Molecular design of interfacial modifications to alter adsorption/desorption equilibria at fluid-adjoining interfaces

by

Nicholas J. Musolino

B.Eng. *summa cum laude*, The Cooper Union
for the Advancement of Science and Art (2006)
S.M., Massachusetts Institute of Technology (2009)

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Science in Chemical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2013
~~September 2012~~

Author .....
Department of Chemical Engineering
August 17, 2012

Certified by .....
Bernhardt L. Trout
Professor of Chemical Engineering
Thesis Supervisor

Accepted by .....
Patrick S. Doyle
Chairman, Department Committee on Graduate Study

Molecular design of interfacial modifications to alter adsorption/desorption equilibria at fluid-adjoining interfaces

by

Nicholas J. Musolino

Submitted to the Department of Chemical Engineering
on August 17, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Science in Chemical Engineering

Abstract

The thermodynamics and mass transfer kinetics of adsorption and desporption at interfaces play vital roles in chemical analysis, separation processes, and many natural phenomena. In this work, computer simulations were used to design interfacial modifications to alter the physical processes of adsorption and desorption, using two different approaches to molecular design.

In the first application, the finite-temperature string method was used to elucidate the mechanism of water's evaporation at its liquid/vapor interface, with the goal of designing a soluble additive that could impede evaporation there. These simulations used the SPC/E water model, and identified a minimum free energy path for this process in terms of 10 descriptive order parameters. The measured free energy change was 7.4 kcal/mol at 298 K, in reasonable agreement with the experimental value of 6.3 kcal/mol, and the mean first-passage time was 1375 ns for a single molecule, corresponding to an evaporation coefficient of 0.25. In the observed minimum free energy process, the water molecule diffuses to the surface, and tends to rotate so that its dipole and one O–H bond are oriented outward as it crosses the Gibbs dividing surface. As the molecule moves further outwards through the interfacial region, a local solvation shell tends to protrudes from the interface. The water molecule loses donor and acceptor hydrogen bonds, and then, with its dipole nearly normal to the interface, stops donating its remaining donor hydrogen bond. After the final, accepted hydrogen bond is broken, the water molecule is free. An analysis of reactive trajectories showed that the relative orientation of nearby water molecules, and the number of accepted hydrogen bonds, were important variables in a kinetic description of the process.

In the second application, we developed an *in silico* screening process to design organic ligands which, when chemically bound to a solid surface, would constitute an effective adsorption for a pharmaceutically relevant mixture of reaction products. This procedure employs automated molecular dynamics simulations to evaluate potential ligands, by measuring the difference in adsorption energy of two solutes which differed by one functional group. Then, a genetic algorithm was used to iteratively improve a population of ligands through selection and reproduction steps. This pro-

cedure identified chemical designs of the surface-bound ligands that were outside the set considered using chemical intuition. The ligand designs achieved selectivity by exploiting phenyl-phenyl stacking which was sterically hindered in the case of one solution component. The ligand designs had selectivity energies of 0.8 to 1.6 kcal/mol in single-ligand, solvent-free simulations, if entropic contributions to the relative selectivity are neglected. This molecular evolution technique presents a useful method for the directed exploration of chemical space or for molecular design.

Thesis Supervisor: Bernhardt L. Trout
Title: Professor of Chemical Engineering

Acknowledgments

I would like to kindly acknowledge the many people who helped make this work possible.

First, I would like to thank my advisor at MIT, Prof. Bernhardt Trout. Prof. Trout has provided invaluable guidance throughout the course of this project. His advice as to the scope and context of this work, and his suggestions for carrying it out at a technical level and for solving technical problems, were always very helpful.

I would also like to thank my thesis committee members, Profs. Alan Hatton and Daniel Blankschtein. Each provided valuable guidance and feedback throughout my tenure here. In particular, Prof. Hatton helped conceptualize the selective adsorption project described in this thesis, and Prof. Blankschtein provided important suggestions about how to compared measured evaporation rates to their actual values, and very valuable encouragement at times when it was very much appreciated..

I would also like to thank Prof. Hatton for encouraging me to participate the School of Chemical Engineering Practice at MIT, and for his efforts as Director there. I learned a great deal, at a technical level and about professional practice, from Profs. Claude Lupis and Robert Fisher while working as a student-consultant at two Practice School field stations.

The School of Engineering at the Cooper Union provides "an education equal to the best," in the words of the institution's founder. I would especially like to thank several faculty members who provided guidance and valuable instruction to me over my four years there. In the Department of Chemical Engineering, Profs. Richard Stock and Irv Brazinsky were excellent teachers in the classroom, and kindly offered professional advice when approached. In the Department of Mathematics, Profs. Om Agrawal, Martin Pincