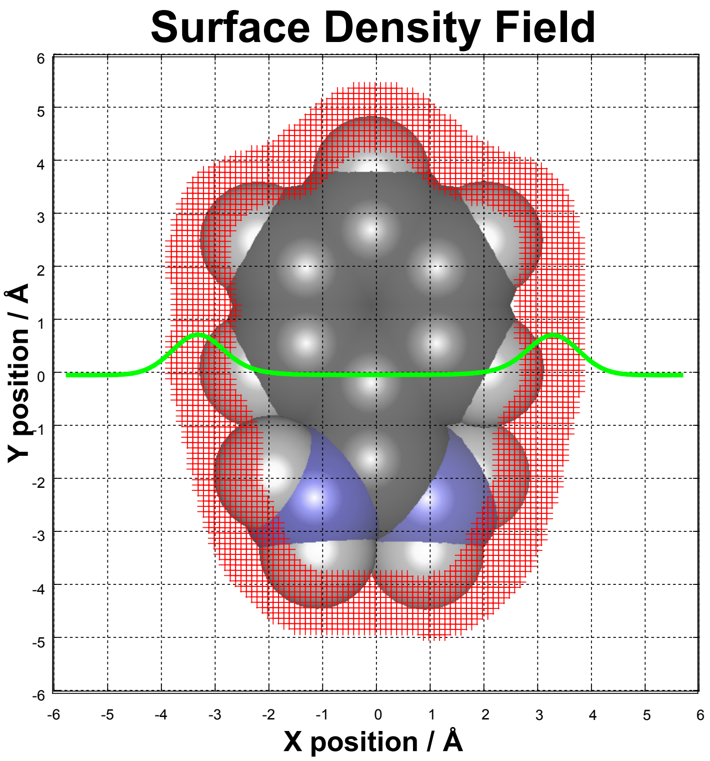


Figure 2.1: Surface-based molecular density functions for benzamidine.

set of "observation" points P, giving rise two the following two density functions.

$$R^A(r) = \left(\sum_i M_i^A\right)\left(\sum_k E_k^{P,A}\right) \tag{2.4}$$

$$R^B(r) = \left(\sum_i M_i^B\right)\left(\sum_k E_k^{P,B}\right) \tag{2.5}$$

Here, the two surface density functions are defined with respect to a *single* set of observation points P. The spheres that "pack" around each of the two molecules A and B share the same centers, but they have different radii, depending on the minimum distance to each molecular surface. One can define a similarity metric in terms of the overlap integral of the product of the two surface density functions.

This function is very closely approximated by the function computed by Surflex-Sim, simplified slightly in what follows.

$$S_k^P(d_k^A, d_k^B) = e^{(-(d_k^A - d_k^B)^2)/\sigma} \tag{2.6}$$

$$S_{A,B}^P = \sum_k S_k^P(d_k^A, d_k^B) \tag{2.7}$$

Here, the Gaussian terms are soft reward functions for concordance of the distances from the observer points P measured to the molecule surfaces of A and B (denoted d_k^A, d_k^B). When σ is roughly twice γ from Eqs. 2.1 and 2.2, the equivalence to the surface overlap integral holds. The intuition behind the metric is simple: when the minimum distances from each observer to each molecule are similar, the molecules must exhibit the same surface shape. The morphological similarity function used by Surflex-Sim[26,28] defines an *infinite* grid of observer points, with weights set such that a shell of observer points around each molecule subject to a comparison contribute. In practice, finite observer sets having significant weight are selected, and alignment optimization is done using the set constructed with respect to the query ligand. Similarity scores are reported using that set, another set constructed with respect to the final aligned new ligand, and a merger of the two.

2.2.2 Electrostatic Similarity

The foregoing has addressed only the molecular shape aspect of 3D molecular similarity, but the degree to which the polar moieties of two molecules are congruent is also important to consider. The surface-based similarity approach of Surflex-Sim explicitly models hydrogen bond donors and acceptors, formal charges, and the directionality of polar interactions. All molecule atoms are labeled as being hydrophobic, hydrogen bond donors, hydrogen bond acceptors, or formally charged atoms (charge is automatically delocalized where needed, as in carboxylates). From each observer point, in addition to computing the distance to the closest atom of any type, which gives the pure shape of the molecule, distances are also computed to nearest polar positive atom (this includes donors and atoms assigned positive charge) and polar negative atom (acceptors and atoms assigned negative charge). For a particular observer point, directionality is treated by comparing the preferred interaction direction of a polar atom to the vector from that atom's surface to the observer. The coincidence of these directions is combined with formal charge to yield a strength value. Similarity from each observer point's perspective is maximized when both molecules produce the same distances *and* strengths. Partial similarity results from, for example, a hydrogen bond donor being aligned with the hydrogen of a charged nitrogen.

Figure 2.2 shows the optimal alignment of two competitive muscarinic antagonists using the full Surflex-Sim function, including both shape and electrostatics. The 2D structures are shown above the optimal mutual alignment, with the quinuclidine derivative shown in cyan carbons. On the right shows the individual molecules in the same pose with atomic surfaces and rods that indicate surface areas that have high similarity between the two. Green rods indicate high shape similarity, red indicate high similarity in the hydrogen-bond acceptor position and directional preference, and blue indicate concordance of the charged amine hydrogen atoms. This is an example where very high 3D similarity (0.82 on a scale of 0 to 1) obtains from molecules having different underlying scaffolds.

2.3 Conformation and Alignment Optimization

All 3D similarity methods are dependent on the conformations of the molecules to be compared, and all in common use are dependent on the particular alignment of

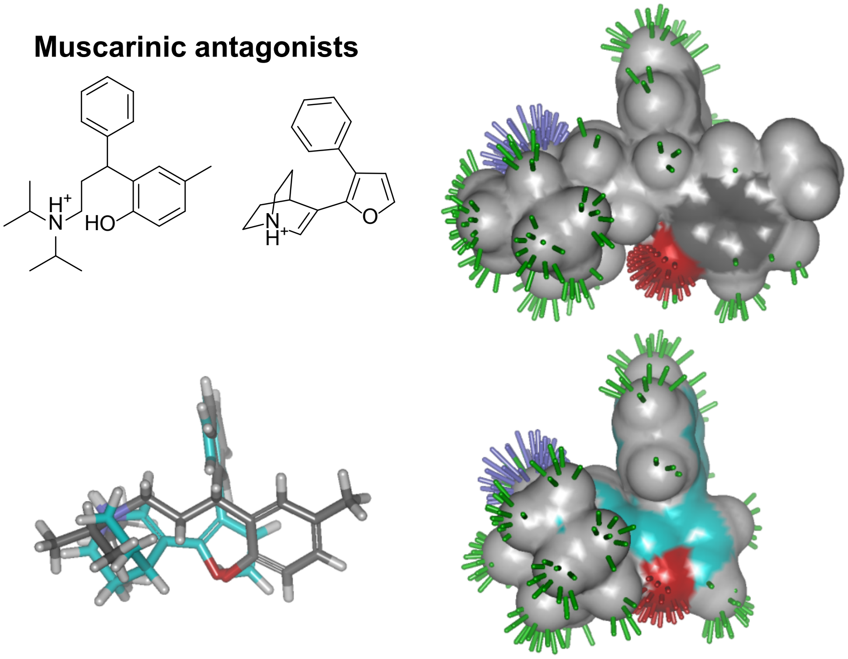


Figure 2.2: Optimal alignment of two muscarinic antagonists using surface shape and polarity.

the molecules as well. This property derives from their direct relationship to what physically makes molecules similar to one another in biological systems. Because of this, in order to compare a molecule to a single pose of another, a 3D similarity approach must identify both the conformation and alignment that yields a maximal similarity value.

Typical drugs and drug-like molecules may have a few rotatable bonds (e.g. aspirin, a COX-1/2 inhibitor) but some can have more than ten (e.g. saquinavir, an HIV protease inhibitor, has thirteen). Methotrexate, an old drug and inhibitor of dihydrofolate reductase, has nine rotatable bonds, and it has been used frequently to test search strategies involving conformation and alignment optimization. Even a very coarse sampling of conformational space of three rotamers per torsion produces roughly 20,000 conformations. A more generous sampling of six rotamers produces over ten million conformations. The issue of alignment generates a multiplicative increase in complexity, because the conformations and alignments must be considered together. Assuming translational uncertainty of ± 5Å and a sampling requirement of 1Å, sampling of translational and rotational space for a molecule such as methotrexate requires over ten million rigid alignments. Clearly, brute-force enumeration of all energetically reasonable conformations (with alignments sampled to roughly 1Å) is not feasible even with a very efficiently computed similarity function. There are two basic strategies that have been taken for the conformational problem and two for the alignment problem.

2.3.1 Conformational Optimization

One way to address the conformation question is to search each ligand of interest *independently* of any other consideration and to retain some maximal number of individual conformers. A typical value for the number of retained conformers is 200, allowing for quite complete sampling of molecules with up to five or six rotatable bonds. This does not provide dense sampling for drug molecules such as methotrexate or saquinavir, but in applications such as virtual screening of very large libraries, the speed requirements may necessitate some tradeoffs.

Figure 2.3 illustrates results for agnostic conformation generation. At left, biotin is shown. It has five rotatable bonds, with two ring conformations of reasonable energy. The lowest-energy conformation is shown in a “canonical” alignment to the Cartesian coordinate system, with the molecular centroid at the origin, the largest radial excursion parallel to the Y axis, and the largest excursion in the XZ plane rotated to be within the XY plane. Conformational expansion of biotin, with a maximal sampling of 200 conformers, contains a conformer that is 0.35Å RMSD different from the conformation of biotin bound to streptavidin (PDB code 1STP). For methotrexate, the conformational variation is clearly much more significant, particularly in terms

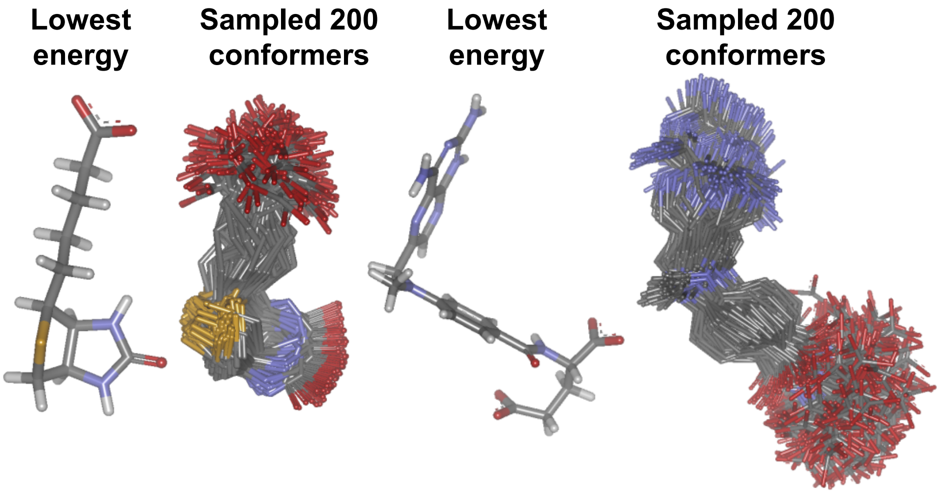


Figure 2.3: Conformational sampling independent of molecular alignment.

of the different conformers that the tail of the molecule can exhibit. In this case, the minimum RMSD from a 200 conformer sample is 1.17Å when compared with the configuration bound to DHFR (PDB code 4DFR). Conformational enumeration of this type typically takes seconds per molecule, and it need be done only once, offering the resulting sampled molecular state for further processing for the negligible cost of retrieving the conformations.

The other approach is to treat the conformational search as *part of* the overall similarity optimization procedure, with strategies to identify only those conformations that are likely to yield good matches to the reference molecule. Such strategies include divide and conquer approaches that fragment molecules into significantly less flexible pieces, search those relatively thoroughly, and either incrementally reconstruct the partial solution or make use of some type of crossover procedure. The idea is that a molecule with, for example, 11 rotatable bonds may be broken into three fragments by severing two torsions. If the torsion-breaks are chosen carefully, the three resulting fragments will each have three rotatable bonds. Coarse sampling of three such fragments yields less than 100 total conformations ($3 \times (3^3)$), and more thorough sampling produces about 600 ($3 \times (6^3)$). This compares to roughly 200,000 for coarse sampling of the unfragmented molecule. If it is possible to produce reasonable alignments for the conformations that will form part of a close to optimal solution *independently* of one another, then such a strategy can be very effective in identifying high similarity poses. The broken torsions receive their configurations during a reconstruction process.

The question of whether independence is a good assumption or not is essentially an empirical one. However, certain molecular types present known difficulties, such as a molecule with a small central scaffold from which emanate multiple substituents that can clash with one another. In such a case, independent alignment of the substituents may lead to incompatible geometries for successful reconstruction of a high similarity final pose. Crossover procedures that recombine full solutions instead of incrementally constructing partial solutions can be effective even in these cases. Such procedures were developed initially for docking,[29,30] and they have been adapted for similarity optimization as well.

2.3.2 Alignment Optimization

Similar to the challenge of conformational optimization, the problem of alignment falls into fast sampling approaches that attempt to make more clever choices based on the context of the particular molecules in question.

Figure 2.4 shows a strategy for generating alignments of a conformation of one molecule A onto a query conformation B. In the example, the query conformation of biotin (as bound to streptavidin in PDB code 1STP) has been canonically aligned within the Cartesian frame. In a case where one has a highly similar molecule (here biotin is shown as the ligand to be aligned) *and* the conformational sampling is adequate, simple canonicalization of A can yield a result very close to optimal. Here,

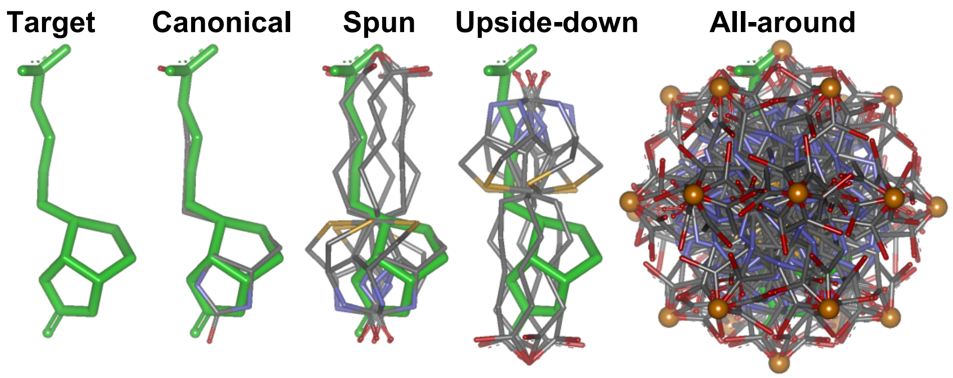


Figure 2.4: Canonicalization of alignment and sampling thoroughness.

beginning with a randomized initial biotin conformation, sampled agnostically as described above, each of the conformations is placed in the canonical alignment to the Cartesian frame. The conformations include the one shown in Figure 2.4, which is clearly very close to the identity alignment. The alignments are generated without any computation of or consideration of molecular similarity. In cases of molecules with highly similar shapes, very limited alignment sampling of a reasonably thorough conformational sample can produce excellent results very quickly. This requires only limited evaluation of the similarity function and local optimization in order to produce close to optimal molecular overlays.

As molecules begin to differ, the major axes may align well, but minor axes may not line up correctly, in which case generating a flip or a sampled spin around the axis will help identify high-similarity alignments. Of course, the same principle applies for the major axis, and vertical flips may also be required. Figure 2.4 shows the systematic spinning of biotin around its major axis. It may also be necessary to consider the "flip" of each conformation along its major axis. But as molecules become less similar, finding close to optimal alignments becomes increasingly challenging. The problem can be treated generally by choosing some minimal spatial sampling interval and ensuring that alignments will be generated to cover the chosen density. This can be done efficiently by treating a conformation as an elliptical body to be spun around its long axis and rotated such that its poles mark out a spherical tessellation.

At the right of Figure 2.4, the uniform sampling of major axis rotation and of axis-direction to sphere tessellation is shown for a single conformation of biotin. In such a sampling, the "nose" of each conformation of the molecule explores each point on a uniformly sampled sphere and also spins around its own axis evenly. One further complication arises with molecules of very different size. An assumption of centroid correspondence may be very poor, and this can be avoided by additional sampling (e.g. along the major axis of the molecule to be aligned). One approach to ameliorate this alignment complexity may involve a coarse sampling within this type of scheme (e.g. the major and minor axis flips of the canonical alignment).

Another approach for addressing alignment optimization is to make specific choices of alignments on the basis of the particular conformation or conformational fragment to be aligned. For example, one can seek to identify matching triplets of points between a conformation of the molecule to be aligned and the query conformation. By ensuring that the triangle edge lengths are similar and possibly that the characteristics of the corresponding points themselves are similar, alignments can be produced quickly by identifying the rigid-body transform that minimizes the least-squares distance differences between the corresponding triangle vertices. The Surflex-Sim approach offers both an aggressive "blind" alignment enumeration as well as a procedure that generates alignments by making correspondences between observer points of each molecule based on conformation-specific information about the molecule being aligned.

2.4 Scoring Function Methodology

There are three broad classes of scoring functions in wide use in the molecular docking field: 1) functions based directly on the theoretical physics that underlie molecular mechanics force-fields; 2) those based on knowledge of contact preferences and related to the statistical physics approach that employs potentials of mean force (PMF); and 3) empirical methods where protein-ligand complexes of known structure and binding affinity are used to directly estimate the parameters of the scoring function. An introduction to physics-based methods was provided in Section 1.3. Knowledge-based methods are based on an idea in classical statistical physics, where one can use observed distributions of geometries in order to deduce the potential that gave rise to the observed distribution. These methods are typified by PMF and DrugScore,[31,32] with the PMF approach representing an alternative scoring function for DOCK and other programs. Related work on such potentials has also been influential in protein folding.[33]

The work presented here builds on a different approach that makes use of an empirically derived function that relates to the processes driving protein-ligand binding. Some of the earliest influential work was by Bohm, resulting in the scoring function used in LUDI.[34] The idea is straightforward. Define a function composed of terms that are related to processes that underlie ΔG_{bind}, and estimate the function's parameters based on experimentally determined protein-ligand complexes with known affinities. Bohm's approach had terms for hydrophobic contact, polar interactions, and entropic fixation costs for loss of torsional, translational, and rotational degrees of freedom. The scoring function used in Hammerhead and Surflex-Dock borrowed heavily from the approach of Bohm.[35–38]

$$\sigma(x,y) = e^{\frac{-x^2}{y}} \tag{2.8}$$

$$\omega(x) = \frac{1}{1+e^x} \tag{2.9}$$

$$\gamma(x) = (\min(0,x))^2 \tag{2.10}$$

$$r = \left((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2\right)^{1/2} - R_i - R_j \tag{2.11}$$

$$d = \omega(-n_3(b_{ij} \cdot v_i)(b_{ij} \cdot v_j) - n_{10}) \tag{2.12}$$

$$q = (1 + n_{11}c_i)(1 + n_{11}c_j) \tag{2.13}$$

$$S(r) = l_1\sigma(r + n_1, n_2) + l_2\omega(n_3(r + n_4)) + l_3\gamma(r + n_5) \tag{2.14}$$

$$P(r) = l_4\sigma(r + n_6, n_7)(d)(q) + l_5\omega(n_3(r + n_8)) + l_3\gamma(r + n_9) \tag{2.15}$$

$$R(r) = l_6\sigma(r + n_{12}, n_{13})(d)(q) \tag{2.16}$$

$$pK_d = \sum_{i=1}^{m}\sum_{j=1}^{n} S(r) + \sum_{i=1}^{m}\sum_{j=1}^{n} P(r) + \sum_{i=1}^{m}\sum_{j=1}^{n} R(r) + (l_7 n_{rot}) + (l_8 \log(MW)) \tag{2.17}$$

The overall function is parameterized in pK_d units and is a combination of Gaussian (σ), sigmoidal (ω), and quadratic (γ) functions of molecular surface distance r. Surface distances (Eq. 2.11) between atoms are defined using fixed radii for each element, with negative values indicating interpenetration. Each atom on the protein and ligand is labeled as being nonpolar (e.g. the H of a C-H,) or polar (e.g. the H of an N-H or the O of a C=O), and polar atoms are also assigned a formal charge, if present. The steric term S (Eq. 2.14) is computed for all pairs of protein-ligand atoms and offers a Gaussian reward for favorable contact and sigmoidal and quadratic penalties for interpenetration. The polar term P (Eq. 2.15) is computed for all complementary pairs of polar protein-ligand atoms, and it has very similar form to the steric term. Note that the Gaussian reward is scaled by a directional attenuation factor (Eq. 2.12) and by a charge scaling term (Eq. 2.13). The term for same-charge interactions (Eq.

Parameter	**Description**	**Value**	**Parameter**	**Description**	**Value**
l_1	Steric Gaussian scale	0.0898	n_1	Steric Gaussian location	0.1339
l_2	Steric sigmoid scale	-0.0841	n_2	Steric Gaussian spread	0.6213
l_3	Steric penetration	-0.9450	n_3	Sigmoid steepness	*10.0000*
l_4	Polar Gaussian scale	1.2388	n_4	Steric sigmoid inflection	$n_1 + o$
l_5	Polar sigmoid scale	-0.1796	n_5	Steric VdW allowance	*0.1000*
l_6	Polar repulsion	-2.5200	n_6	Polar Gaussian location	0.6313
l_7	Conf. fixation	-0.2137	n_7	Polar Gaussian spread	0.3234
l_8	Rigid body fixation	-1.0406	n_8	Polar sigmoid inflection	$n_6 + o$
			n_9	Polar WdW offset	*0.7000*
o	Repulsion offset	0.4880	n_{10}	Polar directionality inflection	0.6139
			n_{11}	Charge scale factor	0.5000
			n_{12}	Polar rep. Gaussian location	0.5010
			n_{13}	Polar rep. Gaussian spread	0.5000

Table 2.1: Summary of scoring function terms in Surflex-Dock.[35, 37, 38]

2.16) is computed for non-complementary pairs of protein-ligand atoms. The interpenetration portions of the steric and polar terms are also computed for intra-ligand atom pairs that are not 1-2 or 1-3 related and have at least one intervening rotatable bond. The final overall score (Eq. 2.17) includes terms for conformational entropy loss and translational and rotational energy loss. Table 2.1 gives the values of the linear and non-linear parameters for the standard Surflex-Dock scoring function.

The original scoring function[35] was parameterized using only *positive* training data, which resulted in very low values for repulsive terms (hard clashes are not typical features of good protein-ligand interactions). Consequently, and *ad hoc* treatment of the parameters for clashes was used. Subsequent work introduced the idea of synthetic *negative* data,[37] where decoy molecules that showed inappropriately high scores were used to estimate values for the repulsive terms, thereby producing the values shown in Table 2.1. The italicized values in the table are constants, so there are a total of 17 real-valued parameters required for the specification of the Surflex-Dock scoring function. The full multiple-instance estimation regime can be used to tune the Surflex-Dock scoring function for particular targets or families of targets[38] using information that includes complexes with known affinities, protein targets with known binders and decoys thought not to bind, and also protein-ligand complexes with known cognate ligand geometry and decoy cognate ligand poses. Figure 2.5 shows plots of the hydrophobic term and the polar term for a hydrogen bond (left). The hydrophobic term (bottom curve, solid line) yields approximately 0.1 pK_d units per ideal hydrophobic atom/atom contact. The top curve (dashed line) shows that a perfect hydrogen bond yields about 1.2 pK_d units and has a peak corresponding to 1.97Åfrom the center of a donor proton to the center of an acceptor oxygen. The value was learned based entirely on the empirical data and corresponds closely to the expected range.

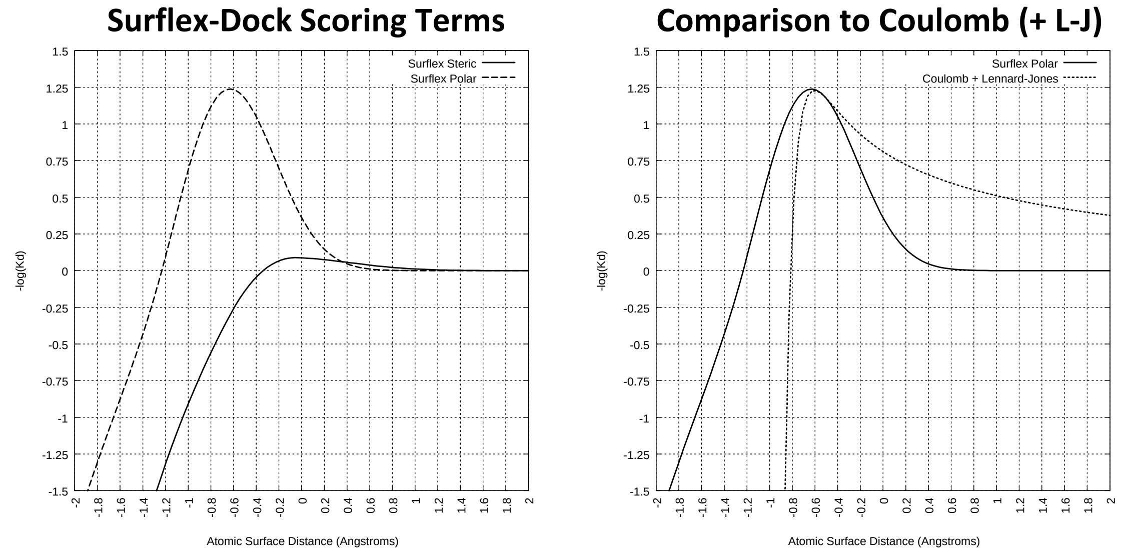


Figure 2.5: Surflex-Dock primary scoring function terms (left) with the polar term compared to a naive physics treatment (right).

Despite the large difference in the value of a single hydrophobic contact versus a single polar contact, the hydrophobic term accounts for a larger total proportion of ligand binding energy on average. This is because there are many more hydrophobic contacts than polar contacts in a typical protein-ligand interaction. This comports well with the understanding that the hydrophobic effect tends to dominate protein-ligand binding. Note that while it may be intuitive to simply equate the empirically derived steric scoring term with the purely enthalpic Lennard-Jones term, the parameters shaping the empirical term have been learned in a regime that tries to estimate ΔG_{bind}. So, the shape and weighting reflect the fact that, for example, a protein-ligand hydrophobic contact *implicitly* suggests the displacement of water.

The preceding discussions of molecular similarity, conformation and alignment optimization, and scoring molecular interactions has introduced the technological framework that has given way to various avenues of computer-aided research that include molecular docking, ligand-based morphological screening, and (relevant to the work presented in this dissertation) quantitative binding affinity prediction to name a few. The following section will discuss quantitative structure-activity relationships, and introduce our contributions to this field and what remains to be the core work of this dissertation.

2.5 Quantitative Structure-Activity Relationships

Quantitative Structure-Activity Relationships (QSAR) is a technique used for quantitatively predicting the interaction between a molecule and binding region of a specific target. A medicinal chemistry lead optimization project will typically require design and synthesis of several hundred to a few thousand small molecules. The design process seeks to increase or maintain potency (affinity for the desired biological target) while simultaneously addressing aspects of selectivity, solubility, absorption, distribution, metabolism, excretion, and toxicity due to undesirable off-target effects. Potency is just one part of the puzzle, but it is a important part, and since it can generally be measured quickly with in vitro assays, it is the subject of direct optimization. Strategies to address other properties require making alterations in either the substituents or basic scaffolding of a chemical series while trying to maintain gains in potency, which in many cases presents a significant challenge. The central importance of the affinity of a ligand for its desired protein target has driven the field of QSAR modeling.

The ligand-based form of the prediction problem is illustrated in Figure 2.6 (left) with a typical series of highly-related analogs, where discrete substitutions at each of a small number of positions on a central scaffold are to be explored. The most widely used methods for addressing the affinity prediction problem are easy to apply in such cases. These methods require a user to provide a three-dimensional alignment of a single conformation of each training and test ligand. For such methods, alignments are typically made using a core ring system and by choosing a single low-energy conformation for each molecule. Given such series of molecules, modeling approaches may be employed that amount to regression analyses of correlations between substituent changes and binding affinity.

Figure 2.6 (right) also shows a more realistic example, where ligands from different chemical series form both the training and testing sets. The top row of compounds are muscarinic antagonists: atropine, azatadine, and oxybutynin. These pre-dated a lead optimization project that involved chemical series exemplified by structures from the bottom six compounds. This case represents the norm in drug discovery, where multiple chemical series have known biological activity and multiple series are under active optimization. In these cases, it is not possible to treat the prediction problem by considering discrete substituents on a common scaffold. There are four critical challenges: 1) choice of the relative alignments and bioactive conformations of ligands (poses) is necessary, but the correspondence of parts between series may

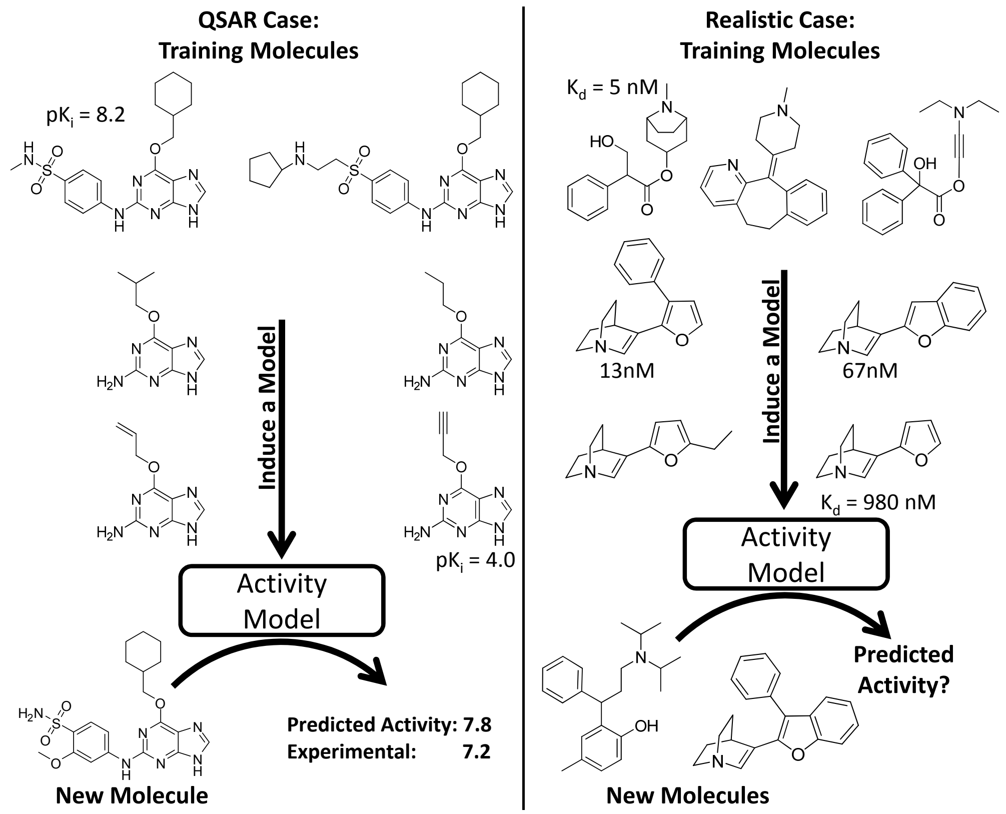


Figure 2.6: A typical case of a congeneric series that can be modeled using common QSAR approaches (left) contrasted with what is seen in practice during a lead optimization project (right).

not be obvious, and the conformations may not be near global energetic minima; 2) the combined effects of substituents may not be additive; 3) changes in ligand structures induce changes in ligand pose relative to a binding pocket; and 4) molecular activity may depend upon the detailed shape of the binding pocket cavity and the complementarity between pocket and ligand.

Most QSAR approaches derive a mathematical relationship between molecular descriptors and activity that is often only tangentially related to the physical process of ligand binding. The most widely used methods for activity prediction include field-based approaches such as CoMFA (and variants such as CoMSIA and Topomer

CoMFA),[39–41] pharmacophoric approaches (e.g. Catalyst and Phase),[29, 42–45] and descriptor-based approaches.[46] These approaches do not meet the four challenges. CoMFA and related methods have three serious limitations. First, they assume fixed alignments that either ignore the effects of substitutions or the potential for the modeled site to influence a ligand's pose. Second, they assume that the effects of multiple substitutions will be additive. Third, they offer no means by which to choose molecular pose based on the modeled binding site. Pharmacophoric methods, while addressing aspects of pose choice and of the influence of an activity model on pose, do not provide models of activity model the effects of the detailed shape of a protein binding pocket. The descriptor-based approaches, as a class, move quite far from physical reality, relying upon correlations between potentially hundreds of descriptors (typically not dependent on molecular pose) and activities. The field of QSAR was the subject of an incisive critique by Stephen Johnson, who suggested that the correlation/causation logical fallacy has been responsible for a great deal of the disappointment in real-world accuracy of QSAR predictions.[47]

Methods such as CoMFA can have utility in organizing and rationalizing large quantities of structure-activity data, but only in cases where ligands share a common scaffold, ligand poses change relatively little with substitutions, the effects of substitutions are close to additive most of the time, and where the substituents offer limited flexibility. Within a limited domain of applicability, such models can also be predictive enough to help facilitate design. Pharmacophoric methods can be of use in helping deduce relative poses of ligands with different scaffolds, but they cannot effectively model the very common case where specific shapes of hydrophobic substituents have a large impact on binding affinity. They also are relatively stronger in identifying *positive* aspects responsible for ligand binding compared with identifying and representing *negative* aspects responsible for non-binding.

Figure 2.7 shows an illustration of the interplay between pocket shape, molecular alignment, and activity. The central scaffold (left), with the unsubstituted furan, is substantially improved by adding a phenyl or by changing to a benzofuran (top and bottom). The compound that combines the two modifications is less potent than either of the parent compounds. A very simple explanation for this is that the binding cavity is simply too small to accommodate both large substituents at the same time. In the general case, additivity should not be expected. Changes in the substituents on a scaffold generally lead to changes in the preferred alignment of the scaffold relative to the binding pocket. If two substituents disagree as to the preference in geometry of the central scaffold, the effects of making a compound with both substituents will not be the sum of the changes in energy of each individually. Nominally, this is a case where standard QSAR approaches can be applied, owing to the common underlying scaffold. However, the most common approaches assume linear additivity of either the effects of substituents or descriptors. Here, the combination of the two substituents from the middle pair of ligands, each of which contributes to a significant potency gain *separately*, together yield *worse* potency when combined. If the example is discretized such that the threshold for active is 100nM, the problem is exactly isomorphic to the XOR problem. In the late 1960s, Marvin Minsky and Seymour Papert[48] used this example to successfully discredit the subfield of machine learning that relied upon linear network models (principally the perceptron approach[49]).

The broad field of QSAR is populated with approaches that rely upon correlations between descriptors and activities that are not based on a physically realistic representation of the protein ligand binding process. When such approaches work, they do so within a narrow domain of applicability. The centrality of ligand potency in drug design, coupled with clear physical requirements for models to address, argues for a methodology that closely resembles the protein-ligand binding processes.

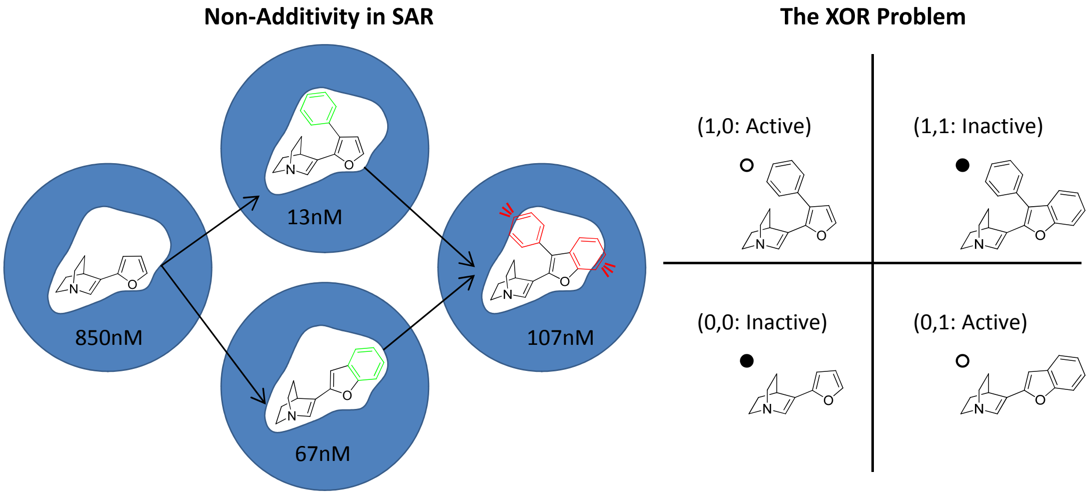


Figure 2.7: The QSAR problem is isomorphic to the XOR problem.

Founded on previous extensive work in surface-based molecular similarity,[26,28,50–53] flexible molecular docking,[30,35–38,54–56] and multiple-instance learning,[6,27,35,37,38,57,58] we have developed an approach to affinity prediction[59–61] that addresses all four of the theoretical challenges listed above. We call the approach Surflex-QMOD ("Quantitative Modeling" built within the Surflex computational platform), or just QMOD. The QMOD approach builds on the earlier Compass approach, which offered solutions to both the pose problem and the detailed shape problem.[6,27,57] However, the models themselves were mathematically abstract and could be physically unrealizable, leading to difficulties in interpretation and visualization. The Compass work introduced the idea of multiple-instance machine-learning. This specifically supported derivation of a virtual binding pocket *at the same time* as the precise relative poses of training ligands were identified. The process iterated between model refinement and pose refinement, where the *model itself* was used to choose the poses. The QMOD approach deviates from Compass by constructing *physical* models that are directly analogous to protein binding pockets.

The QMOD approach transforms the QSAR problem into one of physical binding pocket construction coupled to a fitting process that is similar to molecular docking. By inducing a physical pocket, we naturally obtain the dependence of ligand pose on its structure, non-additivity of substituent activity, and models that reflect the physical complexities and mechanistic behavior of true protein binding pockets. The result is a binding site composed of molecular fragments that can be treated as a target for molecular docking. The binding site model consists of molecular fragments that can account for multiple positions of protein residues. It is not a literal reconstruction of a single configuration of protein residues. New molecules are docked directly into the binding site, with their highest scoring poses serving as the prediction of binding geometry and the corresponding score being the predicted affinity. In deriving a virtual binding pocket at the same time as we identify the relative poses of ligands, the key analogy is that one can treat a computational model of a binding site as one treats a protein binding pocket. We seek the optimal fit of ligands into the binding site. One begins with a guess as to the initial alignment of ligands and then constructs a model of activity that depends on the ligands poses. The model can be thought of as a virtual receptor. Next, poses are explored for each ligand that optimize their interaction with the virtual receptor. The virtual receptor is refined, making use of the new ligand poses, and the process iterates between pose refinement and virtual receptor refinement. As the virtual receptor evolves, the changes in ligand scores due to pose optimization decrease. When the iterative process converges, the final poses of the ligands are optimal with respect to the final virtual receptor. The virtual binding pockets constructed are called pocket models or *pocketmols* for short. Figure 2.8 illustrates the key concepts involved in model construction. Transformation of the QSAR problem into one of pocket construction and ligand fitting directly addresses the four central challenges in a manner that is unique to the QMOD approach within the QSAR field.

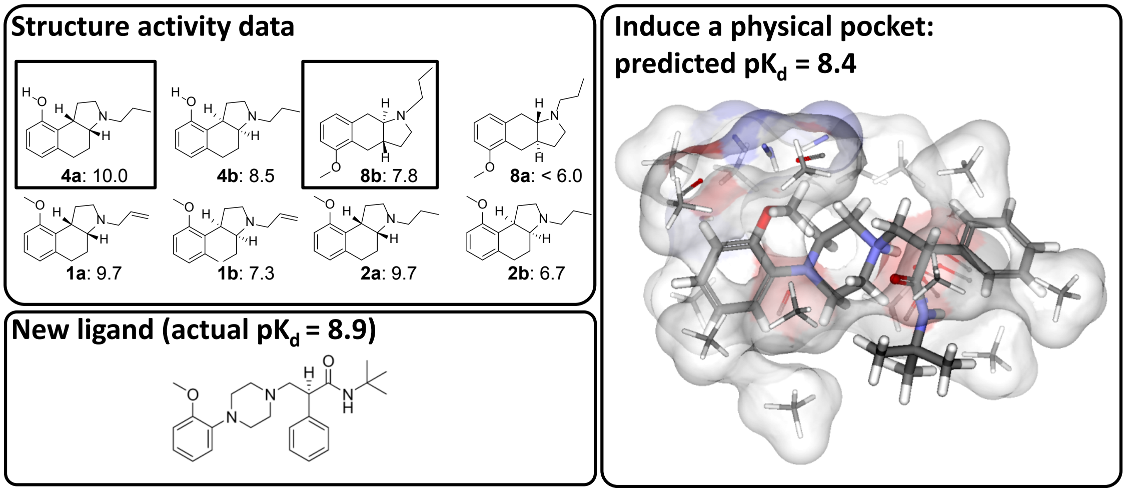


Figure 2.8: The QMOD approach takes structure-activity data and produces a physical model of a binding pocket, to which new ligands can be docked and scored for predictions of both affinity and bound pose.

2.6 Conclusion

The theoretical basis for protein-ligand binding is well understood, but it does not lend itself easily to direct computational simulation that is practical for binding affinity predictions. The challenges of quantifying predictions of ligand binding affinity has many forms. Considerations of molecular representations, conformation and alignment variation, and complexities related to scoring molecular interactions serve as foundational concepts that must be addressed. In this chapter we have introduced our approach to ligand affinity prediction that addresses the major theoretical challenges facing the field of QSAR. The QMOD approach functions by constructing *physical* models that are directly analogous to protein binding pockets. New molecules are flexibly fit directly into the virtual binding site, with their highest scoring poses serving as the prediction of binding geometry and the corresponding score being the predicted affinity. The QMOD methodology forms the core of the work presented in this dissertation, and will be discussed in detail in the following chapters.

Chapter 3

Quantitative Modeling of a Protein Binding Pocket

3.1 Abstract

Computational methods for predicting ligand affinity where no protein structure is known generally takes the form of regression analysis based on molecular features that have only a tangential relationship to a protein-ligand binding event. Such methods have limited utility when structural variation moves beyond a congeneric series or when significant ligand structural novelty plays a hand in molecular design strategies. The focus of this dissertation is centered on developing a novel approach based on the multiple-instance learning, where a physical model of a binding site is induced from ligands and their corresponding activity data. This method is call Quantitative Modeling (QMOD). This multiple-instance learning approach was adopted by the previously established Compass method for ligand affinity prediction. A QMOD model consists of molecular fragments that can account for multiple positions of literal protein residues. In this work we demonstrate the applicability of this method on DNA Gyrase ligands by training on a series with limited scaffold variation and testing on numerous ligands with variant scaffolds. Predictive error was between 0.5 and 1.0 log units (0.7-1.4 kcal/mol), with statistically significant rank correlations.

Accurate activity predictions of novel ligands were demonstrated using a validation approach where a small number of ligands of limited structural variation known at a fixed time point were used to make predictions on a blind test set of molecules discovered at later time point.

3.2 Introduction

All of the molecules used in this chapter were taken from a lead optimization program conducted at Vertex Pharmaceuticals. This program involved the optimization of benzimidazole based inhibitors of the bacterial gyrase heterotetramer.[62] The series used in this study consisted of 426 compounds. The sequence of synthesis and binding activities (pK_i units) for all of the molecules were known *a priori*. The data set was organized in temporal batches with the first 39 ligands used for model induction and subsequent batches organized in groups of 50 molecules. This chapter will focus the basic method for binding pocket model induction and touch on test results for the temporal window of molecules immediately following the training ligands used for model induction. The enzyme target is a type II topoisomerase that alters chromosome structure through modification of double stranded DNA. Antibacterials such as the fluoroquinolones target the non-ATP catalytic sites of gyrase. In contrast, the benzimidazole inhibitors were discovered in a high-throughput ATPase assay of the GyrB subunit. These were then optimized for activity against the ATP-binding site of GyrB, with an eye toward activity against the ATP site of the ParE subunit (topoisomerase IV) as well. Both of these subunits are responsible for supplying energy for catalysis. In the present study, only activity data from GyrB assays were used for modeling and compound selection. Figure 3.1 shows typical examples of structures and GyrB activities from the initial training set. The position 2 substituents of all inhibitors used in this study were either alkyl-urea (e.g. compound **1**) or alkyl-carbamate (e.g. **4**). Structural exploration was predominated by variation in the position 5 substituent of the benzimidazole, with some substitutions also being made at other positions on the central scaffold (especially position 7). The content presented here is in part discussed in our recent publication in the the Journal of Medicinal Chemistry (Varela/Jain[61]).

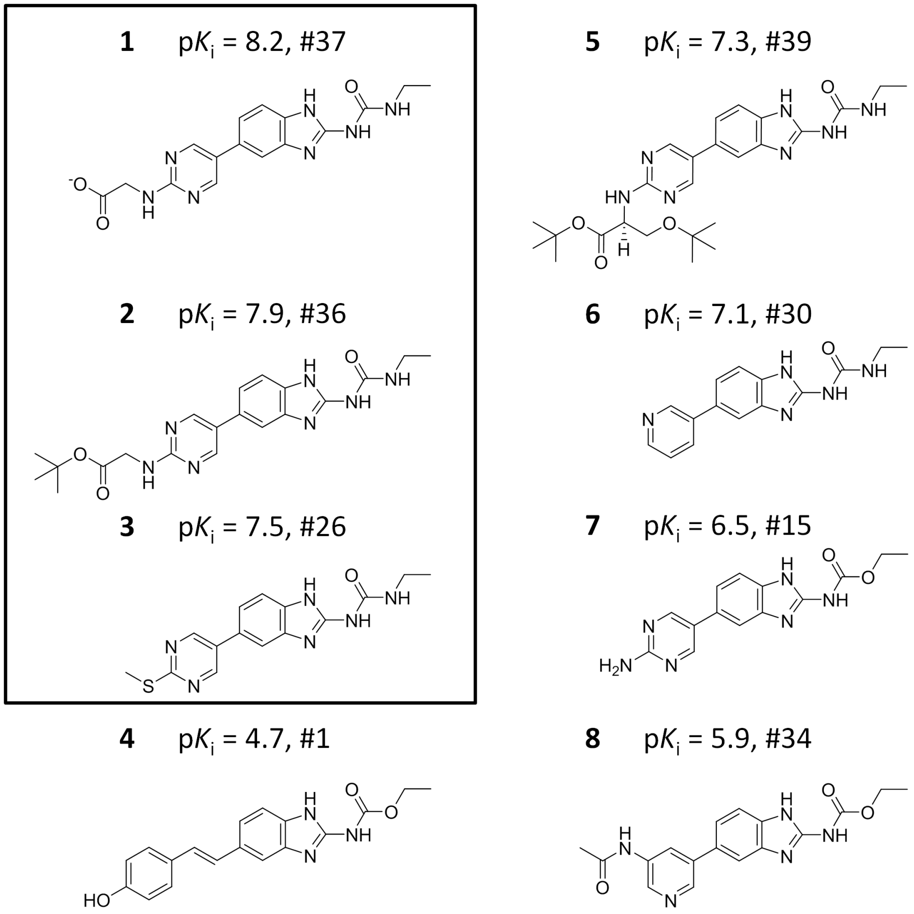


Figure 3.1: Examples of gyrase ligands in the initial training set, which contained the first 39 made from a total of 426 gyrase inhibitors (both pK_i and synthetic sequence number are given). Training molecule activities ranged from a pK_i of 8.2 to 4.7. The 3 most active compounds of the training set (boxed) were used to generate the initial alignment hypotheses.

3.3 Methods

3.3.1 Modeling Procedure

The core computational methods for molecular alignment based upon molecular similarity was presented in Section 2.2 and 2.3 and has been reported in previous papers ([26,28]) and will be described only briefly here. The methods for binding site model induction will be presented in detail. Figure 3.2 shows an example of the steps involved during model induction. Overall, there are six steps to construct and employ a physical binding pocket for activity prediction (a pocketmol):

A) Generation of an initial alignment hypothesis

> Input: structures of the two or three most active training ligands
>
> Output: a mutual alignment that maximizes the overall three-dimensional similarity while minimizing the overall volume

B) Generation of an initial pool of ligand poses

> Input: initial alignment hypothesis
>
> Output: 100-200 alignments for each training ligand are produced

C) Generation of an initial set of molecular probes to form the binding pocket

> Input: pool of poses for each active training ligand
>
> Output: a large set of molecular probes surrounding the ligands, where each probe makes a near-optimal interaction with at least one active ligands pose

D) Selection of an optimal minimal pocket model based on fit to the experimental binding data and model parsimony

Input: the ligand pose pool from step B, the dense set of probes from step C, and activity values for each ligand specified as exact values or inequalities

Output: an optimal set of probes such that nominal interaction scores against this set lie within a specified accuracy

E) Refinement of the pocket model by modifying probe positions interleaved with refining ligand poses

Input: the initial pocketmol from step D, the training ligand pose pool, and the molecular activities

Output: a refined pocketmol with refined ligand poses such that further local optimization of ligand poses against the pocketmol yields little change in scores and where the final scores are close to the experimentally measured ones

F) Testing of new putative ligands within the pocket model

Input: a new molecular structure, the final pocket model, and a selection of training molecules for use in alignment generation

Output: predicted score and pose alternatives for the new ligand using a procedure analogous to docking ligands into a protein active site

The following several paragraphs describe in detail the algorithms and computational procedures used for model building and testing.

3.3.1.1 Ligand Alignment Hypothesis

The initial alignment of training ligands proceeds in two stages. First, a small number of molecules (usually two or three) are selected from which to build an alignment

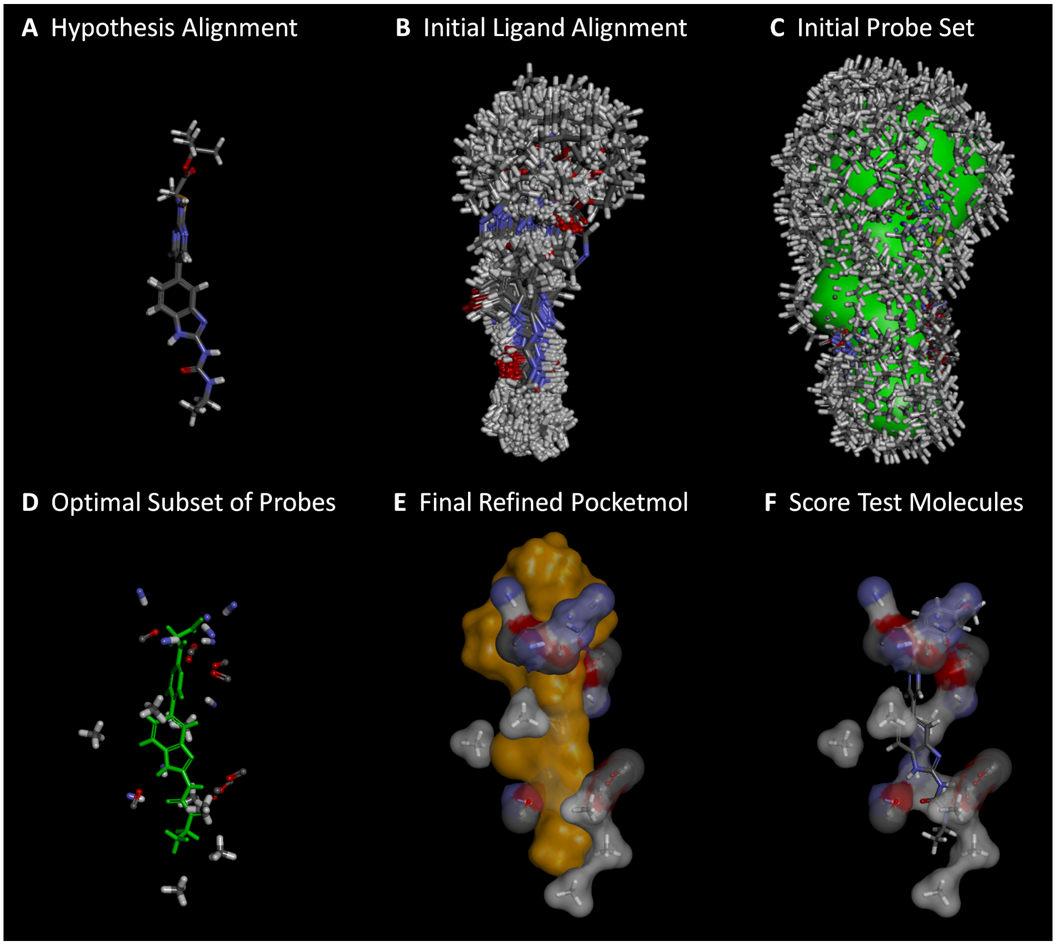

A Hypothesis Alignment
B Initial Ligand Alignment
C Initial Probe Set
D Optimal Subset of Probes
E Final Refined Pocketmol
F Score Test Molecules

Figure 3.2: Derivation and testing of a QMOD pocketmol proceeds in six automated steps: A) an alignment seed hypothesis is constructed from 2-3 ligands; B) 100-200 alignments for each training ligand are produced; C) a large set of probes (many thousands) are created where interactions may exist; D) a small near-optimal set is selected based on fit to experimental binding data and model parsimony; E) probe positions and ligand poses are refined iteratively; F) new molecules are tested by flexible fitting into the pocket to optimize score. The final pocketmol is used in a fixed configuration, but conformational flexibility within the corresponding protein pocket is represented by probes being places in multiple positions.

hypothesis. The methods used for this procedure have been described in previous papers and consist of the morphological similarity algorithm[26] along with the ligand-based structural hypothesis algorithm that depends upon it.[28] The former is a method to compute molecular surface similarity (both shape and polar aspects) between two molecules along with algorithms to enable rapid optimization of conformation and alignment of molecule onto a specific pose of another. The latter uses this procedure in order to produce a joint superimposition of multiple ligands that simultaneously maximizes mutual similarity while minimizing overall volume. Previous in our work in our lab has shown that such superimpositions can yield biologically relevant relative poses and that such joint superimpositions can be used effectively as surrogates for protein structures in virtual screening, even in cases where molecular flexibility is substantial.(Cleves2006,Cleves2008,Jain2000,Jain2004a) Note that it is also possible to select protonation and tautomeric states in this procedure if they are ambiguous.

Figure 3.3 shows the highest-scoring alignment hypothesis of the boxed pair of ligands from Figure 3.1, which served as the seed alignment for the model induction process.

The benzimidazole central scaffold and adjacent urea are in tight alignment, with the pyrimdine rings all aligned such that common polar moieties towards the top of the ligands could interact with a common putative hydrogen bond donor near the cyclic nitrogens or adjacent oxygens. Figure 3.4 depicts the quantitative differences and similarities between these three molecules as computed by the morphological similarity method. The differences include the lack of a donor attached towards the top of the pyrimidine ring of ligand **3**, and an additional steric envelope derived from the tert-butyl on molecule **2** compared to **1** and **3**. The depiction of the similarities shows strong surface concordance over the entire ligands.

The similarity function makes use of observer points (small spheres surrounding each molecule) in order to compare molecular surfaces. In the optimized alignments, the differences are very minor, resulting in a similarity of 0.86 (scale of 0-1) between molecules **1** and **2**, 0.83 between **1** and **3**, and 0.78 between **2** and **3**. These small differences manifest as rods in the upper panels, with length proportional to the

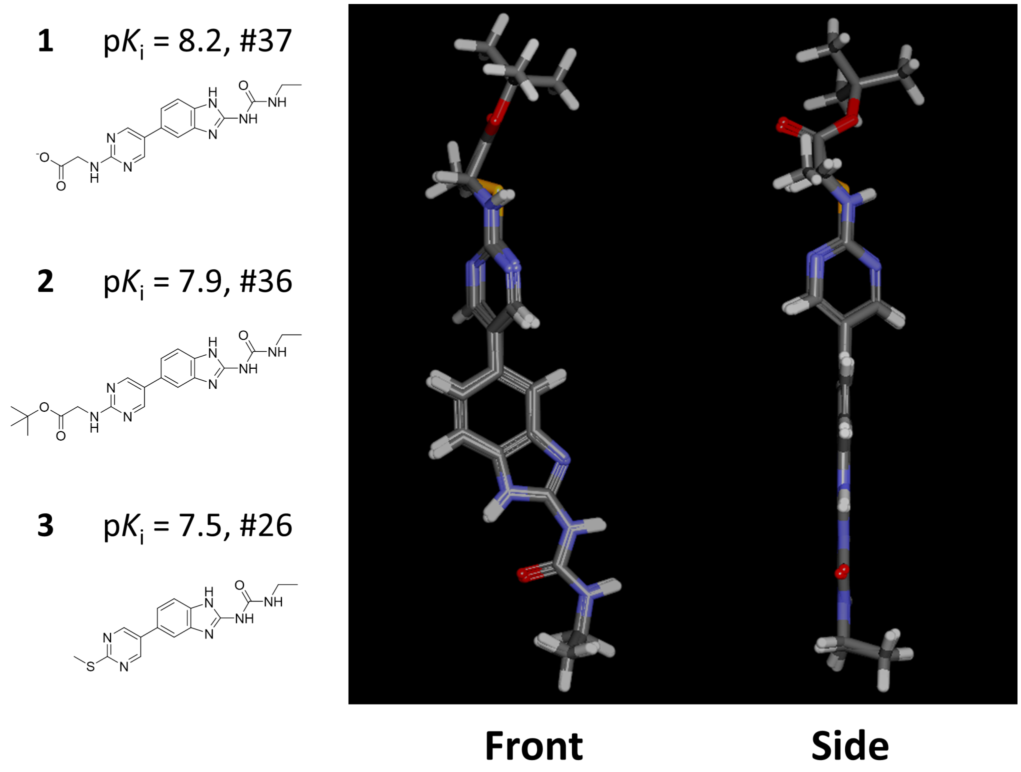


Figure 3.3: The highest scoring mutual alignment of molecule **1**, **2**, and **3** is shown, with the panel at right being the view from the side of the panel at the left. The procedure seeks to maximize joint molecular similarity while minimizing overall volume. The procedure is able to identify joint poses where the steric envelopes are remarkably similar, with the benzimidazole-urea and pyrimidine substructures tightly aligned and the carboxylate and ester oxygens of **1** and **2** being able to accept hydrogen bonds from the same part of space.

magnitude of the difference. Panel A shows the difference and similarities of molecule **1** and **2**. The gray rods indicate steric differences. The longer gray rods stem from the protrusion of the tert-butyl group on **2** relative to **1** at the top of panel A. The surface of molecule **1** is shown in grey skin, and **2** is shown in green skin. Panel B shows the differences between **1** and **3**. The blue rods indicate differences due to positive polar moieties. The longer blue rods are due to a missing hydrogen bond donor on **3** where **1** has a protonated amine. Red rods highlight differences in negative polar moieties. Long red rods are due to missing hydrogen bond acceptor group on **3** relative to the carboxylate on **1**. The other differences are minor. Difference between molecule **2** and **3** highlight both steric and polar differences towards the top of the molecule alignments (not shown). The bottom panels illustrate similarity of the surface of **1** and **2** in A, and **1** and **3** in B. Green rods indicating high shape similarity, red indicating high similarity for negative polar moieties, and blue indicating high similarity for positive polar moieties.

3.3.1.2 Ligand Pose Sampling

This initial alignment hypothesis serves as a template for generation of multiple alternative poses for all training ligands. Since the similarity computation makes no use of activity data and since the relative importance of specific molecular features is not known *a priori*, the gross balance of the importance of shape vs polar characteristics is also not known. The learning paradigm chooses ligand poses as the model of activity is developed, so the issue at the outset is to have a pool of poses for each training ligand that covers the reasonable possibilities. The procedure aligns each training ligand to each of the molecules in the seed hypothesis, using M different weightings of the relative strength of polar versus steric surface features (default weightings: 1.0 and 0.1). For each training ligand, N poses are generated (default 100), which include the N/M highest joint similarity values to the alignment seed hypothesis for each of the M different weighting choices. We have shown previously that the numerical scores of joint similarity to an alignment hypothesis is effective in virtual screening. (REFS 35,28) This amounts to a distinction between ligands with measurable activity (roughly pK$_d$ ¿ 6.0) and nonligands (roughly pK$_d$ ¡ 4.0) that is sufficiently quantitative to yield an enrichment of active compounds at the top of a ranked list.

The poses generated in this manner do provide an adequate pool from which to

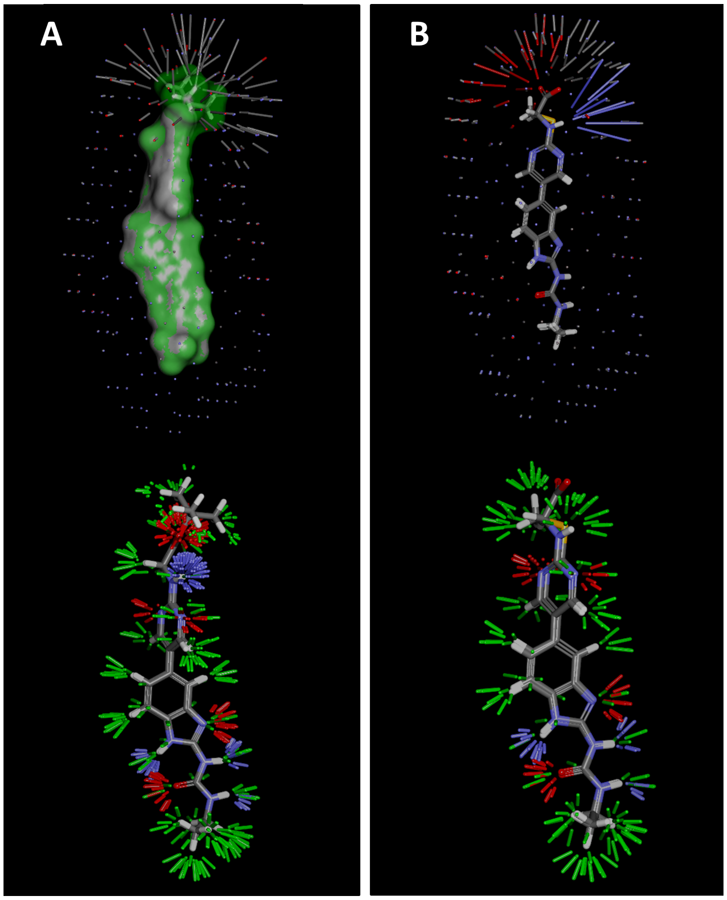


Figure 3.4: The similarity function makes use of observer points (small spheres surrounding each molecule) in order to compare molecular surfaces. In the optimized alignments, the differences are very minor, resulting in a similarity of 0.86 (scale of 0-1) between molecules **1** and **2**, 0.83 between **1** and **3**, and 0.78 between **2** and **3**. These small differences manifest as rods in the upper panels, with length proportional to the magnitude of the difference. Panel A shows the difference and similarities of molecule **1** and **2**. The gray rods indicate steric differences. The surface of molecule **1** is shown in grey skin, and **2** is shown in green skin. Panel B shows the differences between **1** and **3**. The blue rods indicate differences due to positive polar moieties. Difference between molecule **2** and **3** highlight both steric and polar differences towards the top of the molecule alignments (not shown). The bottom views illustrate similarity of the surface of **1** and **2** in A, and **1** and **3** in B. Green rods indicating high shape similarity, red indicating high similarity for negative polar moieties, and blue indicating high similarity for positive polar moieties.

derive a more detailed model of activity. Note that when model construction begins, a particular ligand may have very different alternative poses in the initial pool. Figure 3.5 shows the alignments of training ligands arising from different weights for polar surface features for molecule **5**. For molecules X and Y, the optimal alignments under the different weightings were very similar to one another.

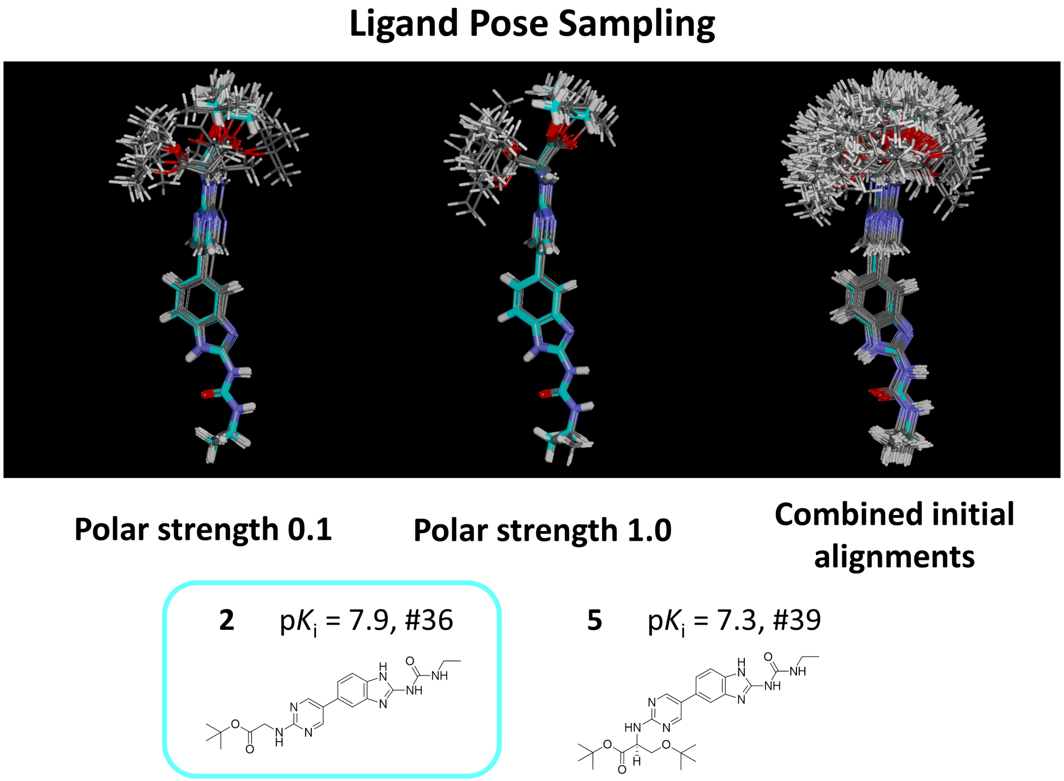


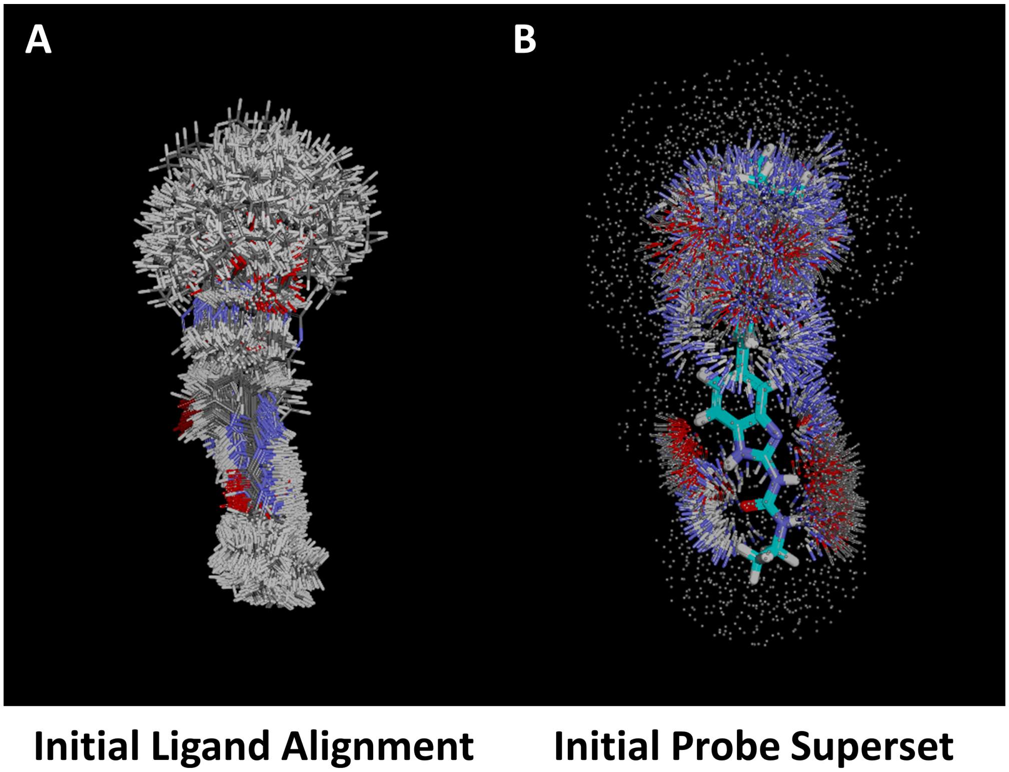
Figure 3.5: Ligand pose sampling is carried out at two polar weightings: 0.1 and 1.0. Shown in cyan is the hypothesis alignment of molecule **2**. Hypothesis molecules **1** and **3** were also used during initial alignment (not shown here for visual simplicity). In atom-colored sticks are the poses sampled for molecule **5** during initial alignment. At left shows a fraction of the poses sampled for molecule **5** at polar weighting 0.1. At center shows a fraction of the poses sampled for molecule **5** at polar weighting 1.0. At right shows the entire pool of poses generated during initial alignment of molecule **5**. Higher polar weighting produces pose samples sharing higher polar concordance near the carboxylate ester near the top of molecules **2** and **5**, while lower polar weighting enables sampling of slightly more discordant alignments in this example. The union of poses resulting from both alignment strategies is used in learning.

3.3.1.3 Probe Generation

Our procedure must accommodate a multiplicity of choices for what the ultimate pocket will look like. In general, we should expect that multiple solutions are possible, all of which may yield equally good fits to the training data. Our approach is to generate a large number of potential probe positions, any of which may be selected and refined in subsequent steps of the procedure. Figure 3.6 illustrates the procedure. We choose a single pose for each active training ligand (defined here as those ligands with $\text{pK}_i > 6.0$) using the highest scoring alignment from the polar surface feature weighting of 1.0 (Figure 3.5). For each ligand, we tessellate its surface using probes of three types: hydrophobic (methane), donor (N-H), and acceptor (C=O). These probes are precisely those used in the Surflex-Dock "protomol" approach for characterizing protein cavities.[28] As in that approach, the probes are subjected to local optimization using the Surflex-Dock inter-molecular scoring function, and probes having high scores are retained unless they are redundant with another probe that has already been accepted. Figure 3.6 shows the resulting probes (with the methane probes devoid of hydrogen atoms for clarity) along with the hypothesis alignments of **1**, **2**, and **3**. The positions of the probes represent reasonable possibilities as to where an interaction from a pocket may lie, but they are too numerous to form a reasonable pocket in a physical sense.

3.3.1.4 Pocket Model Initialization

At this point, we have a pool of poses for each training ligand, both active and inactive, as well as a pool of pocketmol probes. Thus far, activity information has been used only minimally (e.g., by using only the active ligands for probe generation). To select a small set of probes that makes use of the activity data, both active and inactive ligands in the training set are used. We construct a matrix of scores between a pose for



Initial Ligand Alignment

Initial Probe Superset

Figure 3.6: The initial alignments of active ligands are used to produce a large number of molecular probes that interact well with at least one pose of one active ligand (panel B, with hydrophobic probes shown without hydrogens for clarity). Panel A shows the initial ligand alignment of the training molecules. Panel B shows the initial probe superset (atom-color) in comparison to the hypothesis alignments (cyan) used for seeding the initial alignment.

each ligand and each of the pocket probes generated. Given such a matrix, selection of a subset of probes that yields scores close to experimental activities can be treated as an optimal search programming problem, since the predicted activity for a molecule is simply the sum of its scores against a set of chosen probes. To accomplish this task we designed a search procedure that function like a genetic search algorithm (GA), that selects a minimal subset of probes that yields a cumulative interaction score with each ligand such that the mean-squared-error (MSE) among all the training ligands is within 0.05 log units of the experimentally measured activity. This approach enables full consideration of the multiplicity of poses for each ligand while probes in a manner that is agnostic of the number and types of probes selected. The procedure accepts as input the ligand pose pool and dense set of probes from the preceding initialization steps along with the binding activity for each training ligand. The procedure creates an initial population of randomly selected probe subsets (individuals). Every probe subsets is evaluated and the fraction of individuals yielding the best (lowest) MSE are retained for reproduction of progeny subsets. Progeny reproduction involves the random selection of pairs of retained individuals followed by a stochastic combination process that produces a progeny probe subset containing inherited probe selection from both parents. The selection and reproduction process is repeated until a probe set yielding a MSE within 0.05 log units is found or until a predetermined maximum number of progeny generations have been explored. Figure 3.7 shows a schematic of the genetic algorithm search procedure used for probe initialization.

We also introduce methodological enhancements to the model construction procedure that includes multiple biases beyond fidelity to experimental activity values. For the purposes of denovo model induction described in this chapter, the biases include a measure of model parsimony and conservation of the pocket model size. The conservation bias is a linearly-scaled penalty corresponding to the number of probes present in an evolving model. A model with fewer probes is favored than otherwise. Model parsimony is described by similarly active training ligands adopting similar binding poses. Given molecules with similar activity, a parsimonious model is one that explains their binding in geometrically similar ways. The quantitative measure of parsimony is formulated as a weighted sum of pair-wise similarities of all final ligand poses, where molecule pairs with similar activity receive higher weight than those with different activity values. Figure 3.8 shows an example of similarily active ligands represented by a parsimonious model versus a non-parsimonious model. The parsimonious representation provides a physically concordant depiction of how molecules

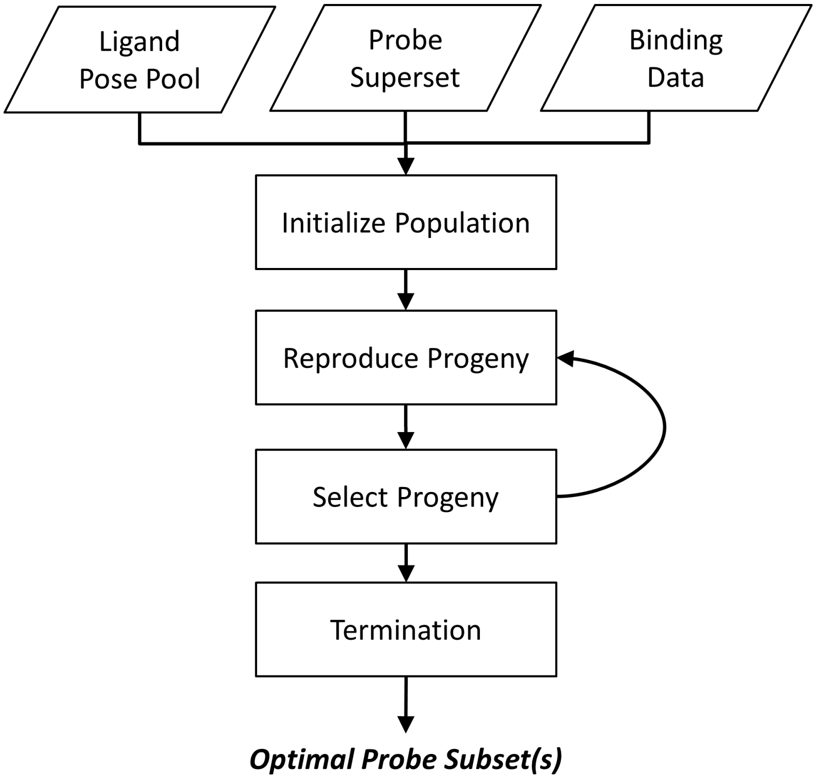


Figure 3.7: The probe initialization search procedure functions like a genetic algorithm, that selects a minimal subset of probes that yields a cumulative interaction score with each ligand such that the mean-squared-error among all the training ligands is within 0.05 log units of the experimentally measured activities.

X and Y may be geometrically similar in their respective bioactive poses.

The resulting small set of probes is shown with thick sticks with skin in figure 3.9. This parsimonious set of probes captures the key interactions that have been described in previous studies of the design and optimization of benzimidazole inhibitors with the GyrB ATP binding pocket (e.g., ref Charifson et al.[62]). Pocket probes complementing the h-bond acceptor, donor, and multiple hydrophobic elements were automatically selected on the basis of the requirement to match the activity pattern of the molecules. The small set of selected probes leaves large gaps in the pocket surface, so the search proceeds in a 3-step process that searches for an optimal probe subset, followed by

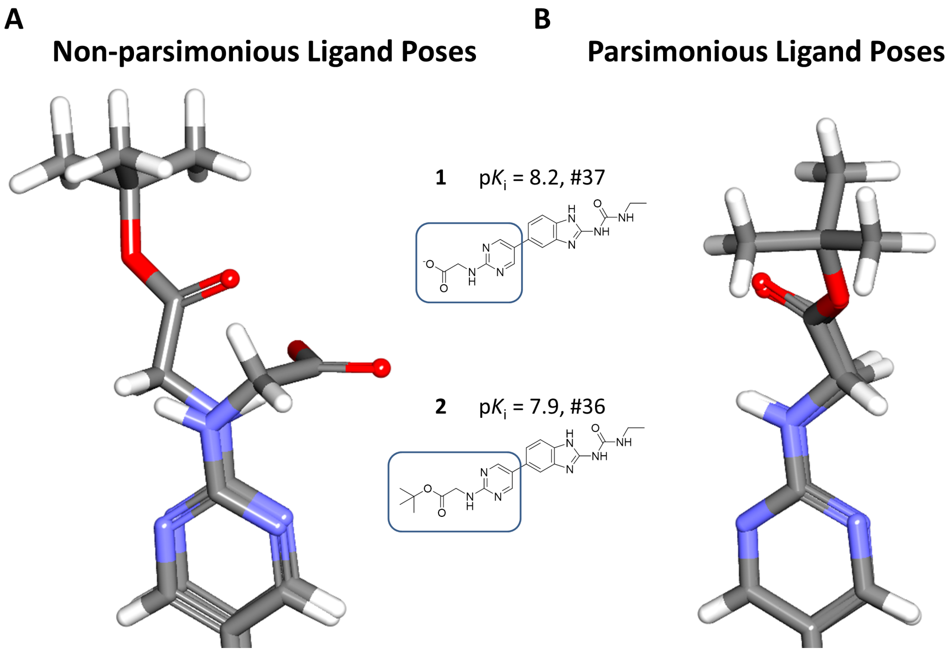


Figure 3.8: Poses of similarly active training ligand exhibiting contrasting levels of model parsimony. Panels A and B show the results of ligand poses derived with and without parsimony consideration during probe initialization. Panel A shows the poses of molecules **1** and **2** derived from an initial probe/model search with the parsimony bias turned off. Panel B shows a contrasting view of molecules **1** and **2** produced from model initialization with the parsimony bias turned on. When parsimony was turned on, the search procedure favored poses that were highly geometrically concordant, as can be observed with the correspondence between the amines and oxygens located on the carboxylate and ester in panel B. A disconcordant representation of these functional groups were observed when parsimony was turned off as seen in panel A.

a quick refinement step that refines the positions and orientations of the probes and poses relative to one another, followed by an additional search that uses the partially refined probes and pose pool to form the bases of the final search. This requirement is a consequence of using the multiple-instance learning approach. Since ligand poses will change in order to optimize activity relative to the model, overly sparse models allow too much freedom in alignment adaptation. For example, a ligand that is inactive because of a large steric protrusion can adapt to occupy empty space with its extra bulk. To allow the model freedom to cover all parts of space surrounding the training molecules, probes are *added back* from the larger pool iteratively, including new probes not near any existing ones. In this work, the tolerances for nearness were 3.5 for hydrophobic probes and 2.0 for polar probes, measured using RMSD in Å. The *add back* search employs an additional search bias that favors the preservation of probes provided as a search template, in this case the partially refined probes from the initial probe search. Figure 3.9B shows the resulting probe set after partial refinement and search. Notice the probe subset derived after quick refinement provides better overall coverage of training molecule **1** while maintaining key polar interactions with the urea towards the lower region of the molecule and preserving key polar interactions with the carboxylate and pyrimidine nitrogens towards the head of the ligand.

3.3.1.5 Pocket Model Refinement

At this point in the procedure, there exists a pool of poses for each training ligand and an initial pocket model consisting of donor, acceptor, and hydrophobic probes (26 total in the gyrase case). The problem now involves local optimization of both the probes and the ligand poses. The goal is a binding site model in which the computed activity is close to experimentally measured activity (while allowing for ligands to optimize their poses within the model). Probe position and orientations are optimized in order to minimize the mean-squared error (MSE) of computed vs actual activity across all ligands. A steepest descent procedure is employed based on the gradient of the MSE with respect to probe positions. Activity is computed for each molecule using the Surflex-Dock scoring function.[35, 63, 64] A single step of gradient descent requires computation of the change in MSE across all ligands with respect to probe position and orientation for each probe in the pocketmol. This gradient computation is made using the maximally scoring pose for each ligand, where the procedure maintains a set of the poses of each ligand explored during model refinement (this is called the

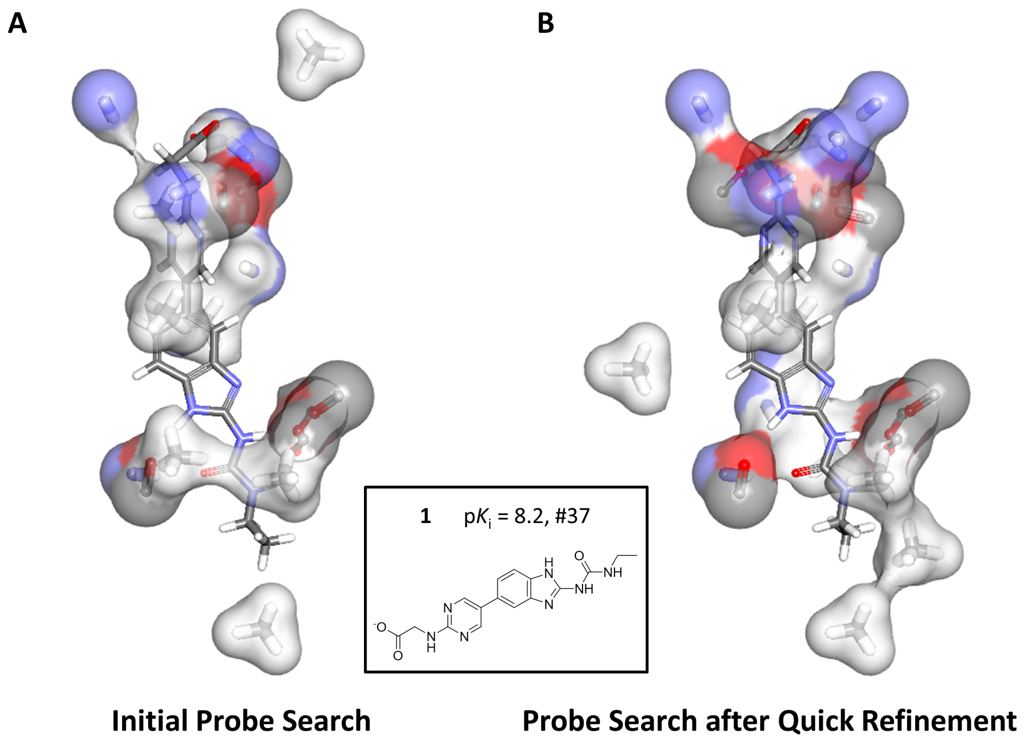


Figure 3.9: The QMOD probe initialization search procedure identifies an optimal subset of probes that fits the ligand binding data. The search procedure is carried out in a three-step process that involves an initial probe search, followed by a quick refinement step, and completed with an additional probe search using the partially refined probe subset as a search seed. A) The subset of probes that optimally fits the training ligand binding data. B) The probe subset derived after quick refinement followed by an additional probe search using the the partially refined probe positions from A and an augmented pose pool that includes partially refined ligand poses.

pose cache). Gradient steps for the probes are made simultaneously for all probes, using a step size that ensures small movements (less than 0.1 Å).

After 100 steps of gradient optimization of probe positions (or if the computed MSE is less than 0.05), the positions are fixed as the current model. Then by use of the current model, each of the ligands pose caches are optimized. Poses for each ligand are subjected to all-atom optimization to maximize the Surflex-Dock scoring function. Note that the scoring function contains both inter-molecular and intramolecular terms, so ligand internal strain is quantitatively traded for interaction with the binding pocket. This iterative process is repeated (probe position optimization followed by ligand pose optimization). During initial learning, the scores of the ligands change substantially during pose cache optimization. In cases where the learning procedure converges, as the learning progresses, these changes decrease in magnitude, and the overall MSE reaches a plateau. The process terminates when the MSE based on ligand scores (after pose cache updating) is less than 0.05. A maximum of 50 rounds of probe and ligand optimization are used. Figure 3.10 shows the final pocket model (atom-colored sticks) along with the initial probe positions (blue). The changes in this example were very subtle, with the most significant rearrangements being in the positions of the acceptor probes that interact with the pyrimidine nitrogens and nearby amine (see discussion for Figure 3.9) of the training ligands. The refinement procedure takes a few hours on standard desktop hardware and is the lengthiest step in the overall process.

3.3.2 New Ligand Binding Affinity Prediction

New ligands are flexibly fit into the final pocket model and are scored as if it were a protein binding site. The scoring procedure enables the direct use of training ligand poses to help guide the alignment of test ligands, which results in more reliable pose generation at a reasonable computational cost. Testing a new ligand takes on the order of minutes. The procedure is analogous to the initial alignment procedure described previously, with the exception that the top scoring alignments based upon similarity to the specified ligands are subject to scoring and local optimization within the fixed pocket model itself. The default parameters make use of 2 similarity weightings (1.0 and 0.1 for polar surface features), 100 best poses from each alignment to each training ligand target, and 5 final poses representing the best optimized fit to the pocket model

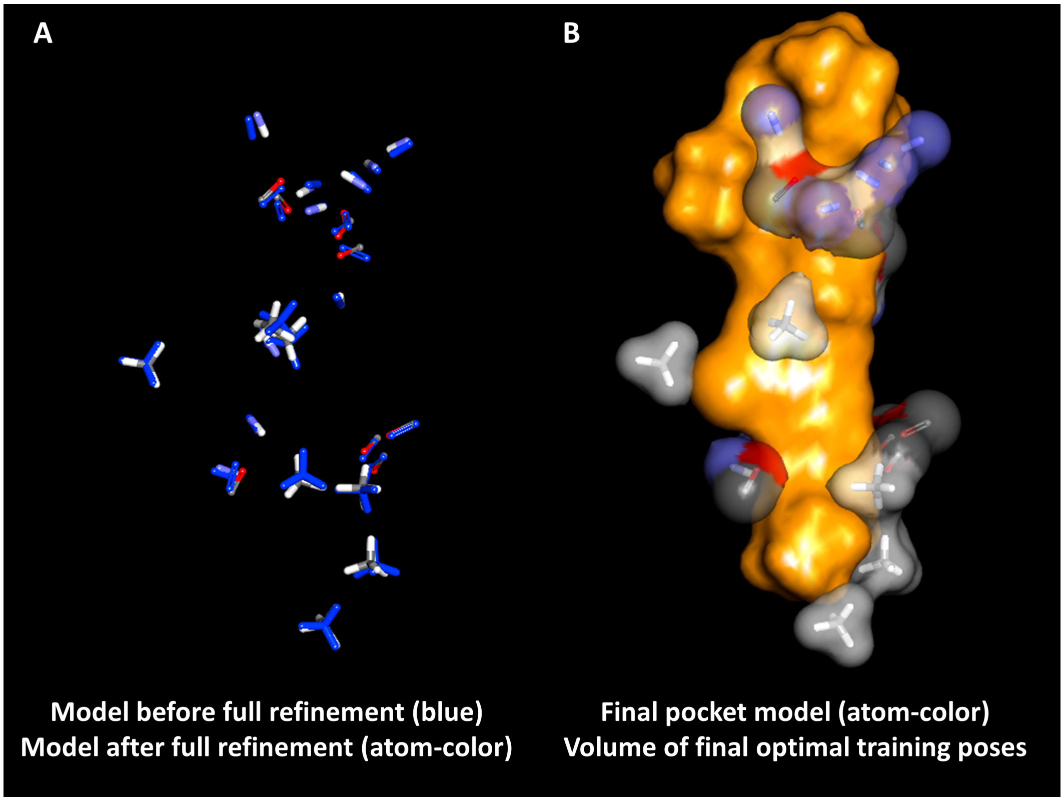

A
B
Model before full refinement (blue)
Model after full refinement (atom-color)
Final pocket model (atom-color)
Volume of final optimal training poses

Figure 3.10: Full model refinement produces an optimal pocket model and poses where optimality is defined by their relative positions and fit of the calculated interaction scores to the known training ligand binding activity. A) The model probes derived from probe initialization (blue) are shown superimposed with the fully refined pocket model (atom-colored). B) The final pocket model (atom-color with skin) is shown with the volume of the final optimal training ligand poses.

for each test ligand.

3.3.2.1 Quantifying Prediction Confidence

In addition to providing a quantitative prediction of activity and physical representation of proposed binding pose, the QMOD method can also provide a measure of prediction confidence. Figure 3.11 shows the initial QMOD pocketmol derived from 39 training molecules (atom-color thin sticks with surface). The pose of compound **2**, which was part of the initial training set, is shown along with the optimal pose of compound **9** (the 47th molecule in the synthetic series). Molecule **9** was predicted with high confidence (0.92/1.0) to have high activity (predicted pK$_i$ of 8.2), yielding an error of 0.3 log units when compared with experimental activity. The confidence measure is defined as the maximal 3D molecular similarity between a test molecule and any of the training molecules (each in its optimal pose according to fit within the pocket model). Here, the most similar training compound to **9** was **2**, with the high similarity obvious in the 2D representations, and with the optimal poses of both molecules being concordant, even including volume overlap of the differing left-hand side substituents.

3.4 Results and Discussion

At a minimum, the goal of QSAR approaches is to make accurate predictions of ligand activity. Preferably, the methods should also yield predictions of relative binding modes, be amenable to visualization, and offer some guidance on the confidence associated with specific predictions. Figure 3.11 depicts the final learned pocket model (atom-colored sticks with skin) along with final optimal pose of the test ligand **9**. As previously described this provides an example of a very accurate prediction made with high confidence. The overall pocket was well enclosed. The interplay between the evolving pocket and the ligand poses was generally subtle, with most ligands showing only minor movement when comparing the initial preferred poses to the final optimal poses.

3.4.1 Performance on the Next Temporal Window of Molecules

Overall the model performed very well on the 50 blind test molecules, producing an average prediction error of 0.52 log units and a Kendall's Tau Rank correlation of

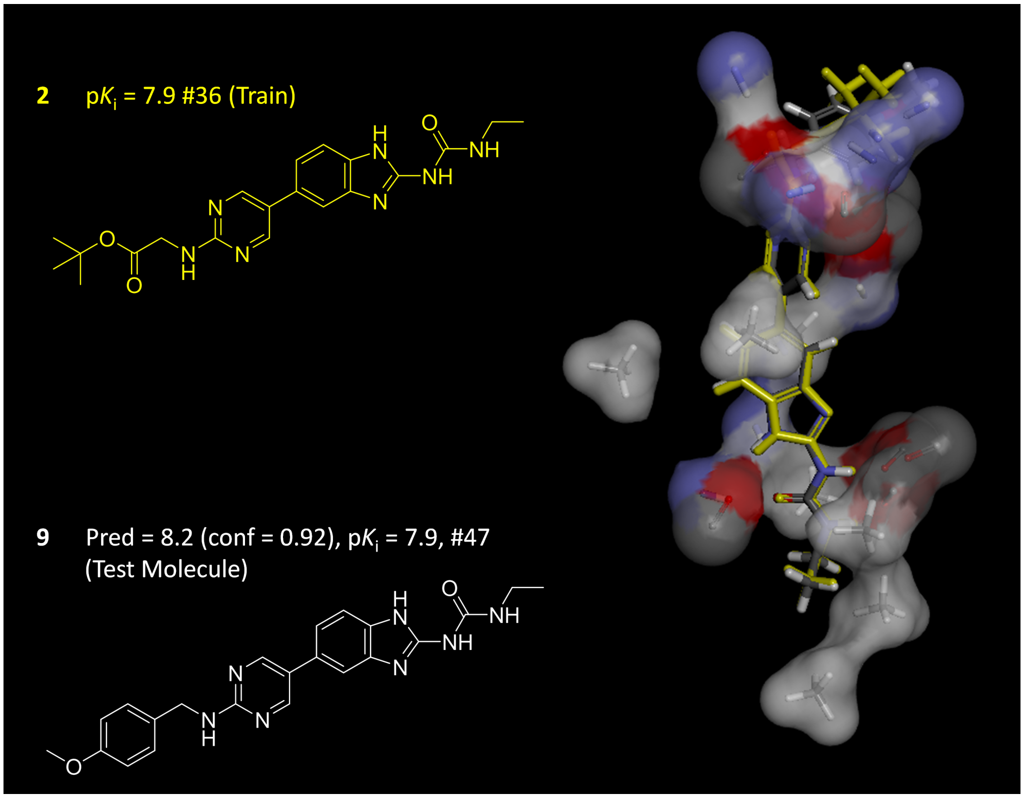


Figure 3.11: The initial QMOD binding site model is shown (right), derived from 39 training molecules. The probes comprising the pocket are shown in atom-colored thin sticks with surfaces. Training compound 2 is shown in yellow, with 2D at left and in its predicted optimal pose at right. Compound 9 (number 47 in the synthetic series) was predicted with high confidence to have a pK_i of 8.2, very close to the experimental value of 7.9 (shown at right in atom colored sticks).

0.35 ($p < 0.01$). The model performed systematically better on predictions made with higher confidence, yielding an average error of 0.4 log units and a Kendall's Tau score of 0.85 ($p < 0.01$) for predictions made with confidence ≥ 0.85.

Figure 3.13 shows molecule **10** in its final predicted pose inside the pocket model. The novel substituent stemming from the nitrogen on the right side provided significant clashes with the pocket model, producing a lower interaction score due to excessive steric clashing with this portion of the model. This scenario highlights a major challenge with ligand-based models that are highly dependent on the breadth

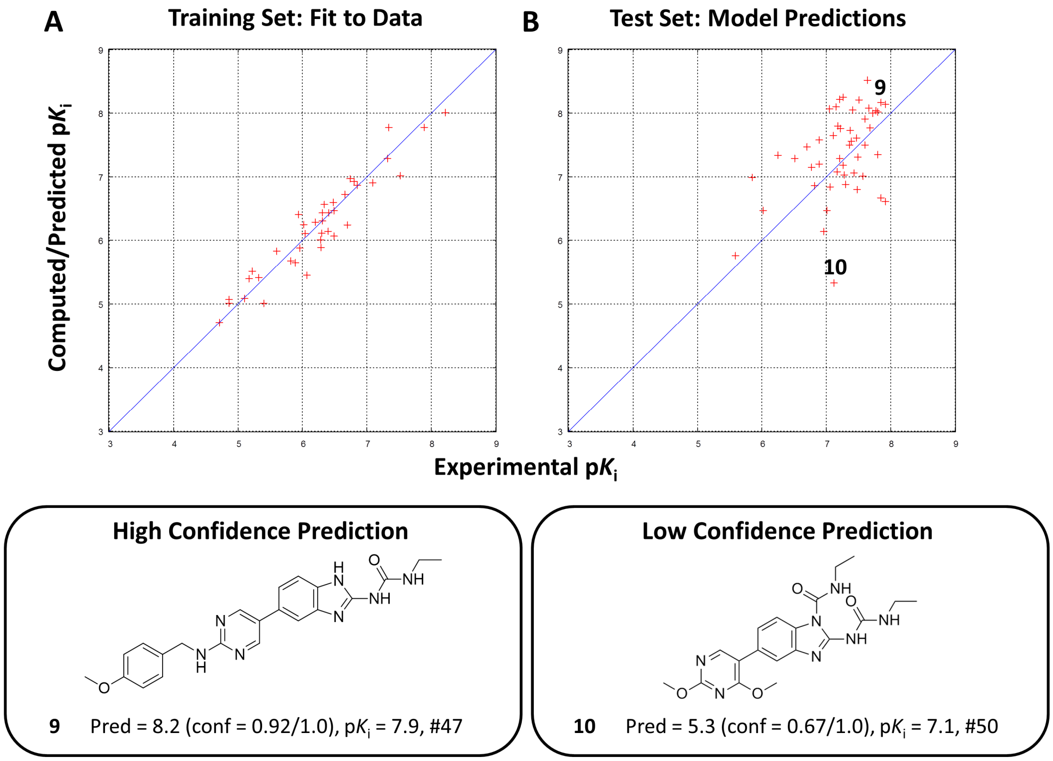


Figure 3.12: Plots of experimental (X-axis) versus computed/predicted pK_i for the final optimized binding pocket for the 39-molecule training set (A) and the 50 blind molecule set (B). Mean error of fit for the training set was 0.2 pK_i units. Prediction error on the 50-molecule holdout set was 0.5 units, with a Kendalls Tau rank correlation of 0.35 ($p < 0.01$ by permutation). Bottom panels highlight two contrasting types of predictions made: those accurate predictions made with high confidence, and those inaccurate predictions made with low confidence.

and quality of training data. In this case a limited number of training molecules with little structural variability near the benzimidazole core scaffold provided negligible information to be learned about potential structural variability in this region. Methods for identifying structurally novel test compounds that may interact with a given model differently that training ligands will be discussed in the following chapter.

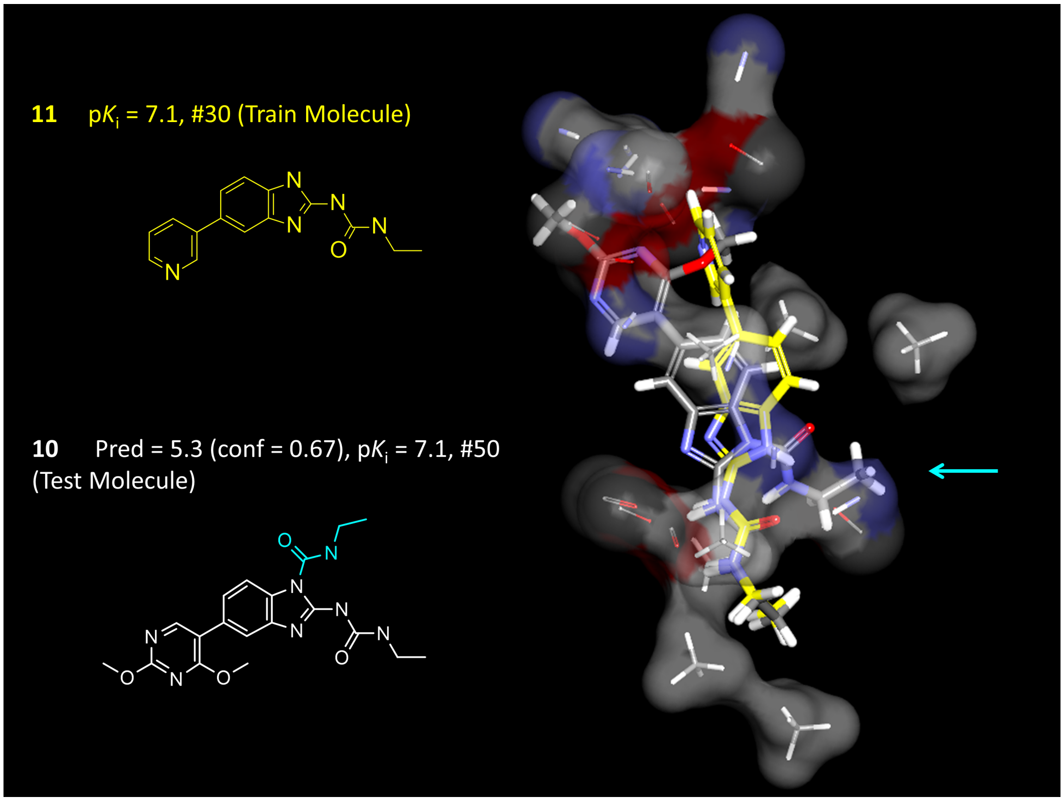

11 pKi = 7.1, #30 (Train Molecule)
N
N
N
N
O
N
10 Pred = 5.3 (conf = 0.67), pKi = 7.1, #50 (Test Molecule)
O
N
N
N
N
N
N
O
N
O
O

Figure 3.13: The initial QMOD binding site model is shown (right), derived from 39 training molecules. The probes comprising the pocket are shown in atom-colored thin sticks with surfaces. Training compound **11** is shown in yellow, with 2D at left and in its predicted optimal pose at right. Compound **10** (number 50 in the synthetic series) was predicted with low confidence to have a pK_i of 5.3, revealing a 1.8 log unit deviation to the experimental value of 7.1 (shown at right in atom colored sticks). Molecule **10** exhibited significant structural novelty compared to the 39 training molecules and consequently interacted very differently with the binding pocket model. Shown in cyan is the novel substituent near the benzimidazole nitrogen that clashed with the right portion of the model causing a significant shift in the final optimal binding pose of test compound **10**.

3.4.2 Relationship to Protein Pockets

The pocketmol that was developed is physical in the sense that it represents real atomic positions of molecular fragments.

The pocket model showed a direct physical relationship to the gyrase subunit

B ATP binding pocket, both in overall shape and detailed accounting of key interaction region within the pocket. Figure 3.14 shows the superimposition of the final pocket model (yellow skin) with the final optimal training poses (atom-colored sticks) compared with the surface of the gyrase binding pocket (blue). The model provides excellent overall coverage of the pocket and the final optimal training poses are physically realistic relative to the concavity of the actual binding pocket. It is important to note, that structure of the gyrase binding pocket was not used in any way during model induction or guidance, but rather served as a validation of model correctness post-facto.

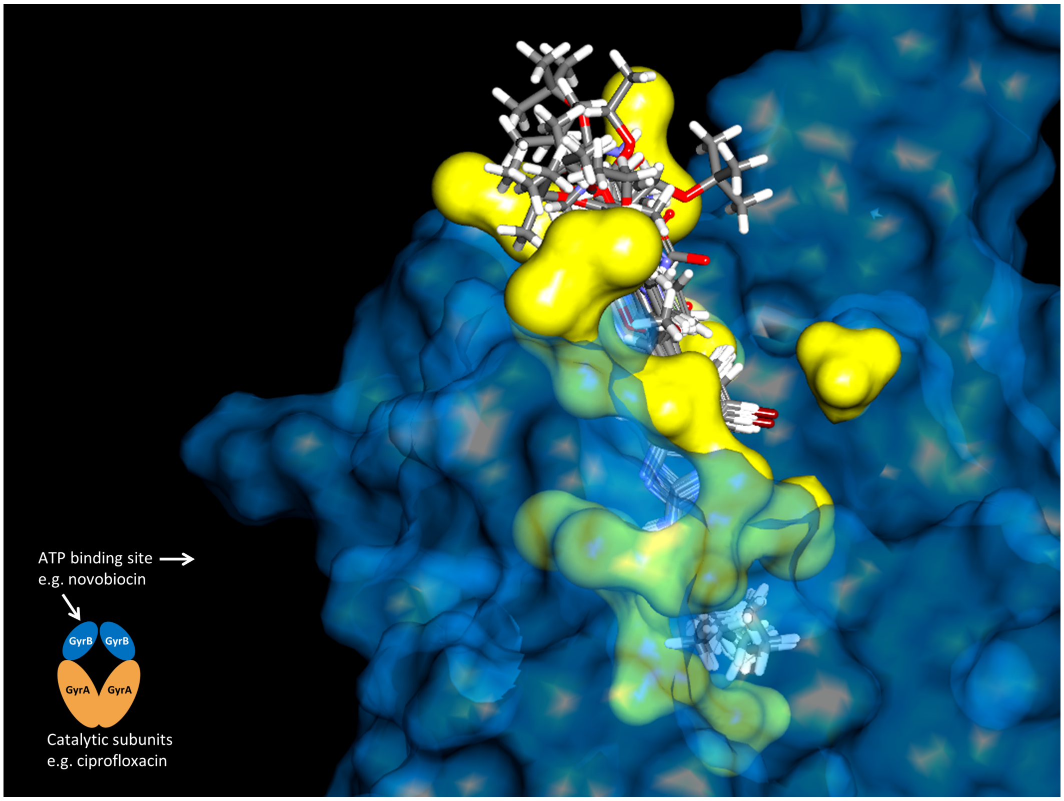


Figure 3.14: The final pocket model (yellow skin) and final optimal training poses (atom-colored sticks) are shown superimposed with the x-ray crystal structure of the ATP binding site of the gyrase B subunit. The model captures the overall shape of the binding pocket while providing physically realistic representations of the final optimal poses of the 39 training molecules.

The group of acceptor probes that interact with the urea nitrogens represents Asp1073 that is known to be critical in binding with this class of compounds.[62] Similarly, several hydrophobic probes captured funnnel-like concavity towards the bottom of the pocket represented by Ala1047, Val1071, Thr1165, ILe1094, ILe1078, and Pro1079. Figure 3.15 shows the correspondence between the model and the binding pocket.

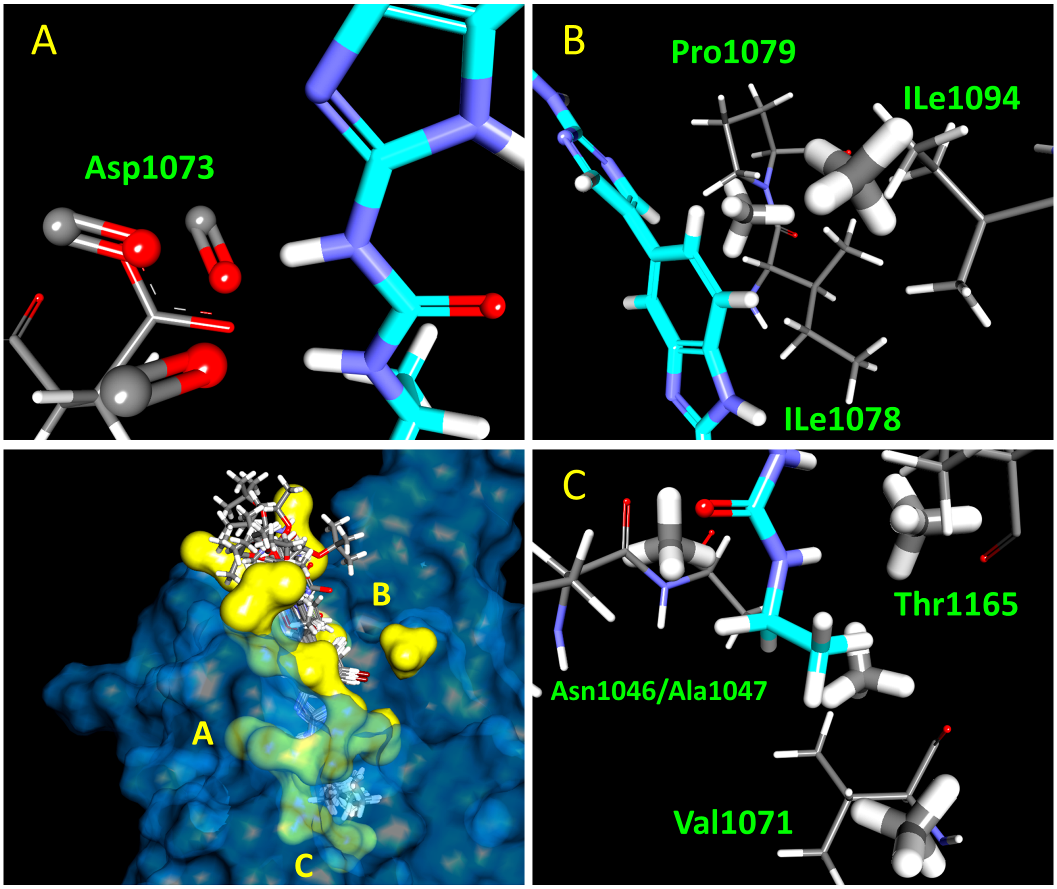


Figure 3.15: The final pocket models capture key residues important for ligand binding. Bottom-left panel shows the final pocket model (yellow) compared with the x-ray crystal structure of the gyrase binding pocket (blue). Training molecule **1** (cyan) is shown to provide a frame of reference. Labels A-C are displayed in zoomed-in view in panels A-C respectively. Acceptor probes represent the Asp1073 known to play a critical interactions with the urea NH groups. Several hydrophobic probes are in good spatial agreement with several residues defining the shape of the binding pocket as shown in panels B and C.

3.5 Conclusion

The Surflex-QMOD approach addresses the physical linkage between activity model and molecular binding mode with pockets having detailed structure comparable to true protein binding sites. Because the model building process results in a model that selects ligand alignments based on mutual interaction, there is a direct correspondence between the physical process of protein-ligand binding and the act of prediction. Notions of model parsimony and prediction confidence are intuitively related to physical notions of shared ligand binding modes and appear to bear directly on the quality of predictions.

Practical approaches for ligand activity prediction in lead optimization that do not rely upon well-determined protein structures must address predictions of ligand pose as well as ligand activity. In this work, a multiple-instance learning approach has been developed for induction of physical models of binding sites. In the challenging test case on gyrase presented here, predictive accuracy on temporarily segregated compounds was excellent, both in terms of numerical accuracy and in terms of geometric concordance with x-ray structure of the binding site. Important challenges remain, including validation on large numbers of targets, identifying structurally novel ligand interactions, development of rigorous approaches to guide model refinement and active learning, and implementation of formal methods to integrate protein structural information in the model induction process. These issue foreshadow detailed discussions in the following chapters.

Chapter 4

Iterative Refinement of a Binding Pocket Model

4.1 Abstract

Computational approaches for binding affinity prediction are most frequently demonstrated through cross-validation within a series of molecules or through performance shown on a blinded test set. This chapter shows how such a system performs in an iterative, temporal lead optimization exercise. A series of gyrase inhibitors with known synthetic order formed the set of molecules that could be selected for "synthesis." Beginning with a small number of molecules, based only on structures and activities, a model was constructed. Compound selection was done computationally, each time making five selections based on confident predictions of high activity and five selections based on a quantitative measure of three-dimensional structural novelty. Compound selection was followed by model refinement using the new data. Iterative computational candidate selection produced rapid improvements in selected compound activity, and incorporation of explicitly novel compounds uncovered much more diverse active inhibitors than strategies lacking active novelty selection.

4.2 Introduction

The field of computational structure-activity modeling in medicinal chemistry has a long history, going back at least 40 years.[65] The preceding chapter discussed the method of *denovo* model induction, highlighting the methodological underpinnings driving the derivation of a pocket model with a single training set of ligand structure-activity data. One main concepts was introduced related to how molecules may be evaluated when *tested* against a given model: prediction of ligand activity with a measure of confidence driven by 3D molecular similarity. This chapter considers this aspect of predictive activity modeling but adds new dimensions. This chapter introduces a method for quantifying molecular novelty in the context of a given model, and shows how considerations of molecular novelty can improve model refinement and predictive power. Rather than focus purely on how well a model can predict activity based on a fixed, particular set of compounds, this chapter instead discusses how a method can guide a *trajectory* of chemical exploration in a protocol that incorporates iterative model refinement. Further, in addition to considering prediction accuracy and the efficiency of discovering active compounds, this chapter discusses considerations of how selection strategies and modeling methods affect the structural diversity of the chemical space that is uncovered over time. The results presented here will show that there is a direct benefit for active selection of molecules that will "break" a model by venturing into chemical and physical space that is poorly understood. The results also show that modeling methods that are accurate within a narrow range of structural variation can appear to be highly predictive but guide molecular selection toward a structurally narrow endpoint. Conservative selection strategies and conservative modeling methods can lead to active compounds, but these may represent just a fraction of the space of active compounds that exist. The content presented here is in part discussed in our recent publication in the the Journal of Medicinal Chemistry

(Varela/Jain[61]).

As previously described Surflex QMOD (Quantitative MODeling) works by constructing a physical binding pocket into which ligands are flexibly fit and scored to predict both a bioactive pose and binding affinity.[59,60,66] Initial QMOD development focused on demonstrating the feasibility of the approach, with a particular emphasis on addressing cross-chemotype predictions, as well as the relationship between the underpinnings of the method to the physical process of protein ligand binding. Those studies considered receptors (5HT1a and muscarinic), enzymes (CDK2), and membrane-bound ion channels (hERG).[59,60,66] The present work addresses two new areas. First, QMOD performance is examined in an iterative refinement scenario, where a large set of molecules from a lead-optimization exercise[62] was used as a pool from which selections were made using model predictions. Multiple "rounds" of model building, molecule selection, and model refinement produced a *trajectory* of molecular choices. Second, the present work considers the effect of active selection of structurally novel molecules that probed parts of three-dimensional space that were unexplored by the training ligands for each round's model. Figure 5.2 shows a diagram of the iterative model refinement procedure. Selection of molecules for "synthesis" for the first round took place from a batch of molecules made after the initial training pool had been synthesized. A full discussion for this initial model build is discussed in Chapter 3. Subsequent rounds allowed for choice from later temporal batches, along with previously considered but unselected molecules. The approach was designed to limit the amount of "look-ahead" for the procedure. The space for molecular selections within each round formed a structural window that reflected the changing chemical diversity that was explored over the course of the project. The iterative procedure was carried out until all molecules were tested. The primary procedural variations involved use of different modeling and selection methods, and the analyses focused on the characteristics of the selected molecular populations, and the relationship of the models to the experimentally determined structure of the protein binding pocket.

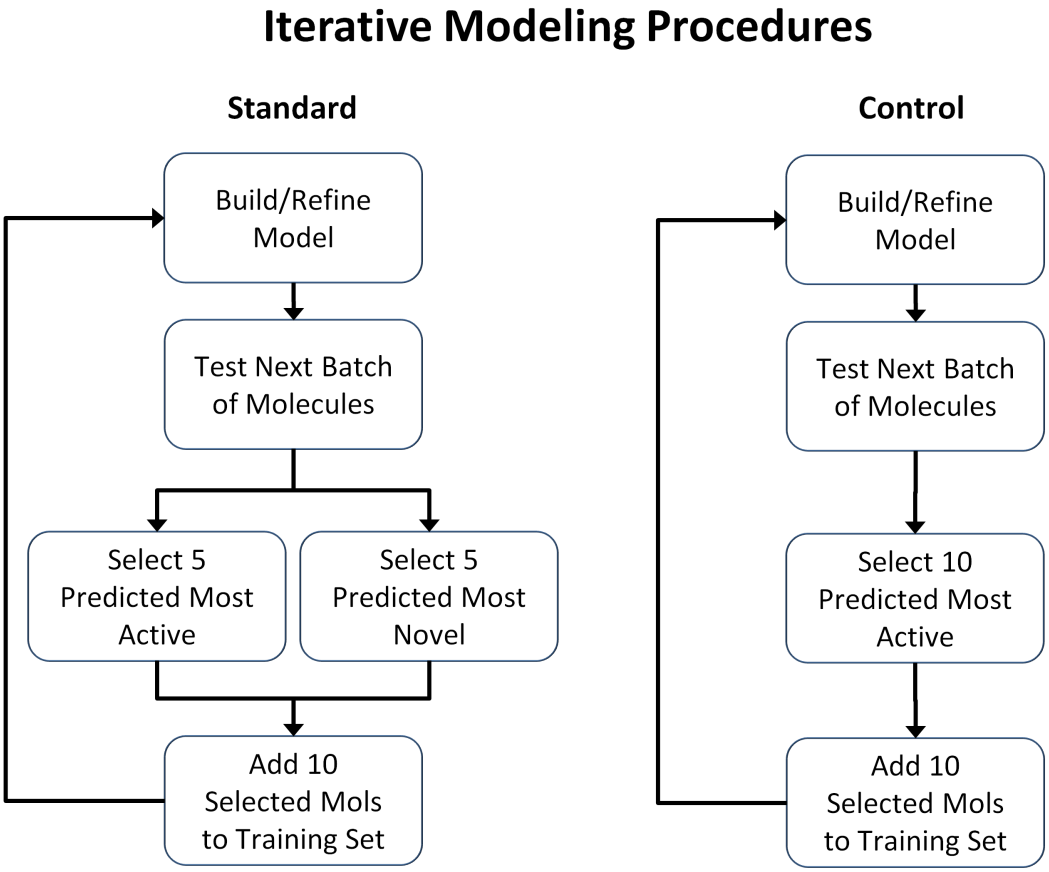

Iterative Modeling Procedures
Standard
Build/Refine Model
Test Next Batch of Molecules
Select 5 Predicted Most Active
Select 5 Predicted Most Novel
Add 10 Selected Mols to Training Set
Control
Build/Refine Model
Test Next Batch of Molecules
Select 10 Predicted Most Active
Add 10 Selected Mols to Training Set

Figure 4.1: The inhibitors first synthesized were used for initial training. All subsequent molecules were divided into sequential batches of 50 candidates each. At the completion of each build/refine iteration, the next sequential batch and all previously considered but unchosen molecules formed a "window" for molecular selections. Based upon model predictions, ten molecules were selected and added to the training set for each round of model refinement. Two selection schemes were employed. The standard method selected molecules based on high-confidence predictions of high activity or based on 3D structural novelty. The control procedure made selections purely based on activity predictions.

All of the molecules discussed in this chapter were taken from a lead optimization program conducted at Vertex Pharmaceuticals. Figure 3.1 from Chapter 3 shows typical examples of structures and GyrB activities from the initial training set.

4.3 Methods

In the preceding chapter several biases were introduced that included that of model parsimony and conservation of a pocket model size. The former relating the 3D geometric concordance between the final optimal poses of similarly active training ligands, and the later describing the preference for smaller and more succinct pocket models. This chapter is a direct extension of the discussion of Chapter 3 with an emphasis on application of the QMOD method in an iterative modeling protocol. In this context, an additional bias (model preservation), is employed during model induction between successive rounds of refinement. Note, model refinement in this context does not describe local pocket model refinement that is carried out immediately following probe search/initialization (Figure 3.2E), but rather refers to the re-application of the entire modeling procedure with the addition of a pocket model serving as an input template guide. Model preservation is expressed as the percentage of model components retained after refinement.

As foreshadowed in the preceding chapter, the initial model derived from the 39 training molecules formed the root of two branches for molecular choice: one making use of a novelty computation and the other focusing only on activity. The novelty evaluation procedure facilitates the identification of molecules that provide novel model-ligand interaction and enables a rigorous interrogation of physical components of an evolving model.

4.3.0.1 Quantifying Molecular Novelty

In addition to predicting ligand affinity, the QMOD procedure can also quantify molecular novelty of a given test molecule. Evaluating molecular novelty extends beyond pair-wise structural evaluation of small molecules. Upon testing a new ligand, the procedure comparatively quantifies the molecule's interactions with a given model. Figure 4.2 depicts an example of the novelty computation relating to a substitution at position 1 of the benzimidazole scaffold. Molecular novelty is a quantitative measure of the degree to which a new molecule explores the space of the binding pocket with new chemical functionality. It is defined using statistics based on the interactions of training molecules with the pocket model and the interactions with unoccupied space near the pocket model (termed the anti-model). The statistics characterize the scores for each probe against the optimal poses for each training molecule and additional poses that sample ligand configurations that are close to optimal. The anti-model is constructed such that it borders on the explored pose pool but excludes the space immediately around the pocket model. Novelty is quantified by comparing the interactions made with the pocket model/anti-model to those made by the training ligands. Compound 10 had the highest novelty score among all 50 molecules in the first batch of compounds from which selections were made. Compound 10 was predicted incorrectly to have low activity, and it was correctly flagged as a low-confidence prediction. Its novelty score was 51.6, corresponding to a normalized Z-score of 5.7 standard deviation units greater than the mean of the remaining pool from those molecules upon which the initial model was tested. The extreme relative magnitude highlights the novelty of the pattern of interaction scores associated with the substitution at position 1 of the central scaffold.

Molecules that interact with the Pocket model and surrounding region differently than the training ligands receive a higher novelty score than otherwise. This defini-

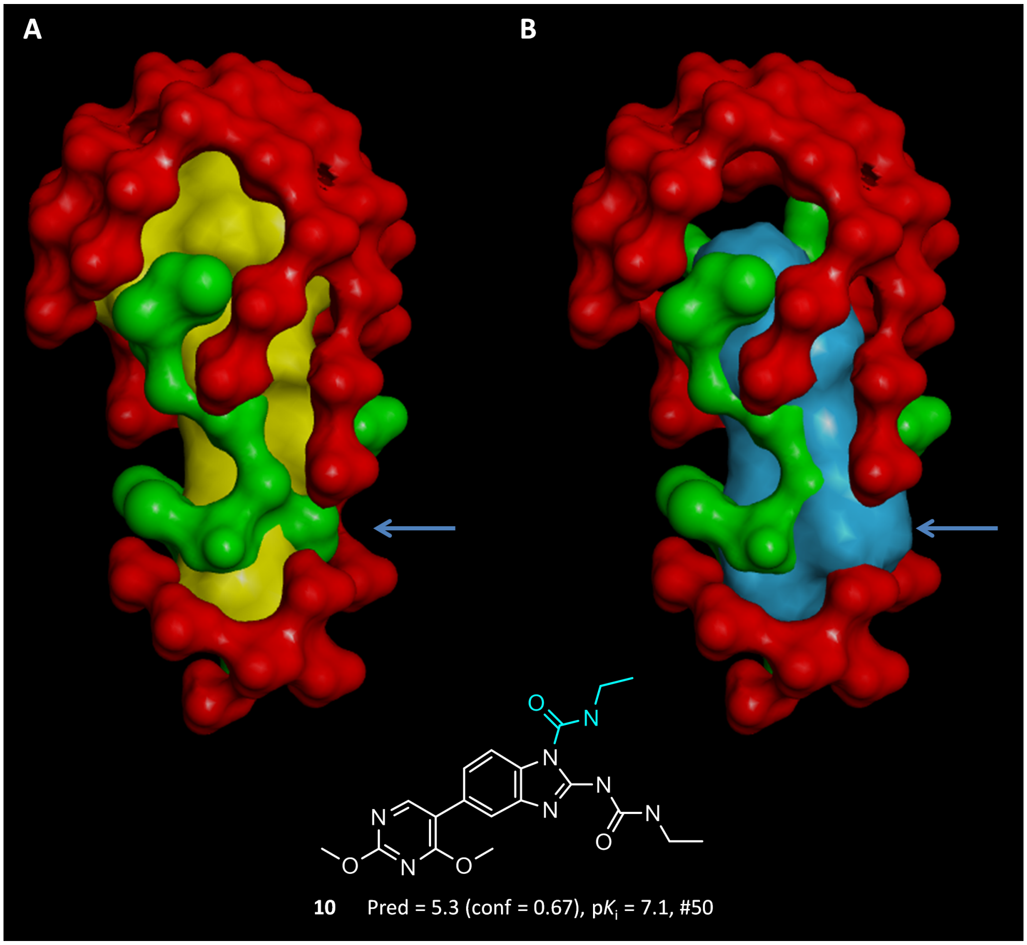


Figure 4.2: The molecular novelty computation compares the interaction score profile of the training molecules in their explored poses (yellow surface, Panel A) to that of a new molecule's probable poses (blue surface, Panel B). The scoring profiles are computed against the the Pocket model (green surface) and anti-Pocket model (red surface), which occupies space that would otherwise be empty. Compound 10, from the initial batch of 50 candidate ligands, contained a novel substitution (shown in blue). This substituent has a natural clash with the Pocket model when aligned to training molecules (blue arrow). The clash produced a tilted pose (not shown), resulting in a low-confidence prediction that was significantly lower than the experimental value.

tion of novelty is highly context dependent and quite different from pure molecular similarity computations. For example, a single methyl group addition to a training molecule will generally have very low impact on a similarity computation. However, if the methyl group pushes into unexplored space (which may or may not contain a pocket model probe), the novelty score will tend to be high.

4.4 Results and Discussion

4.4.1 Effects of Selection Strategy on Experimental Activities of Chosen Molecules.

The ideal experiment in which to assess different design strategies for lead optimization would require independent synthetic teams of equivalent capabilities, each totally isolated from the other. Given an initial starting point, the teams would make a fixed number of compounds over a set time period, with common protocols involving compound testing and provision of assay feedback to the design teams. Although resources needed to carryout such an experiment were not available, a balanced comparison was performed. Here the 39 initial training molecules and their GyrB activities form a common initial starting point, and it is interesting to consider the effects of different computational approaches in terms of the properties of the molecules that are selected from among the remaining 387 that were part of the series. In the standard procedure, half of the molecules selected were chosen to maximize predicted activity and half were chosen as being structurally novel in order to inform the model in areas that had not been explored. In the control procedure, all of the molecules were chosen to maximize activity. Figure 4.3 shows the distributions of experimental activities of molecules chosen using the QMOD standard procedure compared with the QMOD control procedure (recall Figure 5.2). The two distributions within the standard procedure were very different ($p << 0.01$ by Kolmogorov-Smirnov (KS)), with the novelty-driven selections exhibiting a wider dispersion of experimental activity and a much larger proportion of poorly active molecules (roughly 30% with $pK_i <$

6.5 compared with $< 5\%$ from the activity-driven selections). Despite being informed quite differently in terms of structure-activity data, the distribution of activities for molecules selected for activity under the standard protocol were not different than those selected in the control procedure (see Figure 4.3B). The structural characteristics of the resulting pools were very different, and this will be discussed in the next section.

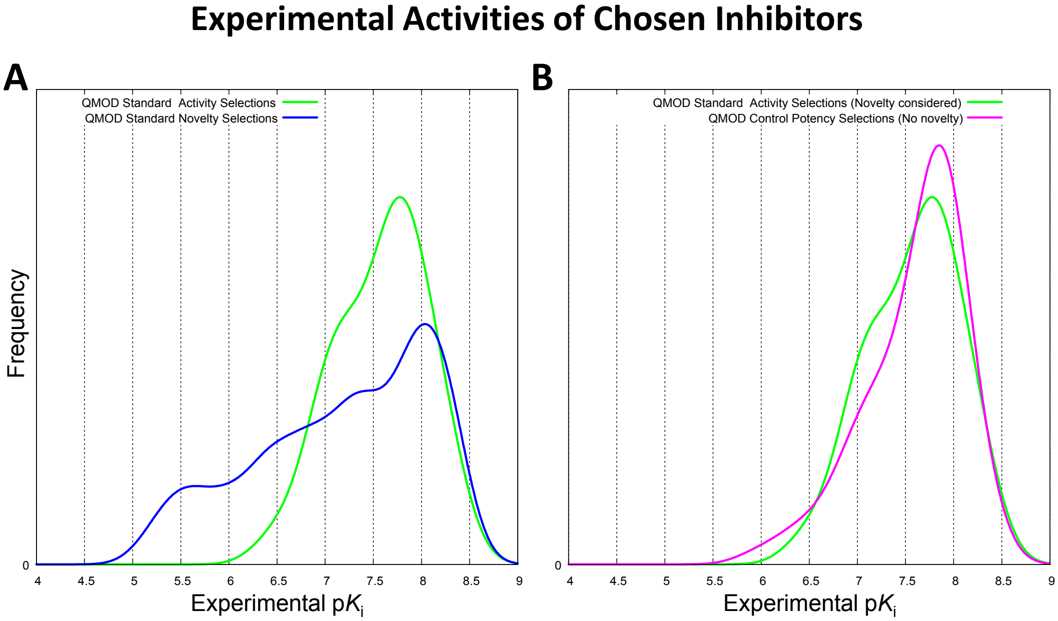


Figure 4.3: Plot A shows the distribution of experimentally measured activity for the QMOD standard procedure, comparing the 40 molecules chosen based on predictions of high potency (green curve) and the 40 molecules chosen based on structural novelty (blue curve). Plot B shows the comparison between the QMOD standard procedure (green curve) and the control procedure (magenta curve), which made selections based solely on potency predictions.

The comparison between the two QMOD procedure variations fits our comparative experiment ideal, with fully independent “synthetic teams” employing different design strategies in isolation. If we consider the distribution of experimental activities of the *next 80 molecules actually made* after the initial 39 in the training set, we deviate from our ideal. First, the project chemists were interested in address-

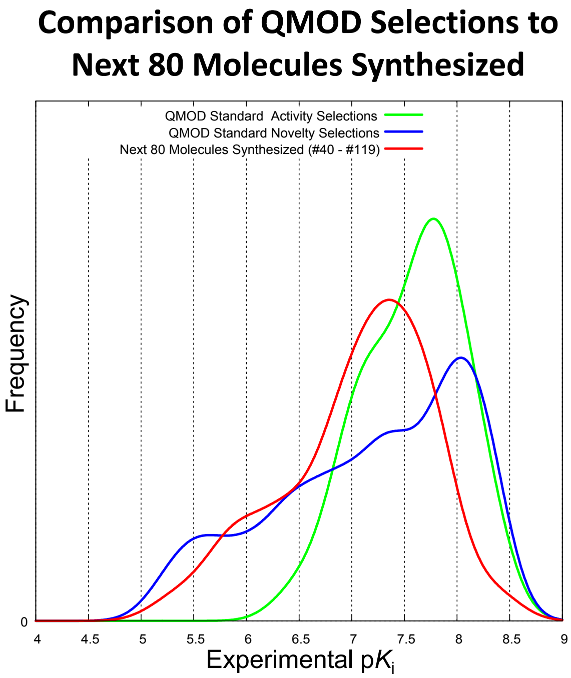


Figure 4.4: The three distributions of experimental activities shown are all highly significantly different from one another: 40 compounds selected for potency (green), 40 selected for novelty (blue), and the next 80 actually synthesized after the 39 that formed the QMOD initial training set (red).

ing issues beyond just potency against GyrB. The considerations included activity against ParE, physical properties of compounds, complexities of synthesis given existing routes and materials, and a host of other items. Clearly, however, they were interested in maximizing potency against GyrB. Second, the project chemists had access to information well beyond what the QMOD modeling procedures had, including crystallographic guidance and knowledge of other inhibitors of the ATP binding sites of gyrase. Bearing this in mind, it is interesting to consider the comparison between the QMOD selections in the standard procedure and the activities of the next 80 molecules actually synthesized after the initial 39. Figure 4.4 shows the three distributions, each of which is highly statistically different from one another. This comparison is not meant to suggest that the QMOD selection approach is definitively “better” in any sense than the efforts of human designers. The comparison provides context for what the space of designable compounds looked like within a fixed frame of temporal exploration measured in numbers of compounds made.

Figure 4.5 provides additional detail, showing the experimental activities in temporal selection order for the QMOD standard protocol, the control protocol with no novelty bias, and the next 80 molecules synthesized. Figure 4.5A shows the trajectory of activity observed with the 40 QMOD standard potency-based selections, nearly all of which had activity greater than 7.0 pK_i. Toward the end of the eight rounds of selection, nearly all molecules had potencies of 8.0 or higher. The corresponding novelty selections (Figure 4.5b) exhibit much wider dispersion, with both “winners” and “losers” being selected across the entire sequence. Notably, maximally active molecules were chosen earlier through novelty-based selection than through potency-based selection in the standard procedure. Again, for contextual purposes, and with the caveats described above, Figure 4.5c shows the sequence of experimental activities for molecules in the synthetic sequence numbered 40–119. The high dispersion and downward trend were probably driven by many factors, but clearly there were challenges in meeting multiple design criteria while maintaining or increasing potency against GyrB. The QMOD control procedure (Figure 4.5d) exhibited stable performance, reliably picking a preponderance of molecules with activity greater than a pK_i of 7.5. Recall that while the distributions corresponding to plots A–C were all significantly different, conditions A and D produced indistinguishable distributions in a statistical sense.

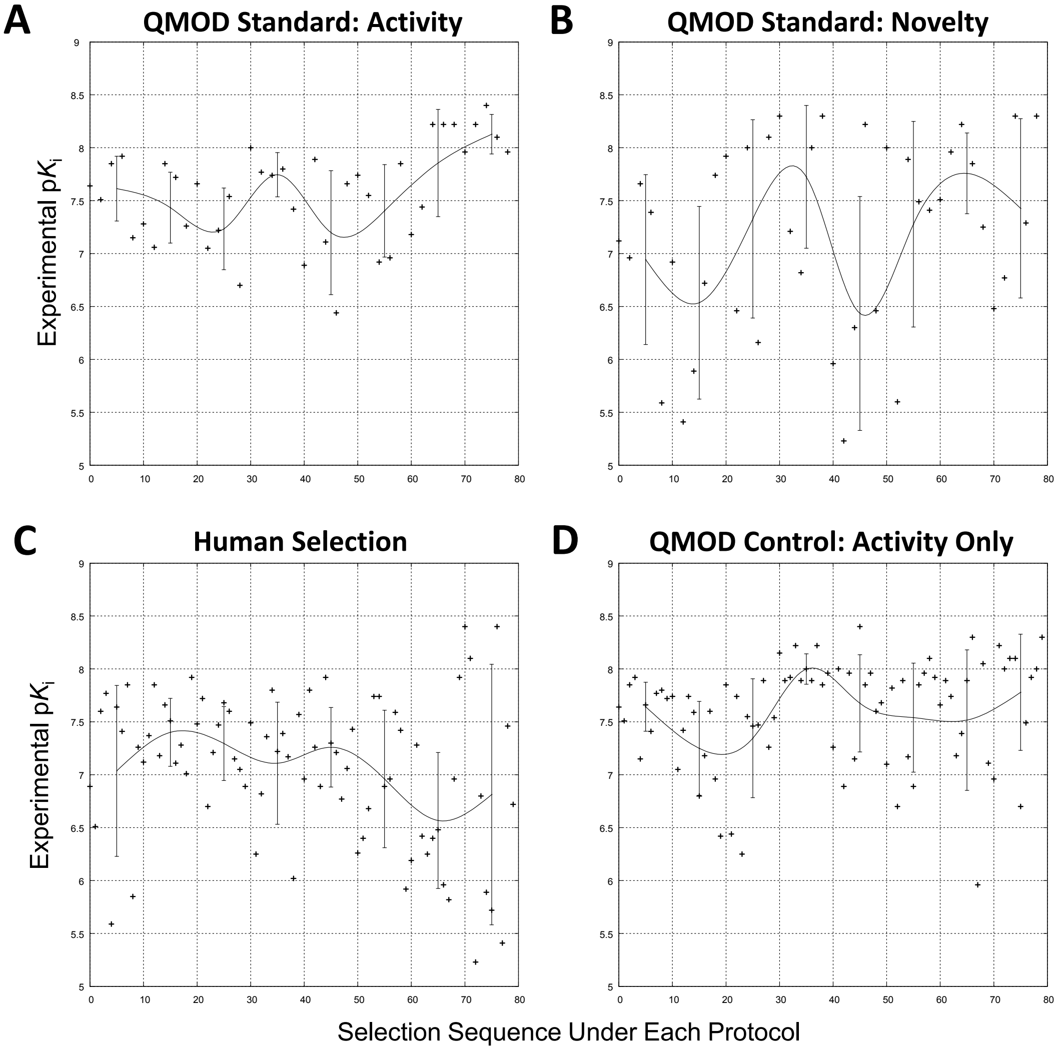


Figure 4.5: The experimental activity of molecules selected is plotted against selection order under different protocols.

4.4.2 Effects of Selection Strategy on Structural Diversity of Chosen Winners

The molecular pools selected using the standard procedure or one without a novelty bias exhibited indistinguishable distributions of GyrB activity. However, the actual value of a given pool of potent inhibitors is affected by chemical composition. A single potent inhibitor along with several nearly identical variants will generally be less useful that the same inhibitor along with several equipotent but structurally different variants. We defined a threshold of p$K_i \geq 7.5$ to identify molecules with desirably high potency ("winners") and compared the structural diversity the winners from the different selection procedures. The standard selection procedure that combined novelty with potency found structurally diverse potent molecules. The plots in Figure 4.6 show the distribution of pairwise 2D (left) and 3D (right) similarities of the winners. The diversity of winners resulting from the standard QMOD procedure is shown in green, and that resulting from the control procedure without novelty is shown in magenta. The distributions of 2D similarity differed primarily in the tails, with the standard procedure showing very few highly similar winning pairs compared with the control procedure. Also, the standard procedure identified a small population of divergent pairs that were missed by the control procedure. The 3D similarity distributions exhibited much more substantial differences, with a very significant shift toward lower mutual similarity within the population of winners from the standard procedure. Figure 4.6 shows an example of a typical highly similar pair (compounds **9** and **12**) from the control procedure along with a structurally divergent pair (compounds **13** and **14**) from the standard procedure. The protrusion of **13** (lower right, in blue) is particularly stark. Notably, inhibitors containing 7-position substitutions also possessed markedly improved potency against ParE,[62] with dual-inhibition of GyrB and ParE being desirable in the context of antibacterial development.

The use of a novelty bias in compound selection drove the exploration of structural diversity. This is easily seen in the evolutionary design tree shown in Figure 4.7. Two selection pathways are depicted that led to two structurally different, yet potent, gyrase inhibitors. In round 2 (left side of Figure 4.7), **15** (dashed arrow) was selected for novelty because of the new interactions made with the model from the benzyl-ester substitution at position 7 of the benzimidazole. In round 7, **16** was selected

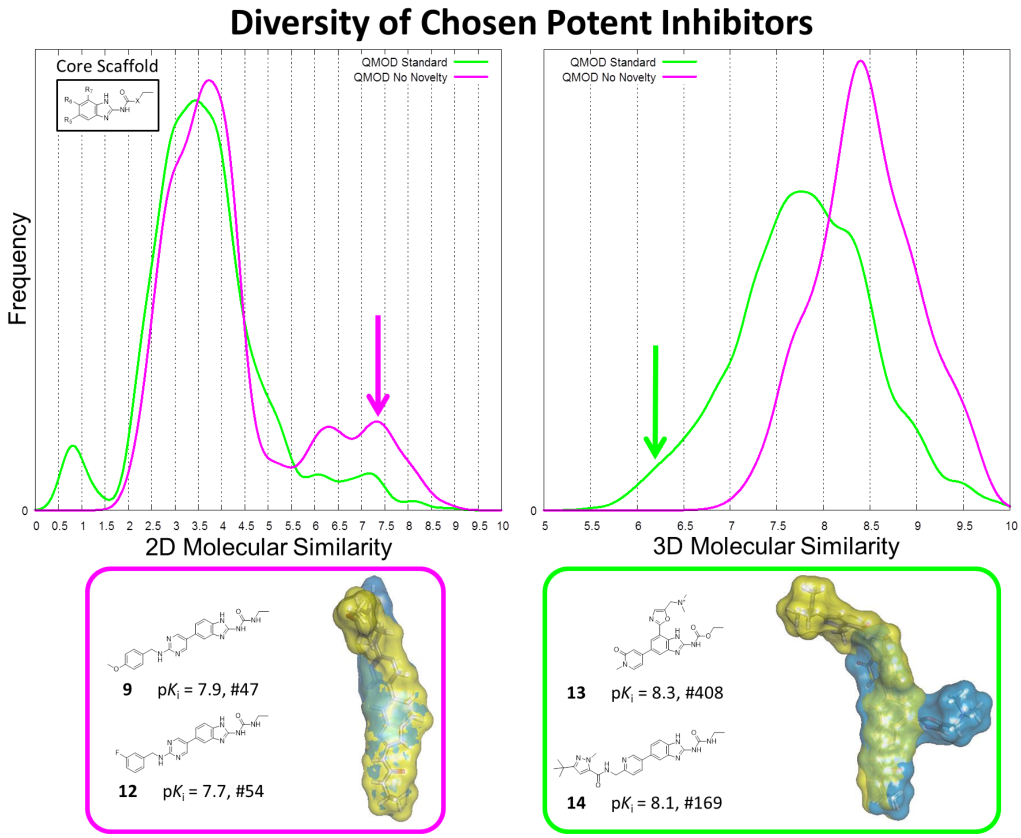


Figure 4.6: The structural diversity among the molecules selected using the QMOD procedure that included an active novelty component was significantly higher in both 2D (left) and 3D (left). At bottom, example pairs of molecules are given from the control procedure (left) and the standard procedure (right).

for potency, where confidence was derived from **15**. In round 8, **17** was selected confidently based on similarity to **16**. QMOD converged on making more accurate predictions on the position 7-substituted molecules over time: errors in prediction for **15**, **16**, and **17** were 0.5, 0.4, and 0.3, respectively. On the right-hand side of Figure 4.7, a separate branch of selections without a substituent at position 7 was also elaborated. In round 3, **18** was selected for potency (similar to **3**). In round 8, QMOD identified one of the most potent compounds in the entire set. Compound **19** was accurately predicted with high confidence (similar to **18**). Molecules **17** and **19** are examples of the most potent and structurally dissimilar molecules in the entire pool.

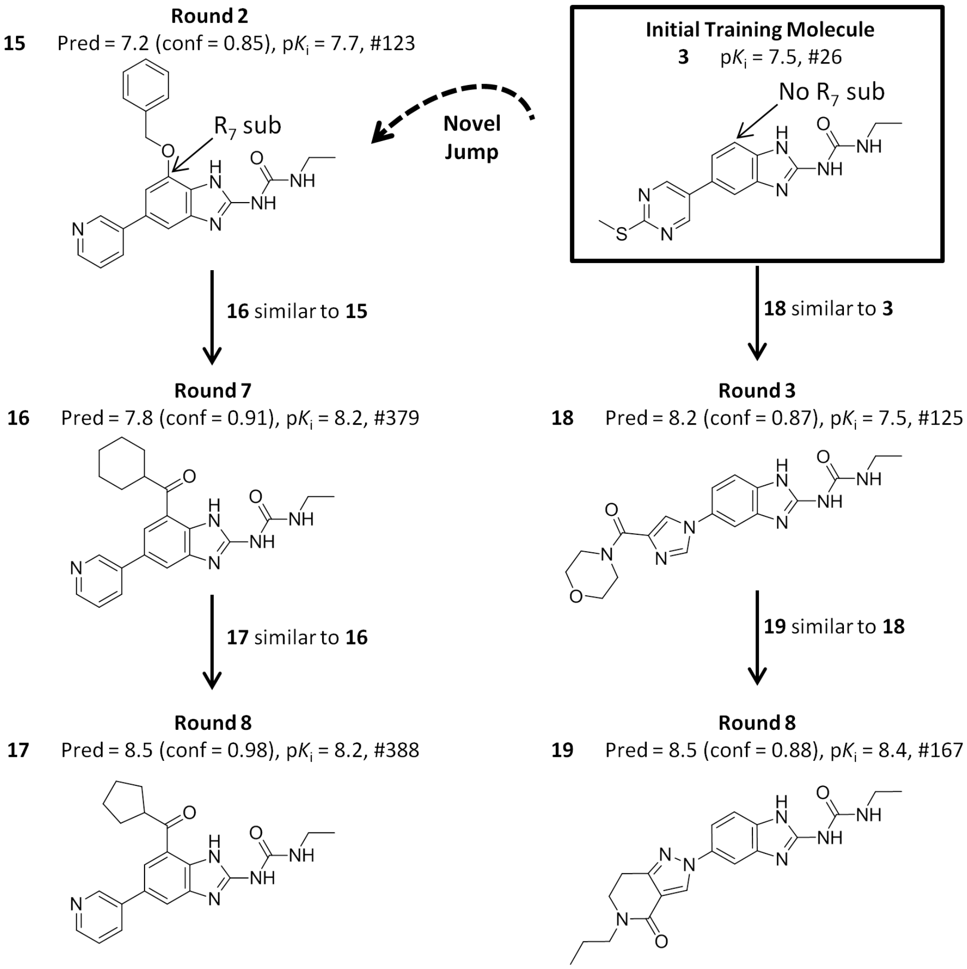

Round 2
15 Pred = 7.2 (conf = 0.85), pKi = 7.7, #123
R7 sub
Novel Jump
Initial Training Molecule
3 pKi = 7.5, #26
No R7 sub
16 similar to 15
18 similar to 3
Round 7
16 Pred = 7.8 (conf = 0.91), pKi = 8.2, #379
Round 3
18 Pred = 8.2 (conf = 0.87), pKi = 7.5, #125
17 similar to 16
19 similar to 18
Round 8
17 Pred = 8.5 (conf = 0.98), pKi = 8.2, #388
Round 8
19 Pred = 8.5 (conf = 0.88), pKi = 8.4, #167

Figure 4.7: Examples of molecular selection based on novelty or on high-confidence predictions of high potency give rise to a branched pattern of chemical exploration.

A significant driver of the 3D structural diversity in the standard procedure arose based on the discovery of multiple potent inhibitors (e.g. compound **13**) with significant 7-position substituents. Figure 4.8 shows the surface envelope of the winners from the standard selection procedure (green) along with that from the control procedure (magenta). These poses were derived by docking into an experimentally determined GyrB protein structure to provide a common target for visualization of the spatial exploration of the binding pocket. The corresponding circled areas identify the binding pocket space that was explored based on active selection of novel molecules that was missed when focusing solely on potency. One of the pitfalls in exploring a binding pocket *without* the benefit of an experimentally determined protein structure is that the degree to which the pocket can be defined is driven purely based on synthesis and assay of compounds. In this purely apples-to-apples comparison of two computationally driven selection procedures, it was clear that a quantitatively driven strategy to explore space *beyond* what had been mapped led to the discovery of a cavity capable of offering increases in inhibitor potency. The class of 7-position substituted inhibitors showed notably better dual-inhibition profiles,[62] illustrating a concrete biological benefit of this type of structural diversity.

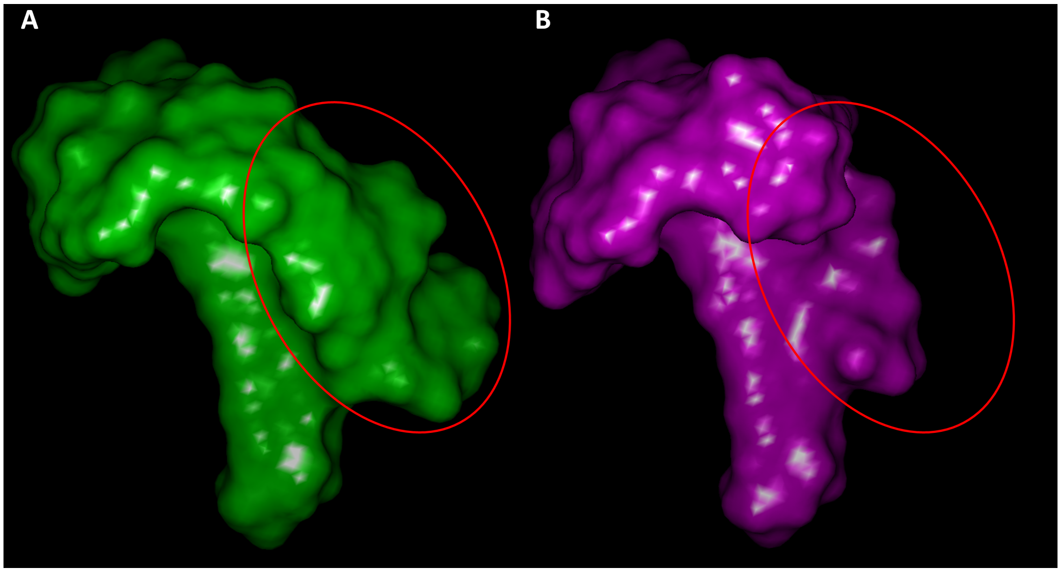


Figure 4.8: The structural diversity among the molecules selected using the QMOD procedure that included an active novelty component was significantly higher in both 2D (left) and 3D (right).

In addition to considering the two variants of the QMOD approach, we also ran a descriptor-based QSAR approach that combined 2D molecular fingerprints with the random forest learning method (termed “RF”).[67–69] Two procedures using of the RF approach were run, paralleling the two procedures used by QMOD (see Figure 5.2). Selection of novel molecules with the RF approach was done by clustering compounds in the selection pool based on their fingerprints and identifying cluster centers. Among the pools of molecules selected for potency by either the QMOD of RF method, whether or not active novelty bias was employed, no significant differences in the distributions of experimental activities were found (KS test p-value > 0.05 in all pairwise comparisons).

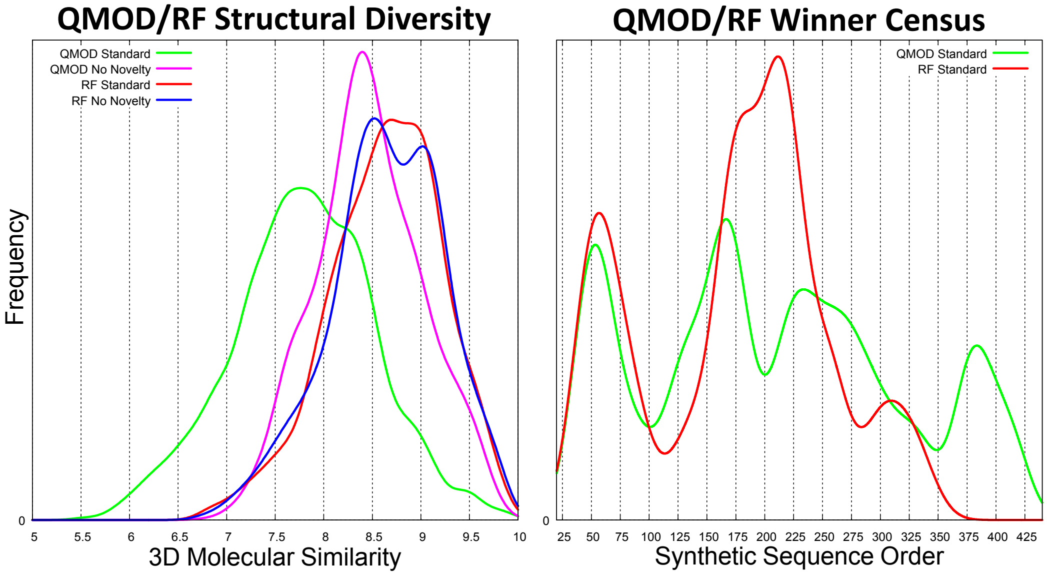


Figure 4.9: Structural diversity among the winners chosen by the RF procedures was much lower than for QMOD (left plot). This lack of diversity stemmed from the lack of diverse selections from the overall project chemical population (right plot).

However, the RF approach, either with or without a novelty component within the selection procedure, produced far less diverse pools of winners. Figure 4.9 shows the 3D similarity distributions of pairwise winner comparisons for the two QMOD variants and the two RF variants. Use of diverse fingerprint cluster centers failed to make an impact on the structural diversity of winners for the RF approach (KS test p-value = 0.33). However, while the QMOD standard approach produced a much more diverse pool of winners than the control approach without active novelty selections, the QMOD control approach produced a significantly more structurally diverse pool of winners than either RF procedure (KS p-value $<<$ 0.01). The lack of diversity is directly evident in the histogram of synthetic sequence numbers shown in Figure 4.9, with the RF approach exhibiting just two primary peaks corresponding to early- and mid-project. The QMOD approach exhibited four peaks, including a set of potent inhibitors from late in the project. Compounds 13, 16, and 17 (Figure 4.6 and Figure 4.7) all corresponded to the rightmost peak, and all of which were made *after* any experimentally potent selections from the RF procedures.

From the middle peak of winners in the synthetic sequence order was a winner shared between the QMOD and RF approaches (sequence #219). Among the winners from the RF protocol, 55% had extremely high 3D similarity to that single compound ($\geq$ 8.50), compared with just 12% of the QMOD control winners. The RF procedure was certainly successful in identifying potent inhibitors, but the procedure, even with a novelty bias, ended up strongly over-represented with multiple examples of highly similar molecules.

One property of sophisticated regression methods such as random forest learning is that many aspects of the population statistics of a training set are well-modeled in order to reduce errors when tested on new data. The models are explicitly affected by both the prevalence of output values and particular features. In a molecular modeling application, it is frequently the case that one specifically designs molecules that literally reach beyond those whose behavior has been modeled. Consider two design candidate molecules, both of which will turn out to be highly potent. Suppose that one of the molecules is highly similar to a pre-existing training molecule in terms of its computed features and one is not. A sophisticated correlative machine such as a random forest predictor will correctly assign a high potency to the former potent ligand. But, it will tend to predict a value for the latter ligand that is close to the maximum likelihood value based on the distribution of training molecules' activities (typically close to the mean or median activity). A mid-range prediction for an "unknown" is a wise play in a probabilistic sense, but it reflects no knowledge of the structure-activity relationship. This "near neighbor" effect manifested itself here very directly. The compounds that were correctly ranked highly during the selection process for the RF method tended to be structurally similar to pre-existing potent compounds.

To test this directly, we constructed an RF model using the same final training molecules as were used for the final QMOD standard model. Both methods identified potent compounds among their top 10 ranked predictions (mean experimental pK_i in both cases of 8.0). However, the 2D structural similarity of the top-ranked RF molecules to the training molecules was much higher than for the QMOD approach (KS p-value $<< 0.001$). This was also seen in the reverse direction. Among the test compounds with pK_i was ≥ 7.9, there was significant variation in the 2D similarity of each compound to its nearest training neighbor. The set of 10 *furthest neighbors* from the training set were arguably the most interesting compounds from the perspective of requiring an accurate computational prediction. They had a mean experimental activity of 8.2. For these, the RF predictions averaged just 7.0, with a single compound predicted to have $pK_i \geq 7.5$. For QMOD, the predictions averaged 7.8, with 7/10 compounds predicted to have $pK_i \geq 7.5$. The full set of training compounds had experimental activity with mean 6.9 ± 0.92 and median activity of 7.1. The RF prediction simply regressed to the wisest *a priori* population-based guess of activity for the most difficult compounds. The QMOD predictive methodology has no ability to make use of population-based information, but despite that, for these difficult compounds, made predictions that correctly identified most as highly active.

One of the surprising aspects of the results is that multiple approaches yielded quite similar population and correlation statistics in terms of the activities of the molecules chosen under different selection protocols. These approaches would all be reasonably characterized as working well on that basis. However, when considering the characteristics of the *structures* of the pool of potent selected molecules, very sharp differences arose.

4.4.3 Active Learning: Abstract versus Physical Models

What we have described in terms of explicit design bias toward novel compounds is related to other active learning approaches, both in the broader machine learning field as well as within computer-aided drug discovery (see the review by Kell[70] for a broad overview). Warmuth et al.[71] used active learning in combination with support-vector machine (SVM) classifiers to iteratively construct QSAR models with the goal of identifying active compounds quickly. They found that a selection strategy of seeking highly confident actives (similar to our potency selections) was effective for finding active ligands and that a strategy of decision-boundary selections was most effective for improving the QSAR models themselves. The study treated activity as a binary variable and did not structure the selection task temporally to mirror lead optimization. The focus was on activity alone and did not assess questions of structural diversity. Fujiwara et al.[72] studied active learning in the context of virtual screening and considered the question of structural diversity. As with the Warmuth study, compound activity was considered as a binary variable and temporal considerations were not taken into account. They showed advantages for combining a diversity-driven model building strategy with a selection method that sought new ligands on which different models produced maximally divergent predictions.

We have explicitly focused on procedures designed to mimic the constraints of a lead optimization exercise, with real-valued compound activities and temporally ordered chemical space exploration. Our direct comparison of the QMOD approach with a parallel random-forest approach exposed differences that relate to the assumptions underpinning a physical QMOD model compared with an abstract mathematical model. The central assumption made by machine-learning methods such as the random-forest approach or support-vector machines is that training and testing examples are drawn randomly from the same population. So, the distributional characteristics of the *activities* of molecules and of the *structural descriptors* are assumed to be the same. Under conditions where these assumptions are true, such methods can produce reliably accurate predictions, where the distribution of test errors will match estimates made by techniques such as cross-validation. The detailed algorithmic underpinnings of such methods actively "game" these assumptions, in order, for example, to reduce the effect of putative outliers in a training set on learned decision boundaries. However, in a lead optimization exercise, both the structural characteristics and activity profiles of compounds made *later* will be quite different (by design!) than those of compounds made *earlier*. With the RF approach, even when making active selection of structurally diverse molecules, *no increase* in structural diversity among the *highly active* selected molecules was observed (see Figure 4.9, red and blue curves in the left-hand plot).

In order for the iterative selection/test/refinement procedure to identify a pool of highly active molecules that are *also* structurally diverse, two things must be true. First, the selection strategy should incorporate structural diversity. Second, the predictive modeling method must be able to incorporate information from novel compounds so as to correctly identify new compounds that are both active and structurally novel compared with previously known actives. Recall from Figure 4.3, the structurally novel molecules included significant numbers with low activity. It is not enough merely to seek novelty in a selection procedure. The predictive models must be capable of making risky “bets” in order to discover a pool of highly active molecules that exhibit a wide range of structural characteristics. A pro-diversity bias alone, as with the novelty-biased RF method, does not guarantee a diverse pool of actives at the end of iterative lead optimization. The QMOD approach makes use of each training molecule to come up with a single physical model. A molecule whose high activity and unusual descriptors might be essentially “shrugged off” by an RF or SVM learning machine will be incorporated into a QMOD Pocket model in a manner that maximizes model parsimony while also explaining the high activity. Because the QMOD model is capable of correctly predicting activity values at or beyond the extremum observed during training, and because it may do so for structurally novel molecules, the iterative procedure that combined predictions of potency with selections of novel molecules produced a diverse pool of winners.

4.4.4 Relationship of the Induced Binding Pockets with the GyrB ATP Binding Site

The foregoing discussion has addressed questions about the numerical and structural qualities of the ligands produced by different selection schemes. While there were clearly benefits to the QMOD approach over the pure machine-learning RF method, perhaps the most salient advantage from a molecular design perspective is depicted in Figure 4.10. The QMOD approach induces the structure of an actual binding pocket, and that pocket has a direct relationship to the true biological active site that was responsible for the activity patterns observed. The QMOD pocket forms a funnel-like shape, with an open area corresponding to where solvent exists. Compound **20** is shown in its predicted conformation along with the experimentally determined one, reflecting no significant deviations and capturing all pendant conformational flips correctly.

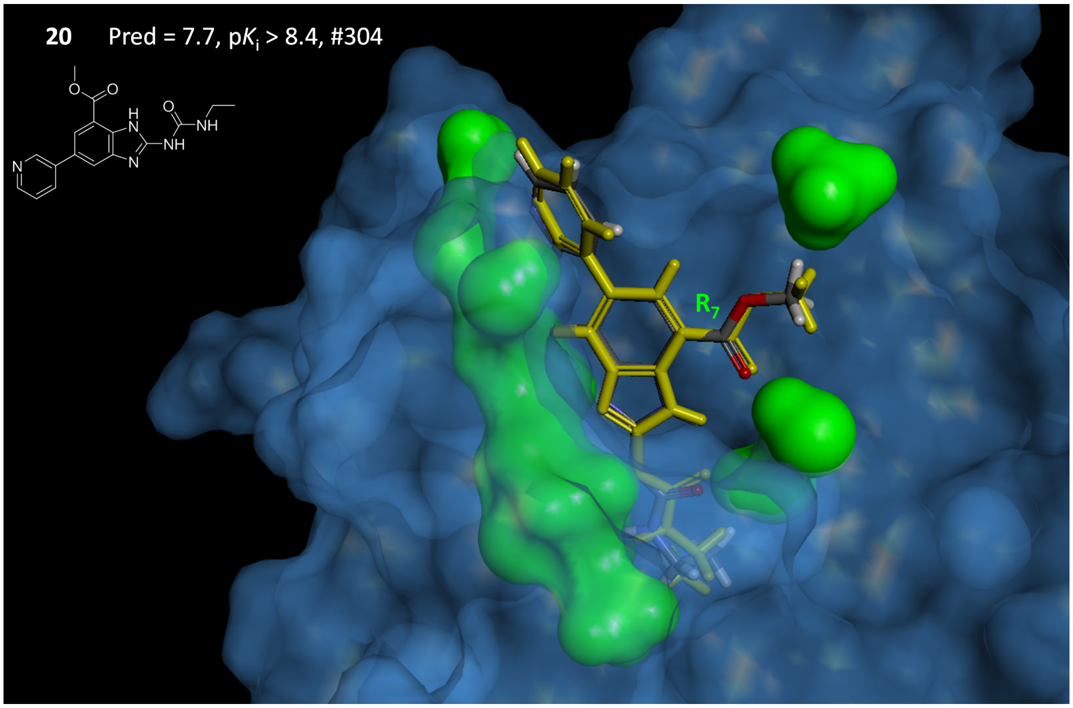


Figure 4.10: The relationship of the final QMOD standard pocket model to the GyrB binding site. Compound **20** in its optimal predicted QMOD pose (atom color) had RMSD of 0.5Å from the experimentally determined bound state (yellow). Alignment of the QMOD Pocket model and optimal ligand poses to the protein structure was done with a single alignment transformation that produced a close alignment of the benzimidazole inhibitor core. Configurational deviations are reflected primarily in the pendant moieties.

In total, 11 structures of bound inhibitors were aligned to one another based on protein pocket similarity,[73] and the predicted poses from the QMOD approach were compared to the bound configurations using the alignment from Figure 4.10. The predicted poses from the QMOD final Pocket model had mean RMSD of 1.2Å, with all but 2 having RMSD less than 1.5Å. Note that RMS deviation is somewhat difficult to interpret here. Barring a grossly different QMOD prediction of the benzimidazole core, which moved very little in the GyrB structures, the measured RMSD would tend to be relatively small. Another measurement of concordance between the Pocket model and protein compares the contact patterns for each ligand to the Pocket model or to the protein. The degree of concordance can be quantified by permutation of atom numbers. In all but three cases, there was a statistically significant relationship in the contact patterns ($p < 0.05$).

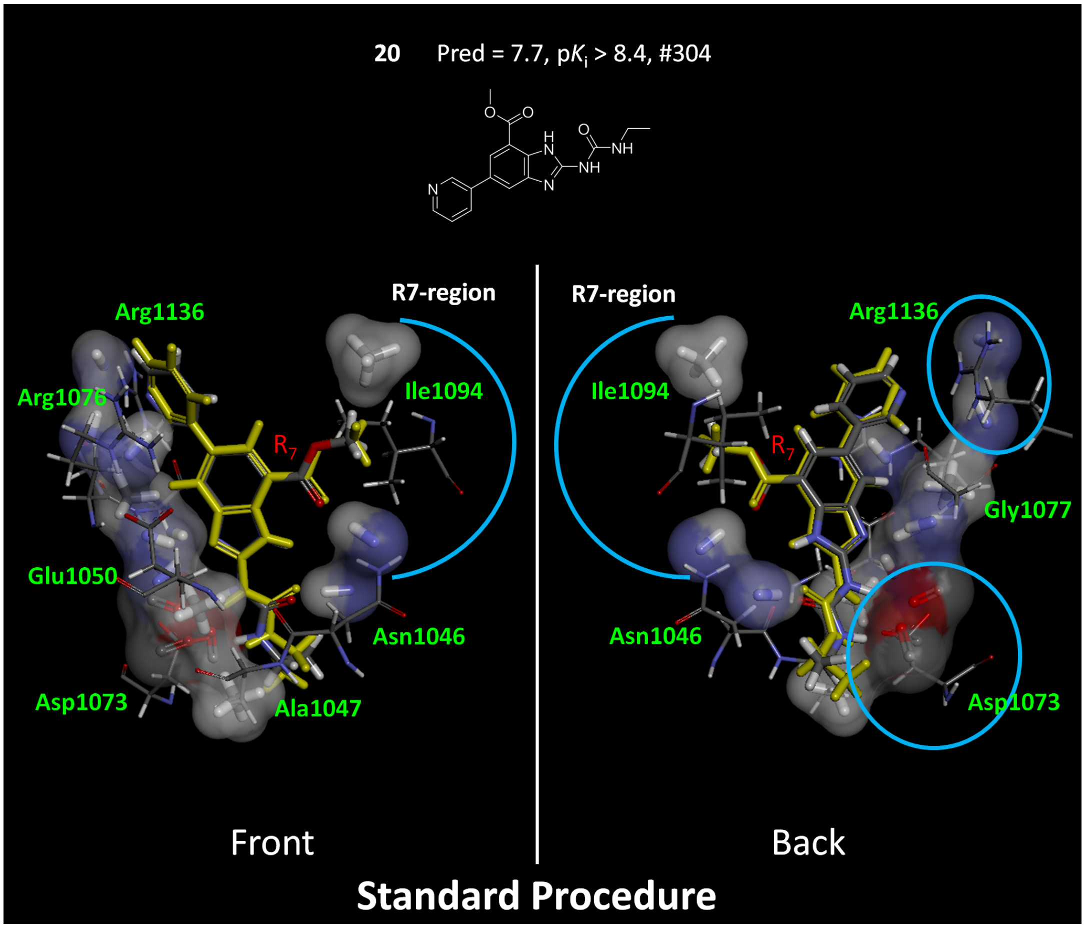


Figure 4.11: The QMOD standard procedure yielded a pocket model where there was a direct correspondence of many probes to particular atoms in the actual GyrB binding pocket. Pocket model probes that do not interact with compound **20** have been omitted from the display for clarity, and the protein has been trimmed to highlight areas of correspondence. The two views shown are flipped front to back.

Figure 4.11 shows additional detail, illustrating the direct correspondence between Pocket model probes and key moieties on the protein. The left-hand view highlights the reason behind the conformational choice for the methyl-ester substituent of compound **20**, which was correctly predicted (marked with a blue arc). The carbonyl ester oxygen makes a hydrogen bond with the N-H probe of the Pocket model, which parallels the same interaction with Asn-1046. The terminal methyl of the ester makes a hydrophobic interaction with a methane Pocket model probe, paralleling an interaction with Ile-1094. The right-hand view highlights two carbonyl probes that mimic the effect of Asp-1073 and two N-H probes that mimic Arg-1136. This degree of qualitative correspondence between Pocket model and protein is typical of our previous work.[59, 60]

Figure 4.12 shows the analogous depiction of compound **20**, but using the final QMOD Pocket model that arose from the control procedure. Recall that the structural variation of the final pool of potent selected ligands was much reduced and that the spatial probing of the binding pocket bordered by Asn-1046 and Ile-1094 was shallow (see Figure 4.8). The prediction for **20** was both numerically poor (low by 2.1 log units) and predicted the incorrect orientation of the 7-position methyl ester. The induced pocket here was unable to correctly accommodate the substituent, also showing a shift of the central scaffold away from its optimal position. While there were areas of good correspondence, especially with respect to the surface shape of the based of the binding pocket, the model induction process is sharply limited by the set of selected compounds. For the 11 inhibitors for which we had bound structures, just 3/11 had concordant contact patterns (compared with 8/11 for the QMOD standard predictions). In operational use of such modeling methods during lead optimization, mindful production of chemical variations that explicitly probe the "edges" of a model can produce significant improvements in the correspondence of refined models with biological reality.

For completeness, because we had *bona fide* structures of the GyrB binding pocket, we also made a comparison of the QMOD predictions to docking and scoring the final pool of unselected molecules. Using a single structure and the score of the top-ranked

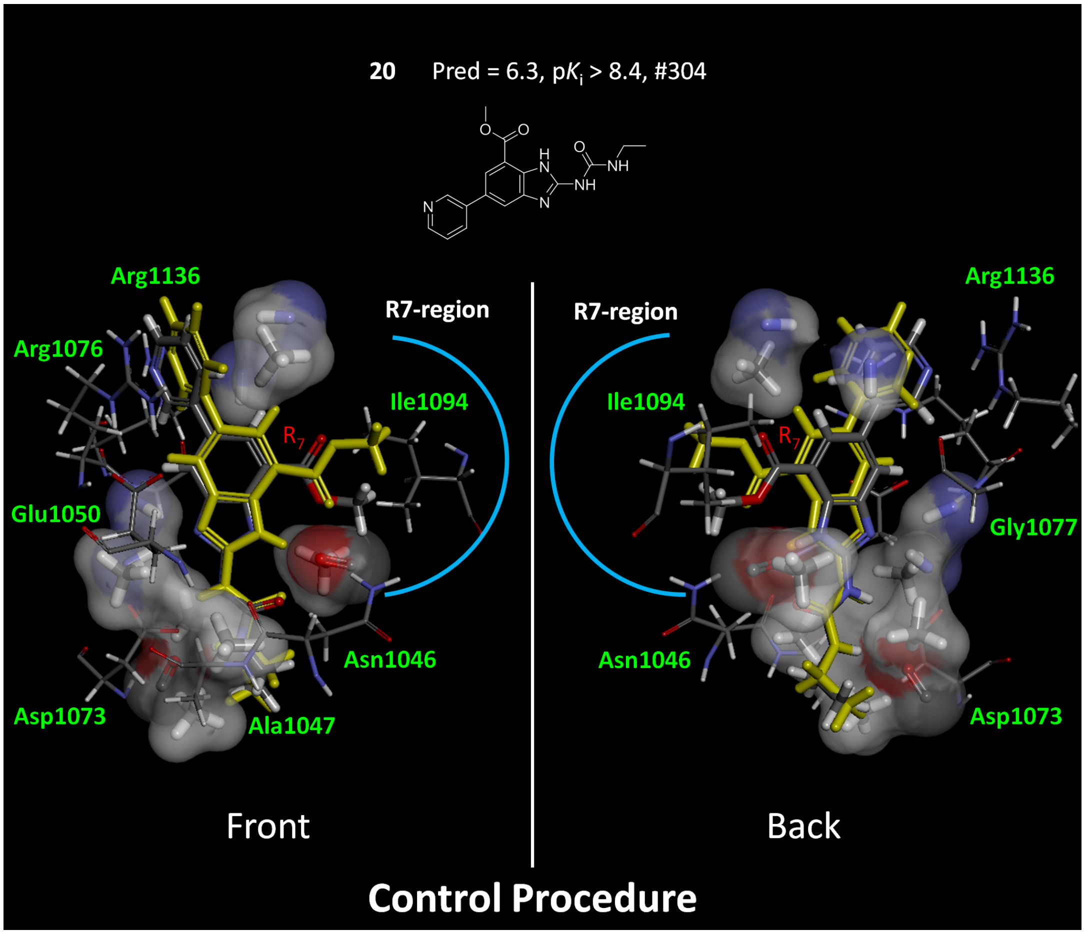


Figure 4.12: The QMOD pocket model that resulted from the procedure lacking an explicit novelty bias produced a poor prediction for compound **20**. The depiction here is analogous to that from Figure 4.11.

docking for each inhibitor did not produce a significant rank correlation. It is conceivable that a more sophisticated procedures such as MM-PBSA[74] might have yielded a reasonable correlation. Brown and Muchmore reported an average RMSE for predicted pK$_i$ using MM-PBSA on three targets of 0.75 (range 0.66–0.89) using linearly rescaled predictions to account for extreme slope and intercept deviations between computation and experimental values. The QMOD final standard model yielded 0.76 RMSE with no scaling correction on the 317 remaining unselected molecules, which is clearly comparable. Molecules pairs whose activity was different by 0.5 pK$_i$ units or greater were correctly ranked more than 70% of the time ($p << 0.01$). Rank correlation of this quality is challenging because over 80% of the experimental activity values fell within 1.5 log units of one another and over 90% within 2.0 units. It is encouraging that a method such as QMOD, with no information of any kind regarding either the bound configuration of ligands or of the actual binding site composition and geometry, could produce predictions of both activity and bound pose that are competitive with sophisticated structure-based methods.

4.5 Conclusion

This study has approached the QSAR modeling question in a novel manner. We explored how different computational selection strategies shaped the observed molecular trajectory derived from a sequence of hundreds of compounds that arose from a real-world lead optimization exercise. There were four primary results. First, the iterative QMOD procedure rapidly converged on models that reliably identified highly potent molecules. Second, explicit computational selection of novel molecules directly lead to a much more structurally diverse pool of potent inhibitors, despite not producing a pool with a different distribution of experimental activities than a control procedure with no novelty focus. Third, the induced binding site model showed strong concordance with the experimentally determined binding site, both in terms of absolute predicted poses as well as ligand/pocket contact patterns. Fourth, direct comparison with descriptor-based QSAR methods showed that while such models yielded similar distributions of activity among selected molecules, the structural diversity of selected potent molecules was much lower than for QMOD. QMOD identified examples of potent molecules across the entire arc of the project's chemical exploration, but the descriptor-based approaches instead produced many examples of highly similar minor variants clustered around the mid-point of the project's history.

There are two major lessons to be learned from this work, which we hope to further validate on additional systems in the future. First, there appears to be a significant hidden cost to reliance upon molecular design strategies that do not actively seek to probe new chemical functionality in a spatial sense. While such strategies may well identify compounds with desirable properties, they may completely miss the identification of entire classes of active compounds. Here, for example, potent activity against GyrB *and* ParE was exhibited by compounds discovered through the selection procedure that sought three-dimensional structural novelty in order to test the physical boundaries of the evolving models. Second, statistical regression methods whose fundamental basis for prediction relies upon correlations between features and desired output values impose hidden costs. They do so by being strongly dependent upon the existence of near-neighbors with known activity in order to accurately predict a new compound to have similar activity. In molecular potency optimization, effort is often placed on design goals toward or even beyond the extreme end of the distribution of known molecular activities. Truly potent molecules that are structurally novel in the descriptor space being used by a correlative machine will be *underpredicted* as a consequence of the gaming strategy employed by statistical regression methods.

The issues of confirmation bias and correlation fallacies discussed in a recent perspective[66] appear *naturally* in the iterative application of predictive modeling for design of potent molecules. Given a method that depends on non-causative correlations to predict potency, selection of the molecules predicted to be potent will tend to *automatically* self-confirm, because only those candidate molecules that are highly similar to known molecules with high potency will tend to be top-ranked. The structurally novel compounds that would have been shown to be potent remain *invisible* in practice, because they will have been predicted to have middling potency. In typical machine-learning problems, inductive bias issues will show up in the distribution of prediction errors on different types of test objects. In the case of medicinal chemistry lead optimization, such bias issues may altogether *suppress* the production of non-confirmatory test objects.

By making use of a different molecular selection strategies, each of which is nominally equally accurate in aggregate behavior, very different outcomes will arise from repeated temporal iteration. The resulting molecules having the high potency sought during optimization will reflect the hidden or explicit biases embedded in the predictive modeling approaches. An approach whose basis for prediction is a close match to the protein ligand binding process, coupled with an explicit selection strategy designed to expand model coverage, will tend to identify a diverse pool of molecules. The structural diversity will most likely manifest itself in properties that were not directly optimized. When making use of purely correlative learning machines, the unseen cost can manifest itself as a numerous but narrow pool of molecules. Given the challenging problem of drug discovery, we would argue that generation of a diverse pool is generally the more desirable outcome.

Chapter 5

Structure-Guided Quantitative Modeling

5.1 Abstract

Binding affinity prediction is frequently addressed using computational approaches demonstrated through cross-validation within a series of molecules or through performance shown on a blind test set. Here, we present a demonstration on how such a system performs when employing an structure-guided modeling strategy that leverages molecular similarity, docking, and multiple-instance learning in cases with limited protein structures but with high ligand diversity. Many QSAR methods have utility in making predictions within a highly related chemical series, but cannot generally be fruitfully applied to novel compounds due to limited domains of applicability. Molecular docking has found utility in applications such as virtual screening, off-target prediction, and structure-based modeling, however it is not generally reliable for affinity prediction toward lead optimization. Previous reports of the Surflex-QMOD approach demonstrated its ability to produce accurate and scaffold-independent predictions of binding affinity by constructing an interpretable physical model of a binding site based solely on the structures and activities of ligands. Here we introduce an enhanced QMOD method demonstrating its ability to integrate protein structure information

as well as ligand structure-activity relationships to construct more robust physical models. The derived structure-guided models are capable of accurately predicting binding affinities over a broad class of compounds while producing more physically accurate representations of the protein pockets and ligand binding modes. Results will be presented establishing significant performance improvements in binding pocket induction and ligand affinity predictions in cases with limited protein structures but with high ligand diversity.

5.2 Introduction

The preceding chapters discussed in detail the underpinnings driving QMOD model induction and its application towards guiding molecule selection/design in a lead optimization effort. The present study considers these aspects of predictive activity modeling but adds new dimensions. Rather than focus purely on how well a method can predict activity based on solely ligand structure-activity data, we instead ask how a method can *integrate* information of protein structures and ligand structure-activity relationships. Such a strategy has produced predictive models that are more widely applicable and accurate for ligand affinity prediction while providing a rational representation of the binding pocket that explains binding activity. Further, in addition to considering prediction accuracy and broad applicability of discovering active compounds, we consider how information usage and modeling methods affect the predictive power attained by a model. We show that there is a direct benefit when integrating structural information of a protein-ligand binding event that guides towards more generalizable predictions of molecules of different types. Traditional use of information can lead towards nominally accurate activity predictions in the case of purely ligand-based modeling within a congeneric series or molecular docking in the context of rank ordering and pose prediction. Integrating protein structures and ligand structure-activity relationships can enable the predictive accuracy necessary for lead optimization while providing broader applicability and increased predictive performance on structurally diverse molecules.

The present work addresses a new area of data integration that utilizes physical constraints provided by protein structures and structure-activity relationships of known competitive small molecules. We examined the performance of such a method in a lead optimization scenario with an eye towards accurate prediction of cross-chemotype ligands on a large scale and the physical relationship between an induced model and a binding site in challenging cases of significant protein flexibility.

A set of 80 congeneric CDK2 inhibitors[75–77] and 26 temporally consistent x-ray crystal structures[78] offered the ability to consider an structure-guided modeling approach for which the therapeutic target of interest provided well studied structure-activity relationships and protein structures. The series was split between 30 for training and 50 for testing, and five protein structures were chosen based on their combined coverage of the structural diversity present among the pool of proteins considered. In addition, a set of 56 PDB co-crystal structures of CDK2 bound to non-covalent inhibitors was identified from Binding MOAD[79] and were mutually aligned in order to provide a direct comparison between the QMOD-generated model and the actual CDK2 binding site under normal conformation variation. A second set consisting of the first crystallographically determined structure of the adenosine A_{2A} receptor (PDB code: 3EML[80]) and 93 A_{2A} antagonists offered an opportunity to examine the integrated modeling approach on a much more challenging case with a therapeutic relevant target for which limited structural information was available. The structural flexibility of the single A_{2A} structure was computationally explored and provided a conformational ensemble from which 5 structures were chosen based on combined structural coverage. The A_{2A} inhibitors were organized temporarily and split between 63 used for training and 30 used for testing. Model predictive performance was examined with respect to activity prediction and model coverage of the A_{2A} binding site.

Figure 5.1 provides an example of the CDK2 data set. Molecules **1-3** are N2, O6 substituted guanines and are the three most active compounds in the CDK2 congeneric series. Molecules **4-9** are molecules inherited from the 5 crystal structures chosen for structural guidance (1QMZ, 1KE6, 1KE8, 1H07, 1JVP). We considered the effect of utilizing protein structures and bound ligand geometries on model induction and predictive performance on the 50 congeneric test set and more widespread applicability on the set of 56 diverse CDK2 inhibitors for which x-ray crystal structures were available for post-facto evaluation. Note molecule number references do not reflect a relationship to the gyrase ligand numbers in the preceding two chapters.

Figure 5.2 shows a diagram of the structure-guided QMOD procedure. QMOD employs a multiple-instance machine learning procedure for model induction where optimality involves fitting the pocket model to binding activity data *and* fitting ligand poses to the model. The QMOD procedure has been described in previous work[59–61, 66] and a brief description here will highlight new developments of the structure-guided approach. The structure-guided procedure begins with multiple protein structures with bound ligands that provide insights to the conformational variability that may exist between the protein pocket and bound molecules. An alignment hypothesis is generated that aims at determining plausible bioactive poses of the ligands used for model induction, the N2, O6 substituted guanines in this case. This is accomplished by docking the most active training ligands (molecules **1-3**) against the 5 representative crystal structures and caching the top 100 highest scoring poses for each ligand. A single docked pose for each molecule is selected such that the combined mutual 3D similarity with the bound crystal ligands is maximized, yielding an alignment

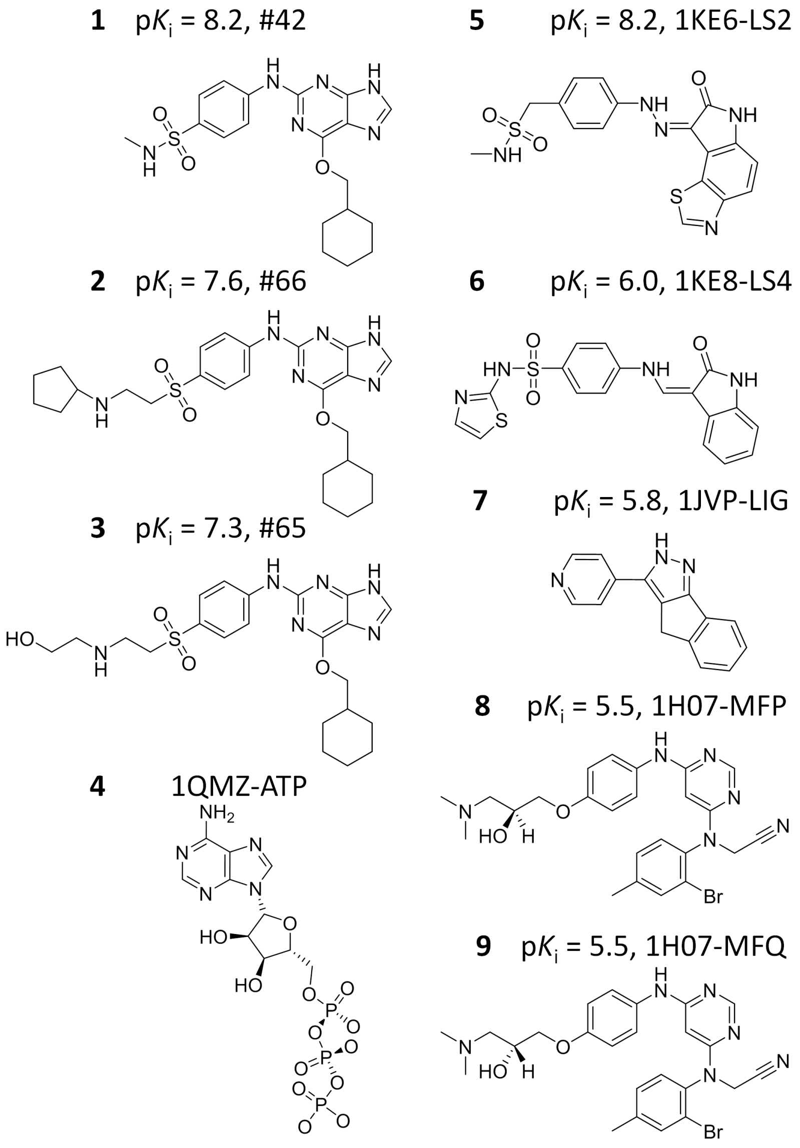


Figure 5.1: CDK2 Ligands used for the alignment hypothesis. Molecules **1-3** are the top 3 most active ligands derived from the CDK2 congeneric series used in this study. Molecules **4-9** are the bound ligands extracted from the 5 protein structures chosen for structural guidance. Crystal ligands for which binding activity was readily available (compounds **5-9**) were included in the training set.

hypothesis that is informed by optimal fitting with the proteins and 3D geometric concordance with bound crystal ligands. The hypothesis alignment seeds the alignment of the remaining training ligands, producing several hundred pose variants per molecule. A rich set of molecular fragments (probes) are generated providing a physical view of where the protein may exist. Probes types include hydrophobic (methane), donor (N-H), and acceptor (C=O). The initial probe set is generated using the standard tessellation procedure that positions probes on the surface of the initial ligand alignment, followed by a process that removes probes that are not within the vicinity of similar type fragments observed in the protein binding pockets. A comparison of panels A and D in Figure 5.2 reveals good overall coverage of the binding site between the protein structures in A and the initial probe set in D. Regions at the front and right-side are adequately covered by the initial rich probe set, while the hinge-biding region at the top is correctly represented by steric and polar probes, and the opening of the pocket at left remains unoccupied. The standard model induction procedure proceeds with a probe search and iterative probe and ligand pose refinement until a fully refined pocket model (termed pocket model) is derived (Figure 5.2E). Testing of new molecules is carried out by flexibly fitting the ligand into the pocket model while keeping the model fixed (Figure 5.2F).

There were three primary results of this study. First, the structure-guided QMOD procedure produces models that are highly predictive within a congeneric series in two contrasting cases where structural information was abundant, and where structural information was severely limited for a suspected highly flexible pocket. The structure-guided procedure performed comparatively well with the purely ligand-base approach with respect to affinity prediction and rank ordering of the CDK2 congeneric test series. In the more challenging A_{2A} case the structure-guided procedure performed significantly better than standard ligand-based approach, highlighting the benefit of integrating structural information in a case where protein flexibility is likely an important contributing factor in accurate activity and pose prediction. Second, the structure-guided modeling procedure is more widely applicable and accurate in activity and pose predictions across a wide variety of *structurally diverse* molecules.

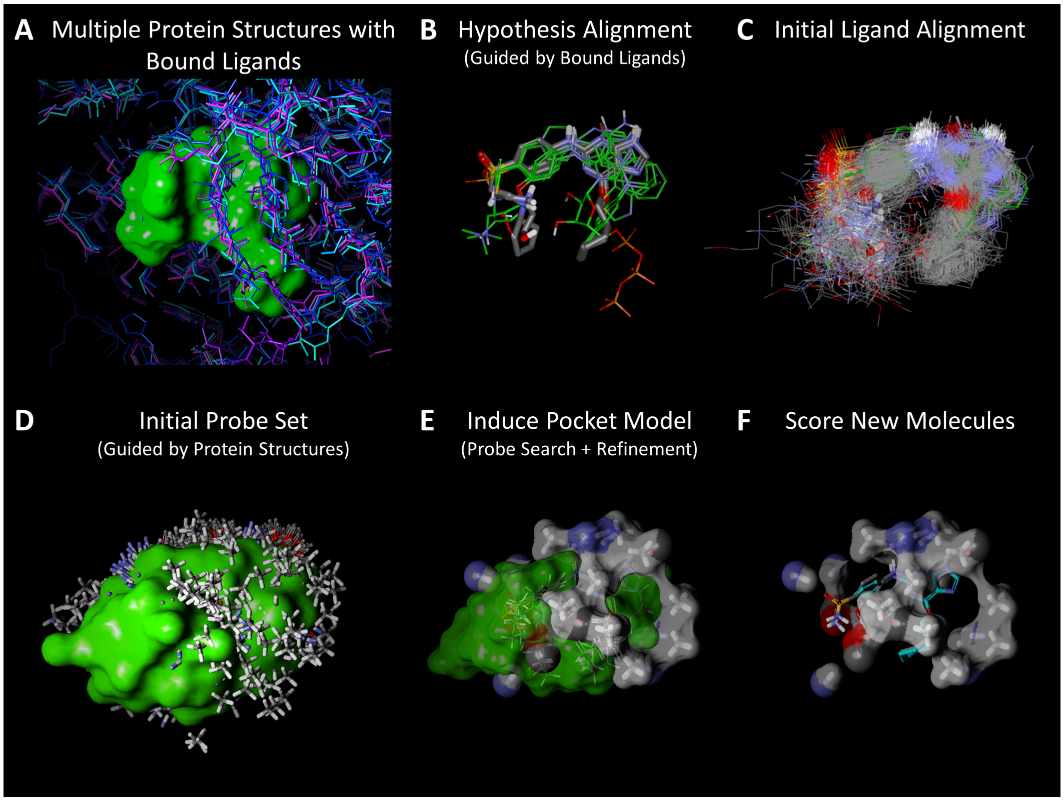

A Multiple Protein Structures with Bound Ligands
B Hypothesis Alignment
(Guided by Bound Ligands)
C Initial Ligand Alignment
D Initial Probe Set
(Guided by Protein Structures)
E Induce Pocket Model
(Probe Search + Refinement)
F Score New Molecules

Figure 5.2: Derivation and testing of a QMOD pocket model proceeds in six automated steps: A) collection of multiple protein structures with bound ligands; B) an alignment seed hypothesis is constructed from the 2-3 most active ligands, guided by their fit to the protein pockets and similarity to corresponding bound poses; C) 100-200 alignments for each training ligand are produced; D) a large set of probes (many thousands) is created where interactions may exist, spatial arrangement is guided by location of similar type fragments in the protein pockets; E) a small near-optimal set is selected based on fit to experimental binding data and model parsimony, followed by iterative probe and pose refinement; F) new molecules are tested by flexible alignment into the pocket to optimize score. The final pocket model is used in a fixed configuration, but conformational flexibility within the corresponding protein pocket is represented by probes being places in multiple positions.

On the structurally diverse CDK2 set the structure-guided QMOD procedure outperforms the standard ligand-based QMOD procedure with respect to rank correlation and activity prediction error. The structure-guided procedure performs equivalently well in ranking diverse molecules compared to molecular docking, but provides the additional benefit of lower prediction errors with better predicted pose performance. Third, the structure-guided procedure produces models that shared high physical concordance with the protein targets under investigation. In the CDK2 case the induced model showed a direct relationship with key binding site elements known for their role in ligand recognition. In the more challenging A_{2A} case, the induced model showed a direct correspondence to the shape and electrostatic characteristics of the pocket while providing a testable hypothesis of protein flexibility and specific interactions with ligand moieties.

The Surflex-QMOD methodology has been validated in prior studies.[59,61,66] The significance here relates to strategic data integration in the context of protein modeling and ligand affinity prediction with an emphasis on binding site elucidation in the presence of structural variability and cross-chemotype predictions on a scale that pushes the boundaries of applicability of traditional 3D-QSAR and molecular docking approaches. There is a dramatic benefit in making use of protein structural information in the presence of significant protein flexibility and ligand structural diversity.

In the case of the congeneric chemical series studied here, it was not surprising that the structure-guided QMOD procedure performed competitively well with the purely ligand-based approach in a purely numeric sense with respect to prediction errors and activity ranking. However, the benefit of utilizing protein structures manifested itself with more accurate representations of the physicality of the protein-ligand binding event. The structure-guided procedure also predicted the structurally diverse molecules more accurately than the standard ligand-based procedure and ranked such molecules equivalently well as molecular docking while yielding more accurate predictions of binding pose. The structure-guided modeling procedure demonstrated the ability to leverage the strengths of both purely ligand-based and structure-based approaches. Structure-guided modeling provides the level of affinity prediction accuracy necessary for lead optimization and more wide spread applicability of ligand ranking and pose prediction on structurally diverse molecules. In addition, the structure-guided procedure provided a physically realistic model of the binding pocket that intuitively highlights physical characteristics of the pocket important for ligand binding.

We believe that this approach of studying protein modeling and affinity prediction, subject to the integration of different computational techniques, offers a means by which to assess the real-world behavior of modeling systems. The results clearly encourage the use of physically sensible approaches that maximizes information usage by means of rational data integration.

5.3 Methods

5.3.1 Quantitative Modeling with Structural Guidance

The structure-guided QMOD procedure is initially guided by structural information through ligand guidance during the alignment hypothesis generation. The top three most active ligands are computationally docked to the protein structures and a single pose for each ligand is chosen such that the combined mutual similarity with the crystal bound ligands are maximized. Figure 5.3 shows the hypothesis alignment of the top 3 most active CDK2 ligands used for model induction. Panel A shows the selected poses in comparison with a bound ligand (compound **5**) derived from one of the protein structures used as input (PDB code: 1KE6). This example highlights the high structural concordance provided by the matching polar moieties at the top-right, matching ring orientation at center, and matching shape and electrostatically exposed profile at the left. Panel B shows the the same alignment hypothesis in comparison to the bound pose of a close analogue (compound **10**). This example shows the structural

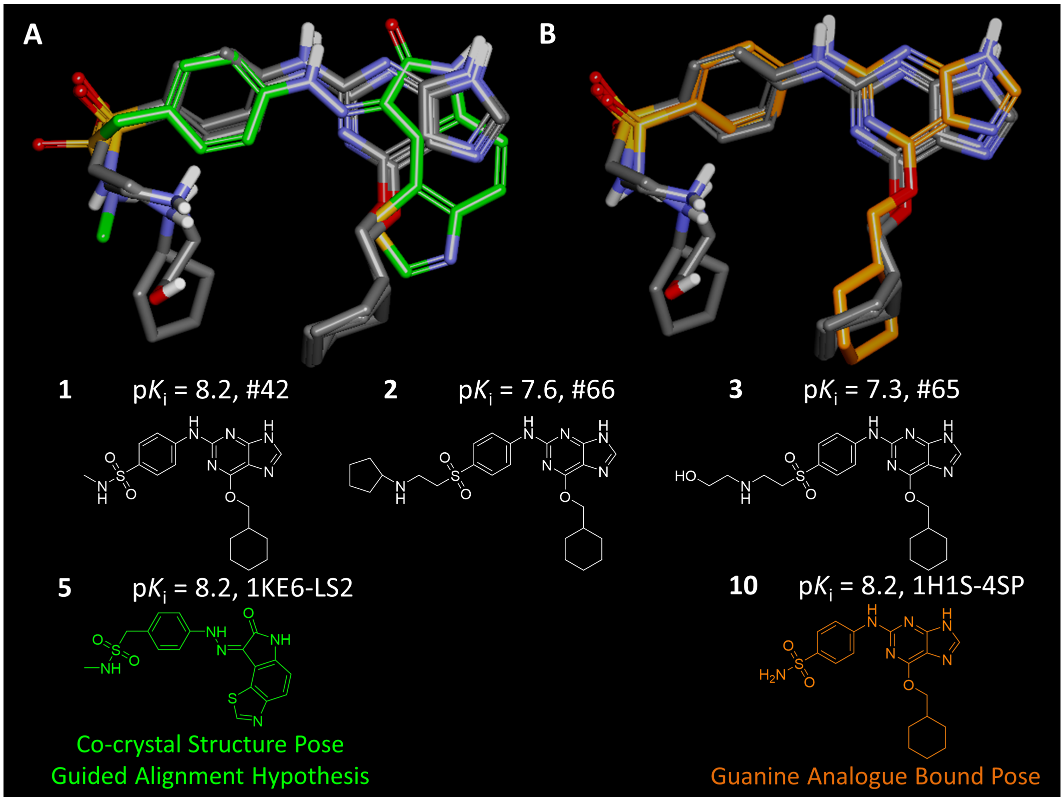


Figure 5.3: Alignment hypothesis yields conformational concordance among highly active CDK2 ligands while satisfying physical constraints of observed bound conformations: A) the hypothesis alignment of the top three most active CDK2 ligands (**1-3**) with crystal structure bound pose of **5**.; B) hypothesis alignment of ligands **1-3** with bound pose of structurally related analog **10** (PDB code: 1H1S). Compound **10** was not used during the hypothesis alignment generation.

relevance provided by a structure-guided alignment hypothesis procedure. Molecule **10** was not used during the hypothesis alignment generation or model induction, but served as an excellent validation of the structure-guided alignment hypothesis procedure. Model induction was carried out as previously described (see Figure 5.2) and predictive performance was evaluated.

5.3.2 Initial Probe Generation Guided by Protein Structures

A rich set of molecular fragments (probes) are generated providing a physical view of where the protein may exist. The initial probe set is generated using the standard tessellation procedure that positions probes on the surface of the initial ligand alignment, followed by a process that removes probes that are not within the vicinity of similar type fragments observed in the protein binding pockets. A detailed comparison of the top-left and bottom-right panels in Figure 5.4 reveals good overall coverage of the binding site between the protein structures and the filtered probe set. Regions at the front and right-side are adequately covered by the initial rich probe set, while the hinge-biding region at the top is correctly represented by steric and polar probes, and the opening of the pocket at left remains unoccupied.

Figure 5.5 provides a detailed look at the hinge binding coverage retained by the filtered probe set. Shown on the right is a zoomed-in view of the top portion of the pocket revealing a multiplicity of polar probe positions covering residues Glu81 and Leu83 in the hinge binding region.

The standard model induction procedure proceeds with a probe search and iterative probe and ligand pose refinement until a fully refined pocket model is derived (see Figure 5.2E). Testing of new molecules is carried out by flexibly fitting the ligand into the pocket model while keeping the model fixed (see Figure 5.2F).

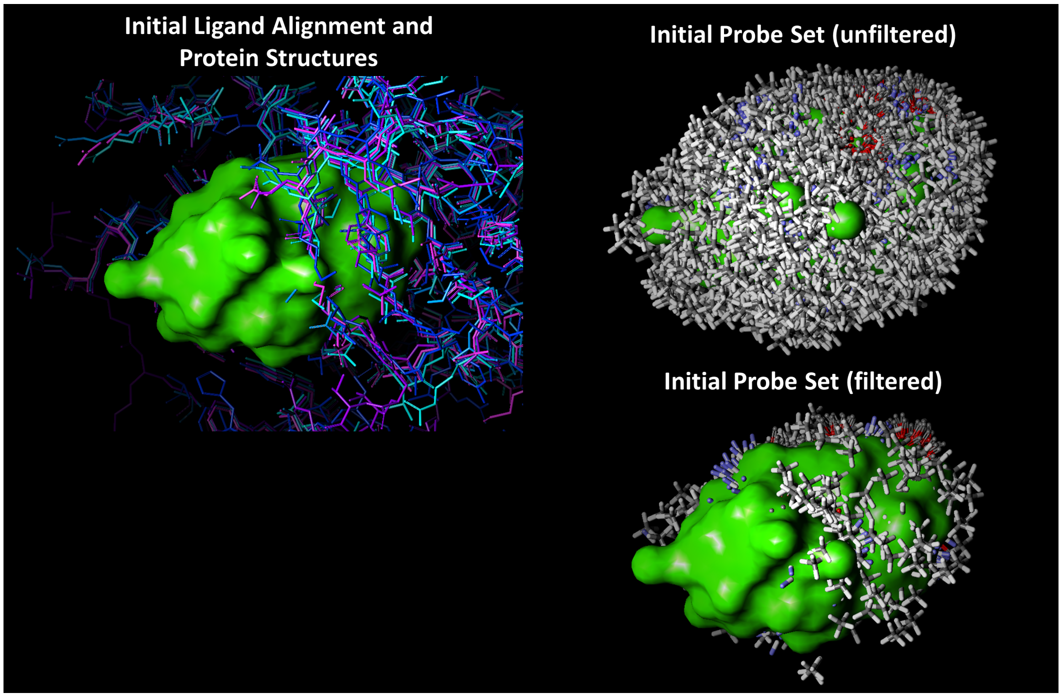

Initial Ligand Alignment and Protein Structures
Initial Probe Set (unfiltered)
Initial Probe Set (filtered)

Figure 5.4: The structure-guided initial probe set provides good coverage of the CDK2 binding pocket. At the top-left shows the 5 CDK2 crystal structures used for model guidance in comparison to the volume occupied by the initial ligand alignment. At the top-right shows the initial probe set immediately following the standard tessellation procedure and at bottom-right shows a sample of the probes remaining after filtering. The filtered probe set provides good overall coverage of the pocket that mimics physical constraints provided by the protein (top-left) while leaving unoccupied regions exposed.

5.4 Results and Discussion

5.4.1 Effects of Integrating Structural Information on Model Predictive Performance

As described above (and shown in Figure 5.2), model induction followed a hybrid approach utilizing molecular docking and 3D similarity during the alignment hypothesis generation. Protein structural guidance was enforced during the initial probe pool generation, and the procedure followed the previously established QMOD learning

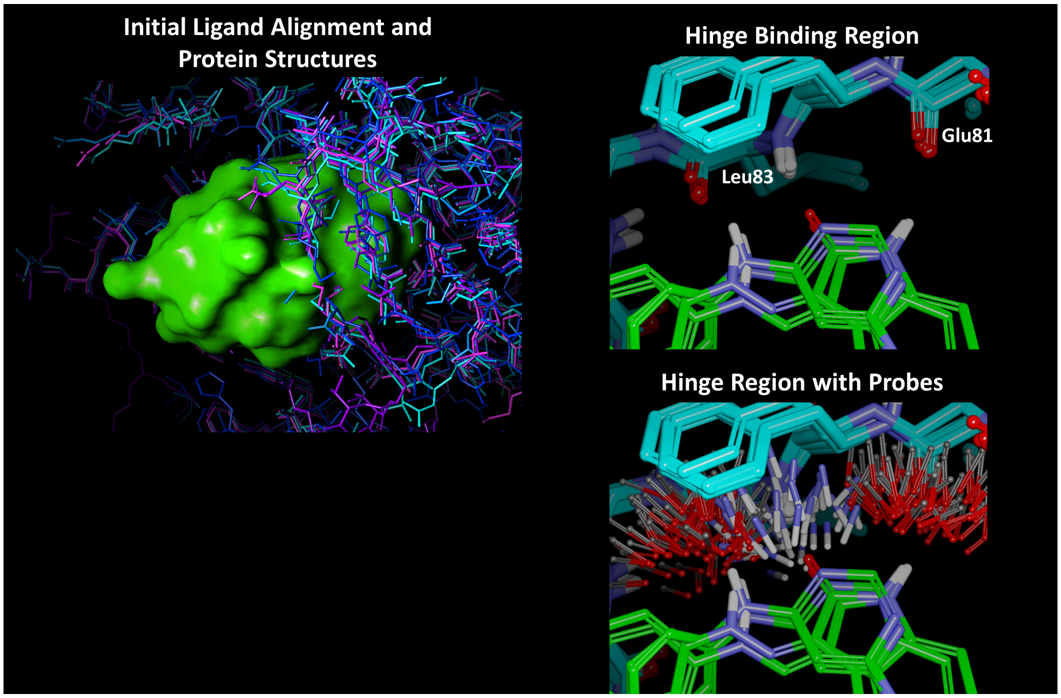


Figure 5.5: The structure-guided initial probe set provides good coverage of hinge binding region known to play a critical role in ligand binding in the ATP binding site of CDK2. Shown on the right is a zoomed-in view of the top portion of the pocket revealing a multiplicity of polar probe positions covering residues Glu81 and Leu83 in the hinge binding region.

protocol to produce a physically realistic model of the protein binding pocket. Figure 5.6 shows the final pocket model at left and prediction performance of the 50 CDK2 congeneric compounds at right. The model was highly predictive within this congeneric series, producing an average error of 0.61 log units and a Kendall's Tau rank correlation of 0.73 ($p < 0.01$). At left shows the final pocket model with the predicted pose of molecule **12** with high confidence stemming from training molecule **11**. The predicted activity of compound **12** was 7.7, a 0.5 log unit deviation of its pK_i of 7.2. At right shows the overall prediction performance on the entire set, highlighting an excellent correlation between the predicted and experimentally determined binding activities.

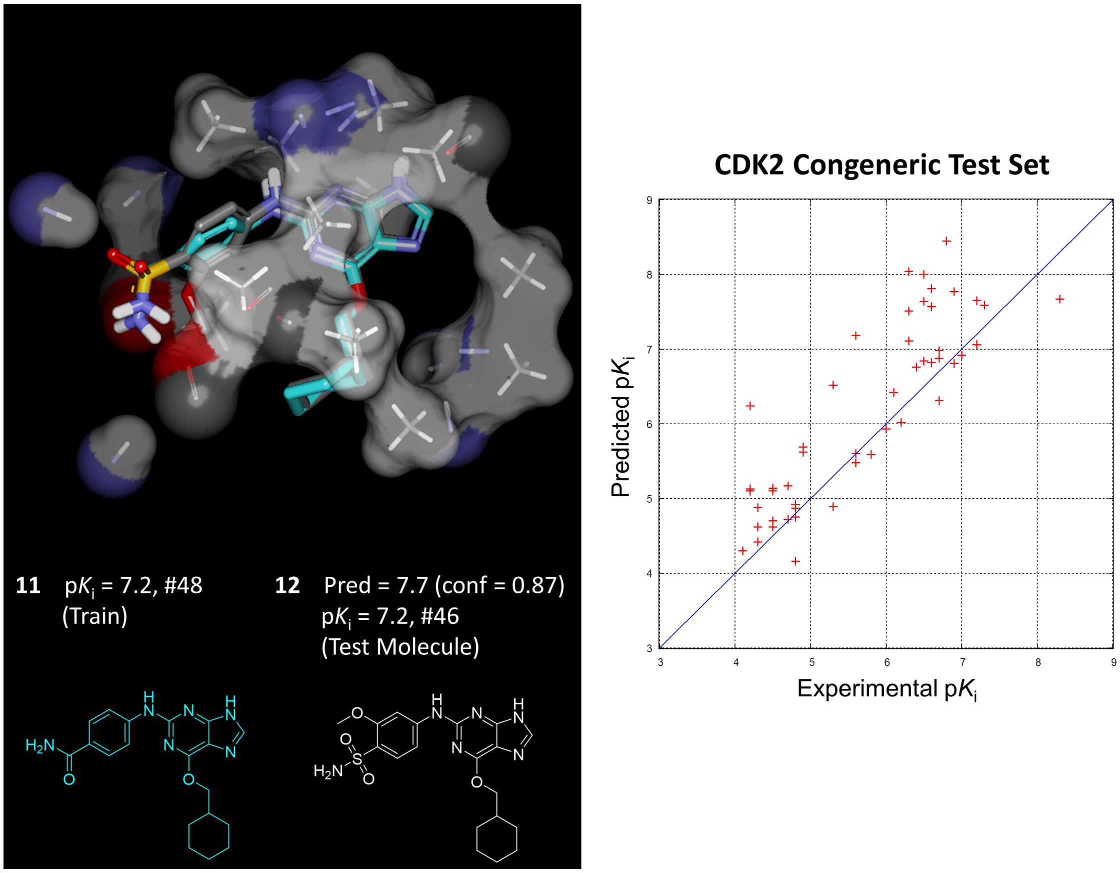


Figure 5.6: The induced CDK2 pocket model produced nominally accurate predictions within the CDK2 congeneric series. At left shows the final pocket model in thin sticks and skin with molecule **12** (atom-colored sticks) in its final predicted pose with high confidence derived from training molecule **11** (cyan). The predicted activity of **12** was 7.7, a 0.5 log unit deviation of its pK_i of 7.2. At right shows the activity prediction performance on the entire congeneric test set. The overall prediction error was 0.61, with a Kendall's Tau rank correlation of 0.73 (p $<$ 0.01, by permutation analysis), and an R^2 of 0.71.

In addition, 7 out of the top 10 *confidently* predicted most active test molecules appeared among the top 10 *bonafide* most potent molecules in the entire test set. Table 5.1 shows a detailed breakdown of the prediction performance on the 50 blind test molecules with the top 10 confidently predicted most active ligands highlighted with a bold underline.

Table 5.1: Performance of the CDK2 pocket model on the 50 blind congeneric ligands (Compound numbers as listed in Reference 3)[a]. 7 out of 10 of the *confidently* predicted most active molecules (boldface underlined) were among the top 10 *bonafide* most active ligands.

Rank	Mol.	Exptl.	Pred.	Error	Conf.	Rank	Mol.	Exptl.	Pred.	Error	Conf.
1	**29**	8.3	7.7	0.6	0.91	26	55	5.8	5.6	0.2	0.72
2	64	7.3	7.6	0.3	0.74	27	33	5.6	7.2	1.6	0.51
3	**46**	7.2	7.7	0.5	0.87	28	34	5.6	5.6	0.0	0.84
4	**44**	7.2	7.1	0.1	0.87	29	31	5.6	5.5	0.1	0.75
5	**45**	7.0	6.9	0.1	0.88	30	56	5.3	6.5	1.2	0.83
6	59	6.9	7.8	0.9	0.74	31	30	5.3	4.9	0.4	0.67
7	**53**	6.9	6.8	0.1	0.91	32	80	4.9	5.7	0.8	0.49
8	58	6.8	8.5	1.7	0.77	33	36	4.9	5.6	0.7	0.78
9	**78**	6.7	9.1	2.4	0.83	34	20	4.8	4.9	0.1	0.63
10	**47**	6.7	7.0	0.3	0.84	35	27	4.8	4.9	0.1	0.51
11	**50**	6.7	6.9	0.2	0.91	36	7	4.8	4.8	0.0	0.67
12	49	6.7	6.3	0.4	0.86	37	15	4.8	4.2	0.6	0.46
13	70	6.6	7.8	1.2	0.50	38	17	4.7	5.2	0.5	0.44
14	74	6.6	7.6	1.0	0.73	39	19	4.7	4.7	0.0	0.44
15	71	6.6	6.8	0.2	0.70	40	24	4.5	5.1	0.6	0.55
16	69	6.5	8.0	1.5	0.71	41	16	4.5	5.1	0.6	0.60
17	**60**	6.5	7.6	1.1	0.82	42	18	4.5	4.7	0.2	0.39
18	72	6.5	6.8	0.3	0.44	43	13	4.5	4.6	0.1	0.64
19	37	6.4	6.8	0.4	0.83	44	3	4.3	4.9	0.6	0.64
20	63	6.3	8.0	1.7	0.75	45	10	4.3	4.6	0.3	0.61
21	**62**	6.3	7.5	1.2	0.80	46	2	4.3	4.4	0.1	0.65
22	61	6.3	7.1	0.8	0.71	47	26	4.2	6.2	2.0	0.51
23	41	6.2	6.0	0.2	0.84	48	23	4.2	5.1	0.9	0.63
24	52	6.1	6.4	0.3	0.46	49	11	4.2	5.1	0.9	0.47
25	28	6.0	5.9	0.1	0.77	50	4	4.1	4.3	0.2	0.61

[a] Experimental, predicted, and error values are units of pK_i.

The significant benefit of integrating protein structural information with the QMOD learning procedure manifested itself with excellent predictive performance on the 56 diverse CDK2 inhibitors. Predictive performance was quantified by deviations of affinity predictions from known pK_i measurements, comparative ranking of molecules by predicted affinity and pK_i, and root-mean-squared deviations between the model's predicted binding pose compared to bound ligand poses determined by x-ray crystallography. On the entire test set the structure-guided model yielded an average activity prediction error of 1.0 log unit and an average RMSD 1.84Å, with a Kendall's Tau rank correlation of 0.26 ($p < 0.01$). The model performed systematically better on ligands for which confidence was higher. At confidence levels of 0.5 and higher the

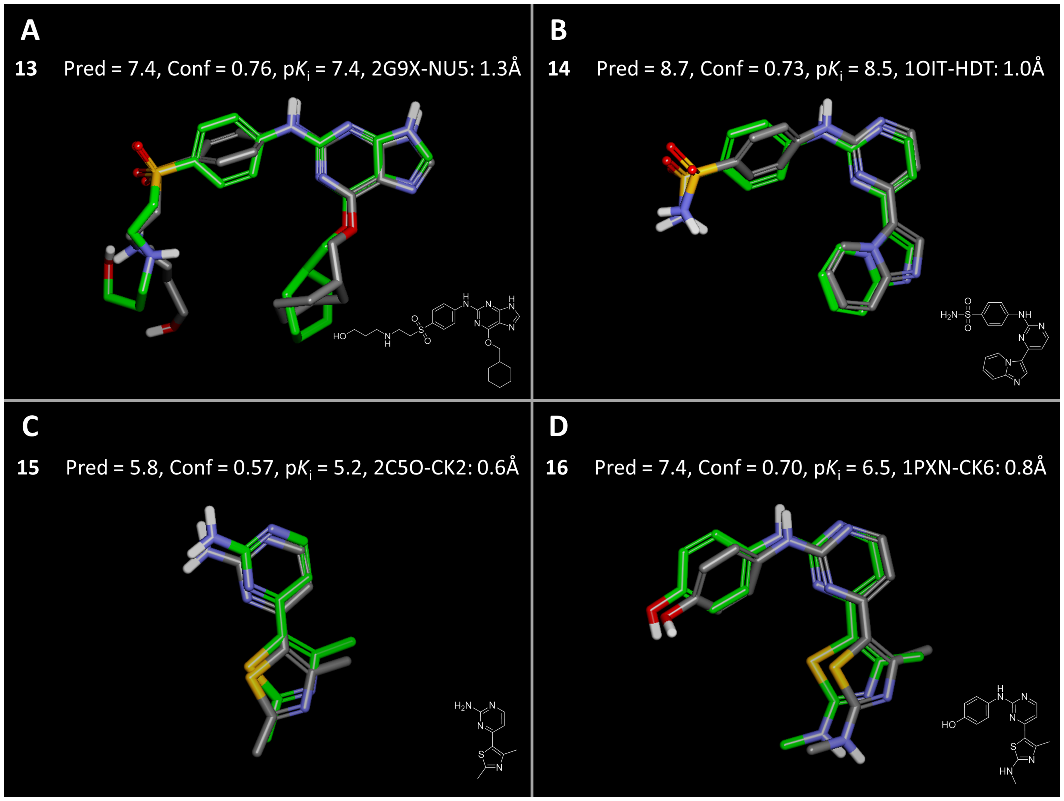

A
13 Pred = 7.4, Conf = 0.76, pKi = 7.4, 2G9X-NU5: 1.3Å
B
14 Pred = 8.7, Conf = 0.73, pKi = 8.5, 1OIT-HDT: 1.0Å
C
15 Pred = 5.8, Conf = 0.57, pKi = 5.2, 2C5O-CK2: 0.6Å
D
16 Pred = 7.4, Conf = 0.70, pKi = 6.5, 1PXN-CK6: 0.8Å

Figure 5.7: The structure-guided modeling procedure produced accurate pose and activity predictions on the diverse set of 56 CDK2 inhibitors. Panels A-D show four examples of diverse CDK2 ligands in their predicted poses (atom-colored) superimposed with their crystal structure bound pose (green).

model produced an average error of 0.9 and an average RMSD 1.57Å, with a Kendall's Tau rank score of 0.32 ($p < 0.01$). At confidence levels of 0.7 and above the model yielded an average error of 0.7 and an average RMSD of 1.2Å, with a Kendall's Tau ranking score of 0.74 ($p < 0.01$). Figure 5.7 shows examples of predictions made on the diverse test set, highlighting excellent performance on predicted affinity and bioactive geometry.

5.4.2 Relationship of the Induced Pocket Model with the Cyclin-dependent Kinase 2 ATP Binding Site

In the forgoing we have discussed the effects of integrating protein structural information on model performance with respect to accurate prediction of ligand activity and bioactive pose with the protein pocket. Another attribute worth evaluating is the physical relationship of the induced model with the protein binding pocket. The pocket model presents a view of the protein that is directly informative of the physical characteristics that best explain the activities of known ligands (i.e. the training set) while highlighting pocket elements that enable broader compatibility with diverse ligands exhibiting potentially significant structural novelty. Figure 5.8 highlights the physical relationship between the model and the crystallographically determined binding site of 2G9X with a bound ligand. Key interaction points on the hinge binding region are well represented by the induced pocket model. Electrostatic interactions provided by Asp86 are modeled by an acceptor probe, and two hydrophobic probes flanking the right and left sides of compound **13** closely match physical constraints provided by Asp86 and ILe10 (Figure 5.8A). The backbone carbonyl of Glu81 is modeled by an acceptor probe and the NH of Leu83 is represented by donor probes (Figure 5.8B). Panel C shows an outward view of the buried portion of the pocket highlighting structural concordance between a series of hydrophobic probes and the arc-like shape of the pocket defined by ILe10, Ala31, Val64 and hinge binding residues 80-83. The electrostatic interaction observed between Lys89 and the sulfonamide groups is modeled by a donor probe (Figure 5.8D).

In addition, the pocket model provided a highly concordant physical shape of the binding pocket, sharing not only key binding elements but the overall structural configuration. Figure 5.9 shows examples of these shared characteristics.

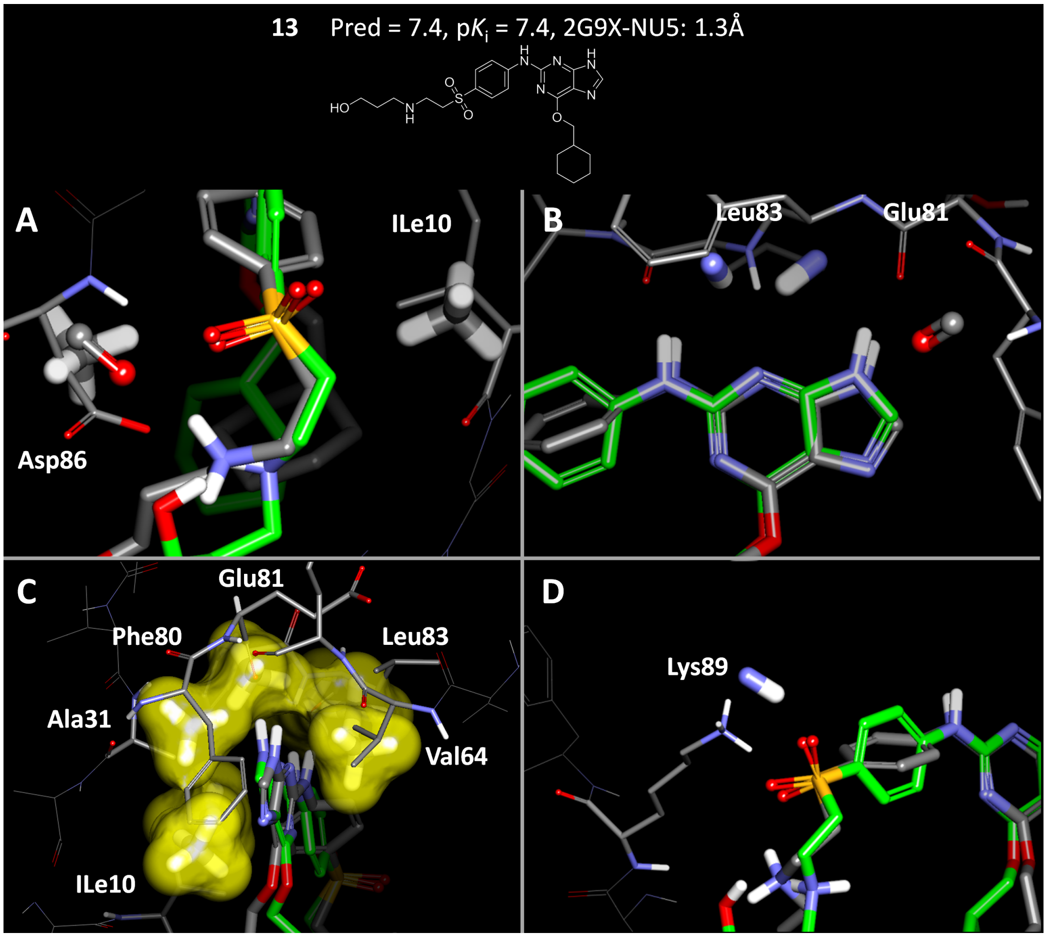


Figure 5.8: The induced pocket model matches key physical characteristics of the binding pocket. The predicted pose of compound **13** (grey) is shown with the bound pose (green) to provide a frame of reference. Panels A-D provide detailed snapshots of key regions of the binding pocket that are well represented by the pocket model. (A) Polar aspects provided Asp86 are captured by an acceptor probe and two hydrophobic probes provide matching physical constraints on the right and left-side of the pocket. (B) The backbone carbonyl of Glu81 is modeled by an acceptor probe and the NH group of Leu83 is captured by two donor probes. (C) Hydrophobic probes model the physical shape of the buried pocket region defined by ILe10, Ala31, Val64, and hinge residues 80-83. (D) Lys89 is represented by a donor probe at the opening of the pocket.

5.4.3 Effects of Structure-guided Modeling compared to Purely Structure-Based or Ligand-Based Approaches

Intuitively using more information gathered from a protein crystal structure may guide one towards developing a more concise understanding of the physical proper-

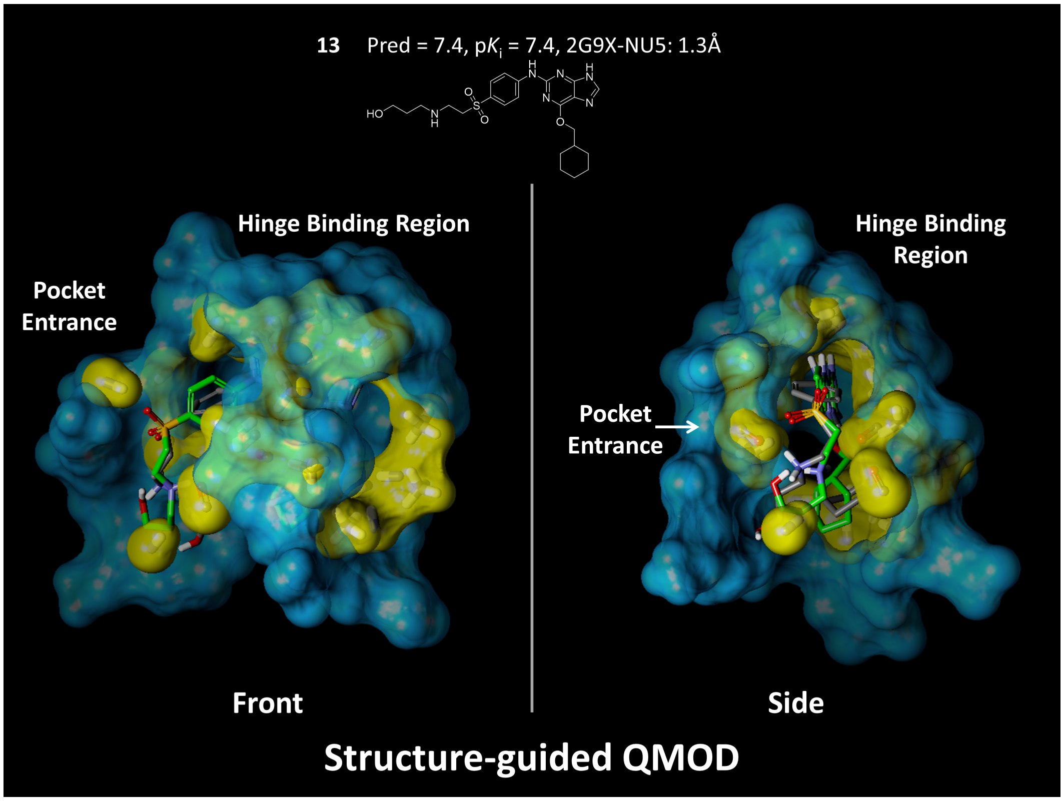


Figure 5.9: The Structure-guided QMOD procedure produces a pocket model that captures the overall shape and electrostatic elements of the CDK2 binding pocket. The 2G9X binding pocket is shown in blue skin with the final pocket model in thick sticks with yellow skin. The predicted pose of compound **13** (atom-colored sticks) is shown with the bound pose (green sticks) for a frame of reference. The front-view highlights the strong physical concordance the pocket model exhibits with the binding pocket, overall coverage of the perimeter of the binding cavity. The side-view shows a rotated clipped view of the pocket model and protein highlighting concordance in overall volume between the pocket model and binding pocket.

ties governing protein-ligand binding. To directly assess this effect we carried out controls that employed purely ligand-based and structure-based approaches individually. We compared the predictive performance of the structure-guided QMOD procedure against the standard ligand-based QMOD approach and results produced from Surflex-Dock. Although employing these methods individually may yield informative results within relevant applications, the structure-guided approach was observed to be more broadly applicable and accurate in predicting ligand affinity, binding modes, and identifying key protein residues responsible for ligand binding. In the ligand-based approach we carried out an analogous experiment to the structure-guided procedure using the same set of 30 CDK2 training ligands and the top 3 most active ligands (see Figure 5.1 compounds **1-3**) for the hypothesis alignment generation. The standard QMOD procedure[61] was carried out using only information derived from the training ligand structures and activities. Within the congeneric test series the standard ligand-based approach yielded excellent predictions with an average prediction error of 0.5 with a Kendall's Tau rank score of 0.73, equivalently powerful compared to the structure-guided QMOD model. The benefits provided by the structure-guided procedure became evident when testing the 56 structurally diverse CDK2 inhibitors. On the entire set of diverse ligands the structure-guided model yielded a Kendall's Tau rank score of 0.26 ($p < 0.01$) whereas the standard ligand-based model produced a 0.07 Kendall's Tau with an insignificant p-value of 0.25. The structure-guided model predicted 75% of the molecules with confidence greater than 0.5 with an average prediction error of 0.9, average RMSD of 1.57Å, and a Kendall's Tau rank score of 0.32 ($p < 0.01$). The standard model predicted 50% of the compounds with confidence greater than 0.5 with an average error of 2.17, average RMSD of 3.84Å, and a Kendall's Tau rank score of 0.08 ($p = 0.30$). A close examination of the prediction errors again revealed the dramatic benefit the structure-guided procedure provided. Figure 5.10 shows a cumulative distribution of the prediction errors produced by the structure-guided and standard QMOD procedures. The structure-guided pocket model produced significantly fewer prediction errors than the standard ligand-based model. The structure-guided model predicted 46% of the inhibitors with errors less than 0.75 log units, 63% with errors less than 1.0, and 77% less than 1.5. The standard ligand-based model predicted 11% of the compounds with errors less than 0.75, 18% with errors less than 1.0, and 30% less than 1.5.

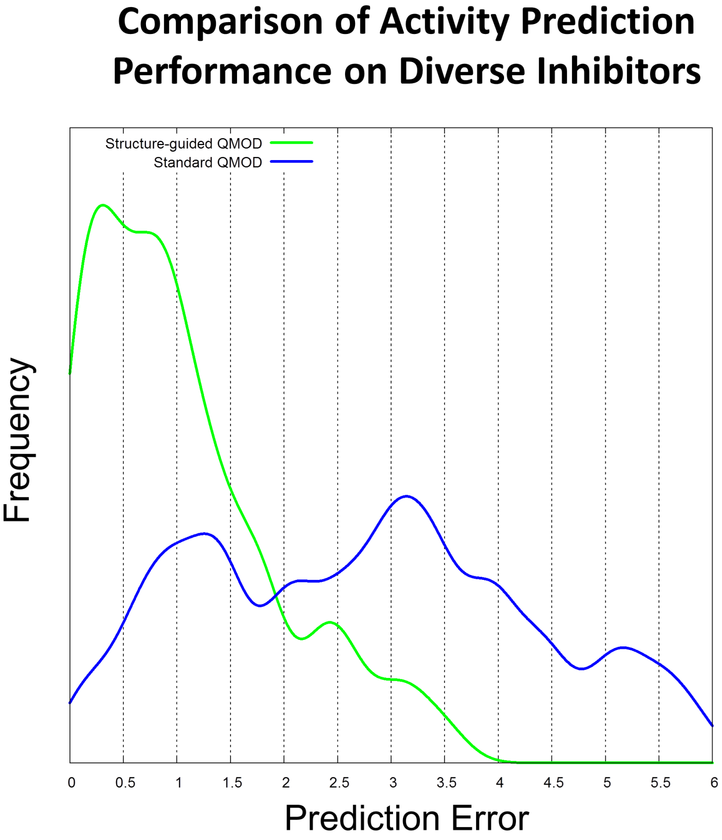


Figure 5.10: The structure-guided QMOD procedure (green) performed more accurately than a purely ligand-based approach on the *structurally diverse* CDK2 compounds. The structure-guided pocket model produced significantly fewer prediction errors than the standard ligand-based model. The structure-guided model predicted 46% inhibitors with errors less than 0.75 log units, 63% with errors less than 1.0, and 77% less than 1.5. The standard ligand-based procedure predicted 11% of the compounds with errors less than 0.75, 18% with errors less than 1.0, and 30% less than 1.5.

From a purely structure-based approach one may nominally resort to molecular docking in attempt to rank order a set of molecules and identify highly active ligands as leads for further optimization. In this light we employed molecular docking with the intention of producing accurate ligand rankings and identifying highly active compounds within our diverse set of 56 CDK2 inhibitors. We started with similar conditions used for the structure-guided QMOD approach which included the same

5 protein crystal structures with bound ligands (see Figure 5.2A). The intention here was to provide a direct comparison of the structure-guided QMOD procedure with a standard docking approach and compare their predictive performance. On the entire set of 56 inhibitors the structure-guided QMOD method produced a 0.26 (p < 0.01) Kendall's Tau rank score with an average activity prediction error of 1.0 log unit and an average RMSD of 1.8Å (see Figure 5.7 for examples). At confidence levels of 0.5 and higher the structure-guided model ranked 42 molecules with a 0.32 Kendall's Tau score with an average error of 0.9 and an average RMSD of 1.4Å. At confidence levels of 0.7 and higher the structure-guided model ranked 11 molecules with a 0.74 Kendall's Tau score with an average error of 0.7 and an average RMSD of 1.2Å. Overall the docking procedure performed notably well in ranking highly active ligands (i.e. $pK_i > 8.0$) over those with low activity (i.e. $pK_i < 6.0$). However, compared to the structure-guided approach, molecular docking was less predictive with molecules for which QMOD predicted with moderate to high confidence. Sets of molecules predicted with higher confidence had notably narrow ranges of binding activity. The entire set of 56 inhibitors provided a pK_i range of 3.5-9.9 spanning 6.4 log units. The set of ligands predicted with 0.5 confidence and higher provided a pK_i range of 4.5-9.9 spanning 5.4 log units, and the set predicted with 0.7 confidence and higher provided a pK_i range of 5.0-8.5 spanning 3.5 log units. On the entire set 56 inhibitors, Surflex-Dock produced a Kendall's Tau rank score of 0.26 ($p < 0.01$) yet with an average RMSD of 2.3Å. On the set of ligands QMOD predicted with at least 0.5 confidence, docking produced a Kendall's Tau of 0.29 ($p < 0.01$) also with an average RMSD of 2.3Å. On the set of ligands QMOD predicted with at least 0.7 confidence, docking produced a Kendall's Tau score of 0.38 with an insignificant p-value of 0.10 determined by permutation analysis, with minimal improvement of pose prediction averaging 2.1Å RMSD. Table 5.2 shows a detailed breakdown of the prediction performance on the 56 diverse CDK2 inhibitors.

Table 5.2: Test summary of 56 structurally diverse CDK2 inhibtors. The test set was evaluated in 3 categories: molecules for which QMOD predicted with confidence greater than 0.7, 0.5, and all molecules. Structure-guided QMOD provided reliably better predictive performance for predictions made with increased confidence. Surflex-Dock ranked the entire set comparably well, yet was nominally less applicable among the sets of molecules with narrow activity ranges (i.e. confidence categories 0.5 and 0.7).

Structure-guided QMOD

Confidence	NMols	pK_i range	Average Error[a]	Kendall's Tau ($p < 0.01$)	RMSD (Å)
0.7	11	5.0-8.5	0.7	0.74	1.2
0.5	42	4.5-9.9	0.9	0.32	1.6
all	56	3.5-9.9	1.0	0.26	1.8

Surflex-Dock

Confidence	NMols	pK_i range	Average Error[a]	Kendall's Tau ($p < 0.01$)	RMSD (Å)
0.7	11	5.0-8.5	2.3	0.38[b]	2.1
0.5	42	4.5-9.9	2.0	0.29	2.3
all	56	3.5-9.9	1.9	0.26	2.3

[a] Error values are units of pK_i. [b] $p = 0.10$

5.4.4 Structural guidance in the Presences of Limited Protein Structure Information

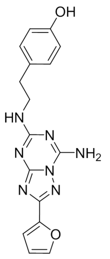
It is often the case that one will possess limited structural information for the protein target under investigation. With limited structures one inherits several challenges with respect to predicting ligand binding behaviors. To demonstrate applicability of the structure-guided modeling procedure in this more challenging scenario we applied the protocol to another therapeutically relevant target, the adenosine A_{2A} receptor. Structural determination of the adenosine A_{2A} receptor has historically proven to be challenging and limited. For structural guidance we used the first determined x-ray crystal structure of the adenosine A_{2A} receptor (PDB code: 3EML[80]). A set of 90 A_{2A} antagonists was gathered from studies carried out in recent design efforts aimed towards developing potent and selective pyrimdine-based compounds as human A_{2A} receptor antagonists.[81–85] The data spanned 5 optimization efforts, training and test ligands were organized temporally, and divided between 60 used for training (pK_i

range 6.0-9.7) and 30 used for testing (pK_i range 7.4-9.7). Figure 5.11 provides examples of the ligands used in this experiment. In addition, three compounds **17-19**[80,86] were included with the intention of providing structurally relevant guidance during the hypothesis alignment and model induction procedure. Molecule **17** was the first A_{2A} bound ligand observed via x-ray crystallography (PDB code: 3EML, 2008). Molecules **18** and **19** were previously identified and well established potent A_{2A} antagonists.

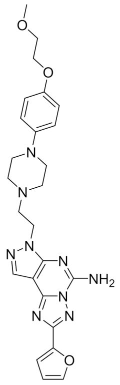
The structure-guided modeling procedure was carried out as described previously (see Figure 5.2) with a slight augmentation of the initial preparation of the protein and hypothesis alignment. We employed the Rosetta Backrub protocol[87] to computationally examine the receptor's structural flexibility. An ensemble of 100 alternative backbrub conformations were generated from which 5 structures were chosen based on their combined structural coverage of the pool of conformations. The hypothesis alignment was generated by employing a similarity-based alignment of compounds **19** and **20** to a joint target that included the crystal ligand **17** and the highest scoring docked pose of molecule **18**. The structure-guided modeling procedure continued as described previously, producing a pocket model that performed well with respect to activity prediction and model coverage of the binding site. Among the 30 test molecules, the structure-guided procedure yielded an average error of 0.85 log units with a Kendall's Tau rank score of 0.60 ($p < 0.01$). 7 out of 10 of the confidently predicted most active ligands were among the top 10 *bonafide* most active molecules. Table 5.3 shows the prediction performance on the entire test set with the top 10 *confidently* predicted most active ligands highlighted with a bold underline.

In addition to excellent numerical predictive performance, the induced model captured the detailed shape of pocket while providing a good representation for specific polar contacts of the A_{2A} binding pocket. Figure 5.12 highlights the relationships

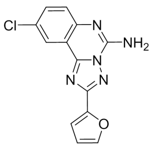
17 pK_i = 8.8, 3EML-ZMA



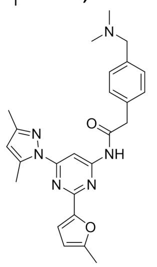
18 pK_i = 9.05, #240624



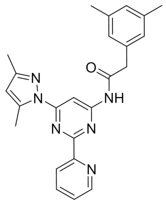
19 pK_i = 8.9, #16687



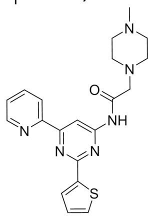
20 pK_i = 9.7, #255739



21 pK_i = 8.5, #429420



22 pK_i = 7.6, #270136



23 pK_i = 6.0, #407711

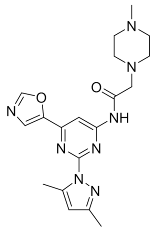


Figure 5.11: Training molecules used for induction of the adenosine A_{2A} pocket model. The A_{2A} data set consisted of a congeneric series of 90 ligands ligands developed during optimization efforts of a pyrimdine-based antagonist of the human adenosine A_{2A} receptor.[81–85] The ligands were organized temporarily and separated into 60 training molecules and 30 test ligands. Molecules **20-23** are examples of this series. Molecules **17-19**[80,86] were included with the intention of providing structurally relevant guidance during the hypothesis alignment and model building procedure. Molecule **17** was the first A_{2A} bound ligand observed via x-ray crystallography (PDB code: 3EML, 2008). Molecules **18** and **19** were previously identified and well established potent A_{2A} antagonists. Compounds **17-20** served as the bases for the structure-guided alignment hypothesis.

Table 5.3: Performance of the induced binding pocket model on the 30 blind A_{2A} ligands (Compound numbers as listed in ChEMBL_13[88]). 7 out of 10 of the *confidently* predicted most active molecules (boldface underlined) were among the top 10 *bonafide* most active ligands.

Rank	Mol.	Exptl.	Pred.	Error	Conf.	Rank	Mol.	Exptl.	Pred.	Error	Conf.
1	429125	9.7	7.8	1.9	0.82	16	**256125**	8.9	8.7	0.3	0.83
2	**256123**	9.5	8.4	1.1	0.80	17	256327	8.9	8.6	0.3	0.71
3	**404863**	9.5	8.3	1.2	0.89	18	255911	8.7	7.9	0.8	0.78
4	**253317**	9.5	8.1	1.4	0.82	19	**256546**	8.6	8.4	0.2	0.76
5	256124	9.5	7.8	1.8	0.83	20	255699	8.4	7.7	0.7	0.83
6	**401895**	9.4	8.8	0.6	0.77	21	255267	8.4	7.9	0.6	0.88
9	403846	9.3	8.7	0.6	0.72	22	404864	8.4	7.2	1.2	0.84
8	**256332**	9.3	8.5	0.8	0.83	23	402781	8.2	7.2	1.1	0.74
7	**256331**	9.3	8.4	0.9	0.86	24	255860	8.2	8.0	0.1	0.76
10	**403845**	9.2	8.5	0.6	0.79	25	255649	8.1	7.0	1.1	0.81
11	253528	9.2	7.8	1.4	0.81	26	404739	8.1	6.8	1.3	0.72
12	429850	9.0	7.9	1.2	0.65	27	255859	7.9	7.1	0.8	0.82
13	256547	9.0	8.1	0.9	0.83	28	255861	7.9	7.6	0.3	0.72
14	256548	9.0	7.8	1.2	0.82	29	255647	7.8	7.0	0.8	0.76
15	**256755**	8.9	8.9	0.1	0.86	30	404744	7.4	7.2	0.2	0.70

[a] Experimental, predicted, and error values are units of $\mathrm{p}K_i$.

between the pocket model and the A_{2A} binding pocket. Two acceptor probes were in excellent spatial agreement with the carboxylate side-chain of Glu169. The alignment of multiple A_{2A} receptor binding pockets revealed significant flexibility of Glu169, and highlighted the potential for interactions with the protonated amine commonly found within this series of compounds. Donor and acceptor probes captured polar interactions provided by Asn253 on the furan oxygen and pyrimidine nitrogen, also common among the ligands used in this study. Hydrophobic probes provided a well defined contour of the binding pocket that was in good spatial agreement with ILe64, Phe159, Met168, Trp384, Leu387, His388, and ILe412 (details not shown).

For the sake of thoroughness, a ligand-based control was employed using the standard QMOD procedure. A model was induced using the same 60 A_{2A} training ligands, using the top 2 most active compounds for the hypothesis alignment. The standard model performed nominally well, yielding an average error of 0.87 and a Kendall's Tau rank score of 0.42 ($p < 0.01$). The performance improvements of the structure-

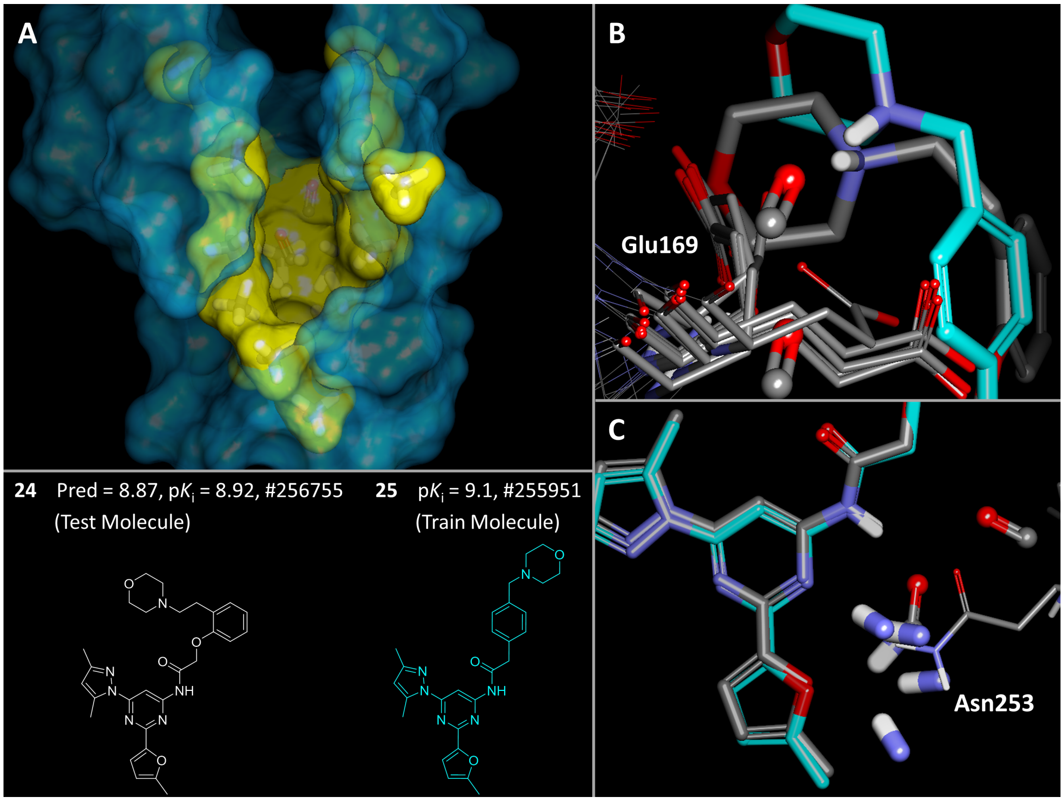


Figure 5.12: The structure-guided model showed a direct relationship with the shape and polar characteristics of the A_{2A} binding pocket. (A) A side clipped view of the the A_{2A} binding pocket (PDB code: 3EML, blue skin) shown in comparison to the induced model shown with sticks and yellow skin. There is high congruence between the shape of the pocket model and the crystal structure of the binding pocket (3EML). (B) Shown is test molecule **24** (atom-colored) in its final predicted pose displayed with training molecule **25** (cyan) from which the confidence measure was derived. The predicted activity of compound **24** was 8.9, revealing a 0.2 log unit deviation from its pK_i of 9.1. Shown in thin sticks are multiple crystal structures displaying the degree of flexibility exhibited by Glu160 (3EML, 2YDO, 2YDV, 3PWH, 3QAK, 3REY, 3UZA, 3VG9, 4EIY) and in thick sticks are the corresponding acceptor probes modeling their potential interactions with the protonated amines of compounds **24** and **25**. (C) Donor and acceptor probes are well matched with interactions provided by Asn253 (3EML, thin sticks).

guided model was not only apparent with respect to ligand ranking but the model showed a higher congruence with the protein binding pocket and more plausible predictions of bioactive poses. Figure 5.13 shows the different model configurations presented by the standard and structure-guided modeling procedures. The standard procedure presents a horizontally extended configuration of the training ligands and pocket model, an arrange that would suggest ligand-induced conformation arrangement of the pocket to accommodate such a binding mode. The structure-guided model presents an upright configuration of the training ligands with the pocket model suggesting a binding geometry that is more consistent with the upright position of the crystal ligand **17** and the optimally docked pose of molecule **18**. To our knowledge a co-crystal structure of the A_{2A} receptor with the pyrimidine inhibitors discussed in this study (i.e. molecules **20-23**) has not been determined. This scenario presents challenges with respect to pose prediction validation for this particular data set; however, this provides an interesting testable hypothesis addressing the general binding configuration of such pyrimdine-based ligands and how the level of protein flexibility plays a factor in such binding.

5.5 Conclusion

We believe that this study has approached the QSAR modeling question with a unique focus of data integration. We explored how different computational modeling strategies performed within narrow and broad chemical classes. There were three primary results. First, the structure-guided QMOD procedure produced models that were highly predictive within a congeneric series in two separate test cases. The structure-guided procedure performed comparatively well with the purely ligand-base approach with respect to affinity prediction and rank ordering of the CDK2 congeneric test series. In the more challenging A_{2A} case the structure-guided procedure performed noticeably better than standard ligand-based approach with respect to ranking ligands and sharing physical congruence with the binding site. These results highlighted the benefit of integrating structural information in a case where protein flexibility is likely

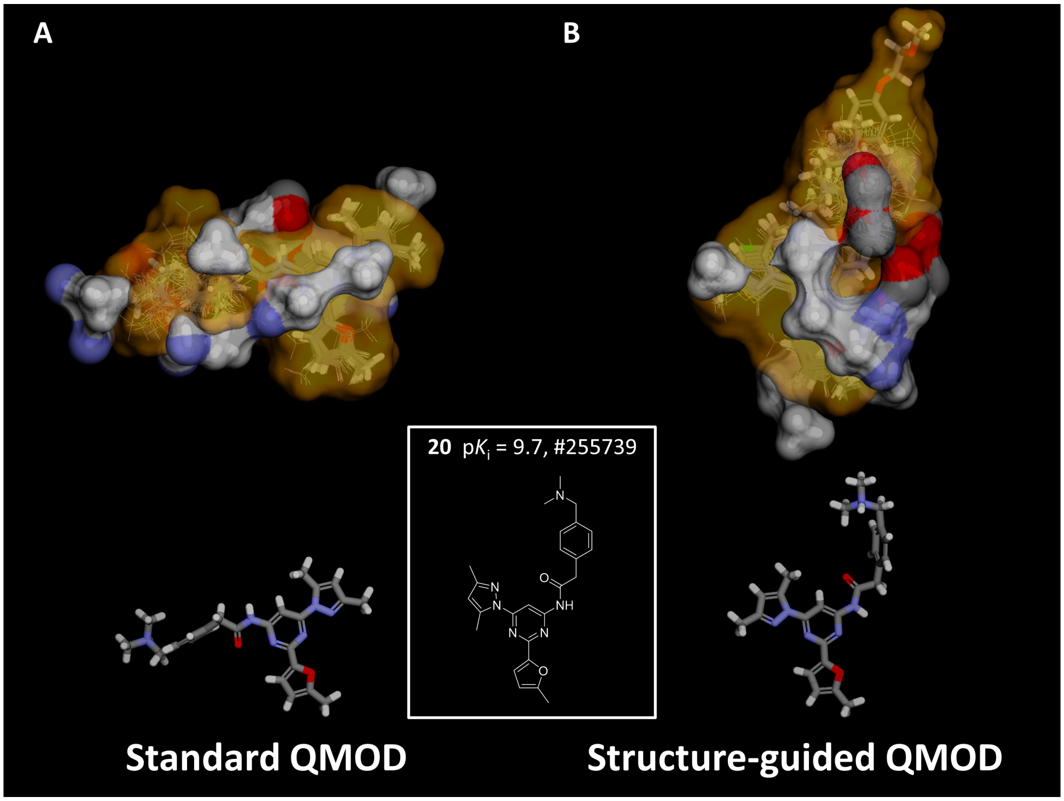


Figure 5.13: The standard and structure-guided QMOD procedures present different binding modes and model configurations of how the training ligands may fit inside the binding pocket. (A) The standard pocket model (atom-colored sticks with skin) and final optimal training poses (wires and sticks with gold skin) present a horizontally extended configuration of the potential binding interaction. Training molecule **20** (thick sticks) is shown below in its final predicted pose. (B) The structure-guided pocket model (atom-colored) and the final optimal training poses (wires and sticks with gold skin) display a vertically extended conformation.

an important contributing factor in accurate activity and pose prediction. Second, the structure-guided modeling procedure was more widely applicable and accurate in activity and pose predictions across a wide variety of *structurally diverse* molecules. On the structurally diverse CDK2 set the structure-guided QMOD procedure outperformed the standard ligand-based QMOD procedure with respect to rank correlation and activity prediction error. The structure-guided procedure performed equivalently well in ranking diverse molecules compared to molecular docking, but provided the additional benefit of better overall performance at higher confidence levels. Third, the structure-guided procedure produced models that shared high physical concordance with the protein targets in this study. In the CDK2 case the induced model showed a direct relationship with key binding site elements known for their role in ligand recognition. In the more challenging A_{2A} case, the induced model showed a direct correspondence to the shape and electrostatic characteristics of the binding pocket while providing a testable hypothesis of protein flexibility and interactions with specific ligand moieties. We demonstrated the applicability of the structure-guided QMOD procedure in two contrasting scenarios: one in which protein structural information was abundant as seen with CDK2, and one in which structural information was very limited as in the A_{2A} case.

Chapter 6

Conclusion and Future Directions

The field of computational structure-activity modeling in medicinal chemistry has a long history, with methods development going back at least 40 years.[65] Such methods, however, have frequently relied upon models that only bear a tangential relationship to the physical process of protein-ligand binding. The QMOD approach, by directly addressing the underlying physical phenomena of protein-ligand binding, is able to provide predictions for both binding affinity *and* binding mode, and it is able to do so with a broad domain of applicability. Models can be induced, and predictions profitable made, on structurally diverse scaffolds.

In the context of real-world application of QSAR methods, modeling is generally done in the context of an iterative process of lead optimization. The models themselves are refined using data that is produced, in part, based on predictions derived from model guidance. Reliance upon models that have a narrow applicability domain have a hidden cost when used iteratively. Sophisticated machine learning methods that rely upon correlation will tend to make "guesses" on structurally novel compounds toward the middle range of activity seen among training compounds (the middle generally being the most statistically likely given the priors). So, a compound that is both structurally novel and is actually among the most potent seen so far will generally be significantly underpredicted. Selections made using methods that have an implicit bias against novel compounds will guide a chemical trajectory that will make incremental modifications to known active molecules and will passively steer away from many profitable avenues of exploration.

The QMOD approach has no such bias, because it is a physical model. Molecules whose optimal pose produces a high score need not be statistically “similar” to any training compounds. Further, the QMOD approach allows for the explicit computation of structural novelty: the degree to which a new molecule physically explores space that has not been explored before. By making use of this novelty computation to explicitly select some molecules that will broaden the structural scope of the model, it is possible to dramatically increase the diversity of potent molecules that result from synthetic exploration. The use of QMOD in such a scenario was explored on a series of gyrase inhibitors, for which synthesis order was known.[61] The molecules used in the study were taken from a lead optimization program conducted at Vertex Pharmaceuticals. While there are many details to the study worth highlighting,[61] the impact of active selection of structurally novel compounds is particularly striking. Active pursuit of structurally novel compounds uncovered a series of potent compounds, sharing a core scaffold with others, but branching so as to occupy an unexplored pocket. These branched compounds also yielded novel biological effects in terms of their selectivity profiles when compared with the linear family.

In the broader aspect of computer-aided drug design, ligand structures and associated activities can be profitably exploited to make better use of experimentally determined protein structural information. In this work, we have also shown how to construct QMOD pocket models to represent protein binding sites in a manner that is constrained to make use of direct structural information. The clear extension to the method is to dispense with the pocketmol formalism and instead to refine the structures of an ensemble of aligned protein binding pockets. The goal would be to use the refined ensemble directly, with a simple docking-based scoring scheme, for affinity prediction. This requires a simple extension to the multiple-instance learning formalism, where in addition to the ligands having the potential for variation, the binding site itself would also be represented as variants. The score for a ligand given an ensemble of protein pocket variants would simply be the one resulting from the optimal fit to any of the variants. Such an approach fits in the gap between the approach described here and the purely physics-based simulation-oriented methods.

In any event, the results reported here encourage the development and use of hybrid methods that maximize information gleaned from different sources, including both biophysical information on protein structure and activity information from experimental determination of ligand activities. We believe that a shift in QSAR modeling that embeds more physical realism is both feasible and is a valuable direction for the field to take.

References

[1] F.S. Collins, M. Morgan, and A. Patrinos. The human genome project: Lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

[2] F.S. Collins, E.D. Green, A.E. Guttmacher, and M.S. Guyer. A vision for the future of genomics research. *Nature*, 422(6934):835–847, 2003.

[3] An integrated map of genetic variation from 1,092 human genomes. 491(7422):56–65.

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[5] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J.P. Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012.

[6] A. N. Jain, T. G. Dietterich, R. H. Lathrop, D. Chapman, Jr. Critchlow, R. E., B. E. Bauer, T. A. Webster, and T. Lozano-Perez. A shape-based machine learning tool for drug design. *J Comput Aided Mol Des*, 8(6):635–652, 1994.

[7] T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[8] A.N. Jain and A.E. Cleves. *Algorithmic Molecular Modeling: Applications for Pharmaceutical Research*. BioPharmics LLC, To be published.

[9] Joseph A. DiMasi and Henry G. Grabowski. The cost of biopharmaceutical R&D: is biotech different? *Managerial and Decision Economics*, 28(4-5):469–479, 2007.

[10] C. Lipinski and A. Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, 2004.

[11] T.E. Creighton. *Proteins: Structures and Molecular Properties*. W. H. Freeman, 1993.

[12] M.K. Gilson and H.X. Zhou. Calculation of protein-ligand binding affinities*. *Annu. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.

[13] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J Mol Biol*, 161(2):269–288, 1982.

[14] T.J.A. Ewing, S. Makino, A.G. Skillman, and I.D. Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, 15(5):411–428, 2001.

[15] M. Totrov and R. Abagyan. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins Structure Function and Genetics*, 29(s 1):215–220, 1997.

[16] S. P. Brown and S. W. Muchmore. Large-scale application of high-throughput molecular mechanics with poisson-boltzmann surface area for routine physics-based scoring of protein-ligand complexes. *J Med Chem*, 52(10):3159–65, 2009.

[17] A. N. Jain and A. Nicholls. Recommendations for evaluation of computational methods. *J Comput Aided Mol Des*, 22(3-4):133–9, 2008.

[18] I.J. Enyedy and W.J. Egan. Can we use docking and scoring for hit-to-lead optimization? *Journal of Computer-Aided Molecular Design*, 22(3):161–168, 2008.

[19] N. Huang, M.P. Jacobson, et al. Physics-based methods for studying protein-ligand interactions. *Current opinion in drug discovery & development*, 10(3):325, 2007.

[20] R.E Bellman. *Adaptive control processes: A guided tour.* Princeton University Press, 1961.

[21] Thomas G. Dietterich, Richard H. Lathrop, Tomas Lozano-Perez, and Arris Pharmaceutical. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

[22] S. Dill, K.A. & Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology.* Garland Science, 2003.

[23] A.J. Hopfinger. A qsar investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J Am Chem Soc*, 102(24):7196–7206, 1980.

[24] 3rd Rush, T. S., J. A. Grant, L. Mosyak, and A. Nicholls. A shape-based 3-d scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem*, 48(5):1489–95, 2005.

[25] B.B. Masek, A. Merchant, and J.B. Matthew. Molecular skins: A new concept for quantitative shape matching of a protein with its small molecule mimics. *Proteins: Structure, Function, and Bioinformatics*, 17(2):193–202, 1993.

[26] A. N. Jain. Morphological similarity: A 3d molecular similarity method correlated with protein-ligand recognition. *J Comput Aided Mol Des*, 14(2):199–213, 2000.

[27] A. N. Jain, N. L. Harris, and J. Y. Park. Quantitative binding site model generation: Compass applied to multiple chemotypes targeting the 5-HT1a receptor. *J Med Chem*, 38(8):1295–1308, 1995.

[28] A. N. Jain. Ligand-based structural hypotheses for virtual screening. *J Med Chem*, 47(4):947–61, 2004.

[29] G. Jones, P. Willett, and R. C. Glen. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J Comput Aided Mol Des*, 9(6):532–49, 1995. Jones, G Willett, P Glen, R C Research Support, Non-U.S. Gov't Netherlands Journal of computer-aided molecular design J Comput Aided Mol Des. 1995 Dec;9(6):532-49.

[30] A. N. Jain. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem*, 46(4):499–511, 2003.

[31] I. Muegge and Y.C. Martin. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal of Medicinal Chemistry*, 42(5):791–804, 1999.

[32] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions1. *Journal of Molecular Biology*, 295(2):337–356, 2000.

[33] T. Kortemme, A.V. Morozov, and D. Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *Journal of Molecular Biology*, 326(4):1239–1259, 2003.

[34] H.J. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, 8(3):243–256, 1994.

[35] A. N. Jain. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J Comput Aided Mol Des*, 10(5):427–440, 1996.

[36] W. Welch, J. Ruppert, and A. N. Jain. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol*, 3(6):449–62, 1996.

[37] T. A. Pham and A. N. Jain. Parameter estimation for scoring protein-ligand interactions using negative training data. *J Med Chem*, 49(20):5856–5868, 2006.

[38] T. A. Pham and A. N. Jain. Customizing scoring functions for docking. *J Comput Aided Mol Des*, 22(5):269–286, 2008.

[39] R. D. Cramer, D. E. Patterson, and J. D. Bunce. Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroid to Carrier Proteins. *J Am Chem Soc*, 110:5959–5967, 1988.

[40] R. D. Cramer. Topomer comfa: A design methodology for rapid lead optimization. *J Med Chem*, 46(3):374–88, 2003.

[41] R. D. Cramer and B. Wendt. Pushing the boundaries of 3d-QSAR. *J Comput Aided Mol Des*, 21(1-3):23–32, 2007.

[42] Y. C. Martin, M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico, and P. A. Pavlik. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J Comput Aided Mol Des*, 7(1):83–102, 1993.

[43] P. Willett. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *J Mol Recognit*, 8(5):290–303, 1995.

[44] O. F. Guner. History and evolution of the pharmacophore concept in computer-aided drug design. *Curr Top Med Chem*, 2(12):1321–32, 2002. Guner, Osman F Historical Article Review Netherlands Current topics in medicinal chemistry Curr Top Med Chem. 2002 Dec;2(12):1321-32.

[45] D. Zampieri, M. G. Mamolo, E. Laurini, C. Florio, C. Zanette, M. Fermeglia, P. Posocco, M. S. Paneni, S. Pricl, and L. Vio. Synthesis, biological evaluation, and three-dimensional in silico pharmacophore model for sigma(1) receptor ligands based on a series of substituted benzo[d]oxazol-2(3h)-one derivatives. *J Med Chem*, 52(17):5380–93, 2009.

[46] S. Ekins, K.V. Balakin, N. Savchuk, and Y. Ivanenkov. Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and kohonen and sammon mapping techniques. *J Med Chem*, 49(17):5059–5071, 2006.

[47] S. R. Johnson. The trouble with QSAR (or how i learned to stop worrying and embrace fallacy). *J Chem Inf Model*, 48(1):25–6, 2008.

[48] M. Minsky and S. Papert. *Perceptrons*. MIT press, 1969.

[49] F. Rosenblatt. A comparison of several perceptron models. *Self-Organizing Systems*, pages 463–484, 1962.

[50] A. N. Jain. *Chemical Analysis by Morphological Similarity (US Patent 6470305)*. 2002.

[51] A. N. Jain. Virtual screening in lead discovery and optimization. *Curr Opin Drug Discov Devel*, 7(4):396–403, 2004.

[52] Ann E. Cleves and Ajay N. Jain. Robust ligand-based modeling of the biological targets of known drugs. *Journal of Medicinal Chemistry*, 49(10):2921–2938, 2006.

[53] A. E. Cleves and A. N. Jain. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J Comput Aided Mol Des*, 22(3-4):147–59, 2008.

[54] J. Ruppert, W. Welch, and A. N. Jain. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci*, 6(3):524–33, 1997.

[55] A. N. Jain. Surflex-dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J Comput Aided Mol Des*, 21(5):281–306, 2007.

[56] A. N. Jain. Effects of protein conformation in docking: Improved pose prediction through protein pocket adaptation. *J Comput Aided Mol Des*, 23(6):355–74, 2009.

[57] A. N. Jain, K. Koile, and D. Chapman. Compass: Predicting biological activities from molecular surface properties. performance comparisons on a steroid benchmark. *J Med Chem*, 37(15):2315–2327, 1994.

[58] T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[59] J. J. Langham, A. E. Cleves, R. Spitzer, D. Kirshner, and A. N. Jain. Physical binding pocket induction for affinity prediction. *J Med Chem*, 52(19):6107–6125, 2009.

[60] A. N. Jain. QMOD: Physically meaningful QSAR. *J Comput Aided Mol Des*, 24(10):865–878, 2010.

[61] R. Varela, W. P. Walters, B. B. Goldman, and A. N. Jain. Iterative refinement of a binding pocket model: Active computational steering of lead optimization. *Journal of Medicinal Chemistry*, 55(20):8926–8942, 2012.

[62] P.S. Charifson, A.L. Grillot, T.H. Grossman, J.D. Parsons, M. Badia, S. Bellon, D.D. Deininger, J.E. Drumm, C.H. Gross, and A. LeTiran. Novel dual-targeting benzimidazole urea inhibitors of dna gyrase and topoisomerase iv possessing potent antibacterial activity: intelligent design and evolution through the judicious use of structure-guided design and stucture- activity relationships. *J Med Chem*, 51(17):5243–5263, 2008.

[63] T. A. Pham and A. N. Jain. Customizing scoring functions for docking. *J. Comput. Aided Mol. Des.*, 22(5):269–286, May 2008.

[64] A. N. Jain. Scoring functions for protein-ligand docking. *Curr Protein Pept Sci*, 7(5):407–20, 2006.

[65] C. Hansch and A. Leo. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Wiley New York, 1979.

[66] A.N. Jain and A.E. Cleves. Does your model weigh the same as a duck? *J Comput Aided Mol Des*, 26:57–67, 2012.

[67] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[68] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comp Sci*, 43(6):1947–1958, 2003.

[69] B. Chen, R.P. Sheridan, V. Hornak, and J.H. Voigt. Comparison of random forest and Pipeline Pilot Naive Bayes in prospective QSAR predictions. *J Chem Inf Model*, 52(3):792–803, 2012.

[70] D.B. Kell. Scientific discovery as a combinatorial optimisation problem: How best to navigate the landscape of possible experiments? *BioEssays*, 34(3):236–244, 2012.

[71] M.K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta, and C. Lemmen. Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, 43(2):667–673, 2003.

[72] Y. Fujiwara, Y. Yamashita, T. Osoda, M. Asogawa, C. Fukushima, M. Asao, H. Shimadzu, K. Nakao, and R. Shimizu. Virtual screening system for finding structurally diverse hits by active learning. *Journal of Chemical Information and Modeling*, 48(4):930–940, 2008.

[73] R. Spitzer, A. E. Cleves, and A. N. Jain. Surface-based protein binding pocket similarity. *Proteins*, 79(9):2746–63, 2011.

[74] S.P. Brown and S.W. Muchmore. Large-scale application of high-throughput molecular mechanics with Poisson-Boltzmann surface area for routine physics-based scoring of protein-ligand complexes. *J Med Chem*, 52(10):3159–3165, 2009.

[75] A.E. Gibson, C.E. Arris, J. Bentley, F.T. Boyle, N.J. Curtin, T.G. Davies, J.A. Endicott, B.T. Golding, S. Grant, R.J. Griffin, P. Jewsbury, L.N. Johnson, V. Mesguiche, D.R. Newell, M.E.M. Noble, J.A. Tucker, and H.J. Whitfield. Probing the atp ribose-binding domain of cyclin-dependent kinases 1 and 2 with o6-substituted guanine derivatives. *Journal of Medicinal Chemistry*, 45(16):3381–3393, 2002.

[76] I.R. Hardcastle, C.E. Arris, J. Bentley, F.T. Boyle, Y Chen, N.J. Curtin, J.A. Endicott, A.E. Gibson, B.T. Golding, R.J. Griffin, P. Jewsbury, J. Menyerol, V. Mesguiche, D.R. Newell, M.E.M. Noble, D.J. Pratt, L. Wang, and H.J. Whitfield. N2-substituted o6-cyclohexylmethylguanine derivatives: Potent inhibitors of cyclin-dependent kinases 1 and 2. *Journal of Medicinal Chemistry*, 47(15):3710–3722, 2004. PMID: 15239650.

[77] R.J. Griffin, A. Henderson, N.J. Curtin, A. Echalier, J.A. Endicott, I.R. Hardcastle, D.R. Newell, M.E.M. Noble, L.Z. Wang, and B.T. Golding. Searching for cyclin-dependent kinase inhibitors using a new variant of the cope elimination. *Journal of the American Chemical Society*, 128(18):6012–6013, 2006.

[78] PDB codes of the pool of protein structures considered for model guidance in the CDK2 study: 1AQ1, 1FVT, 1FVV, 1GIH, 1B38, 1B39, 1FIN, 1FQ1, 1GY3, 1HCK, 1JST, 1QMZ, 1H08, 1DI8, 1H01, 1H00, 1E9H, 1JVP, 1KE5, 1KE6, 1KE7, 1KE8, 1KE9, 1H07, 1E1X, 1JSV.

[79] M.L. Benson, R.D. Smith, N.A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin, and H.A. Carlson. Binding moad, a high-quality protein-ligand database. *Nucleic Acids Research*, 36(suppl 1):D674–D678, 2008.

[80] V.P. Jaakola, M.T. Griffith, M.A. Hanson, V. Cherezov, E.Y.T. Chien, J.R. Lane, A.P. IJzerman, and R.C. Stevens. The 2.6 angstrom crystal structure of a human a2a adenosine receptor bound to an antagonist. *Science*, 322(5905):1211–1217, 2008.

[81] D.H. Slee, X. Zhang, M. Moorjani, E. Lin, M.C. Lanier, Y. Chen, J.K. Rueter, S.M. Lechner, S. Markison, S. Malany, T. Joswig, M. Santos, R.S. Gross, J.P. Williams, J.C. Castro-Palomino, M.I. Crespo, M. Prat, S. Gual, J.L. Daz, J. Wen, Z. OBrien, and J. Saunders. Identification of novel, water-soluble, 2-amino-n-pyrimidin-4-yl acetamides as a2a receptor antagonists with in vivo efficacy. *Journal of Medicinal Chemistry*, 51(3):400–406, 2008. PMID: 18189346.

[82] D.H. Slee, Y. Chen, X. Zhang, M. Moorjani, M.C. Lanier, E. Lin, J.K. Rueter, J.P. Williams, S.M. Lechner, S. Markison, S. Malany, M. Santos, R.S. Gross, K. Jalali, Y. Sai, Z. Zuo, C. Yang, J.C. Castro-Palomino, M.I. Crespo, M. Prat, S. Gual, J.L. Daz, and J. Saunders. 2-amino-n-pyrimidin-4-ylacetamides as a2a receptor antagonists: 1. structure-activity relationships and optimization of heterocyclic substituents. *Journal of Medicinal Chemistry*, 51(6):1719–1729, 2008. PMID: 18307292.

[83] D.H. Slee, M. Moorjani, X. Zhang, E. Lin, M.C. Lanier, Y. Chen, J.K. Rueter, S.M. Lechner, S. Markison, S. Malany, T. Joswig, M. Santos, R.S. Gross, J.P. Williams, J.C. Castro-Palomino, M.I. Crespo, M. Prat, S. Gual, J.L. Daz, K. Jalali, Y. Sai, Z. Zuo, C. Yang, J. Wen, Z. OBrien, R. Petroski, and J. Saunders. 2-amino-n-pyrimidin-4-ylacetamides as a2a receptor antagonists: 2. reduction of herg activity, observed species selectivity, and structure-activity relationships. *Journal of Medicinal Chemistry*, 51(6):1730–1739, 2008. PMID: 18307293.

[84] M. Moorjani, X. Zhang, Y. Chen, E. Lin, J.K. Rueter, R.S. Gross, M.C. Lanier, J.E. Tellew, J.P. Williams, S.M. Lechner, S. Malany, M. Santos, P. Ekhlassi, J.C. Castro-Palomino, M.L. Crespo, M. Prat, S. Gual, J.L. Daz, J. Saunders, and D.H. Slee. 2,6-diaryl-4-phenacylaminopyrimidines as potent and selective adenosine a2a antagonists with reduced herg liability. *Bioorganic & Medicinal Chemistry Letters*, 18(4):1269 – 1273, 2008.

[85] X. Zhang, J.K. Rueter, Y. Chen, M. Moorjani, M.C. Lanier, E. Lin, R.S. Gross, J.E. Tellew, J.P. Williams, S.M. Lechner, S. Markison, T. Joswig, S. Malany, M. Santos, J.C. Castro-Palomino, M.L. Crespo, M. Prat, S. Gual, J.L. Daz, J. Saunders, and D.H. Slee. Synthesis of n-pyrimidinyl-2-phenoxyacetamides as adenosine a2a receptor antagonists. *Bioorganic & Medicinal Chemistry Letters*, 18(6):1778 – 1783, 2008.

[86] B.B. Fredholm, A.P. IJzerman, K.A. Jacobson, J. Linden, and C.E. Müller. International union of basic and clinical pharmacology. lxxxi. nomenclature and classification of adenosine receptors–an update. *Pharmacological Reviews*, 63(1):1–34, 2011.

[87] Colin A. Smith and Tanja Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology*, 380(4):742 – 756, 2008.

[88] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012.

Appendix A

QMOD Usage

Descriptions for all the command-line options discussed in this dissertation are shown here. Surflex Library v2.716, QMOD v1.500

Command	sf-qmod runsetup QMF	QMF has parameters and pointers to other files. The runsetup command produces scripts to drive the training procedure. The TrainMols argument contains pathnames to molecule files and activity data.
Input	*QMF TrainMols*	
Output	*RunSetup-qm*	
	RunTrain-qm[012]	
Command	source RunSetup	Hypothesis generation, initial ligand alignement, probe set generation
Input	*[Files auto-generated by runsetup]*	
Output	*qm-loghypo-hypo*.mol2*	
Command	source RunTrain-qm0	Pocketmol 0 build, optimal training poses, assessment of model parsimony. Note that the same procedure would be followed for RunTrain-qm1 and RunTrain-qm2, producing analogous files.
Input	*[Files auto-generated by runsetup]*	
Output	*final-qm0-best-pocketmol.mol2*	
	final-qm0-best-poses.mol2	
	final-parsim-qm0	

Table A.1: Generating an initial set of models from data with Surflex-QMOD.

Command Input Output	sf-qmod runtest QMF 0 round0-TestList source RunTest-qm0 *round0-TestList* *qm0-selectmols-report.mol2* *qm0-topresults.mol2*	Initial test of the round0 model, specifically selecting model 0 (of three).
Command Input Output	sf-qmod runrefine QMF 0 round0-NewTrain source RunRefine-qm0-ref01 *round0-NewTrain* *final-qm0-ref010-best-pocketmol.mol2* *final-qm0-ref010-best-poses.mol2* *QMODRunFile-qm0-ref01*	Refinement of the original round0 model proceeds in the same folder and requires new training data and *explicit* specification of the desired model number.
Command Input Output	sf-qmod runtest QMF-qm0-ref01 0 round1-TestList source RunTest-qm0-ref01 *round1-TestList* *qm0-ref010-selectmols-report.mol2* *qm0-ref010-topresults.mol2*	Testing of the round1 model is analogous to testing of the original model. Model numbers for *refined* models are always 0.
Command Input Output	sf-qmod runrefine QMF-qm0-ref01 0 round1-NewTrain source RunRefine-qm0-ref02 *round1-NewTrain* *final-qm0-ref020-best-pocketmol.mol2* *final-qm0-ref020-best-poses.mol2* *QMF-qm0-ref02*	Refinement of the round1 model requires the automatically generated QMOD refinement run file and requires new training data.
Command Input Output	sf-qmod runtest QMF-qm0-ref02 0 round2-TestList source RunTest-qm0-ref02 *round2-TestList* *qm0-ref020-selectmols-report.mol2* *qm0-ref020-topresults.mol2*	Testing of the round2 model is analogous. Iteration repeats the runrefine and runtest commands while incrementing round number.

Table A.2: Iterative predictive testing and refinement of a QMOD pocketmol.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

05/22/2013
Date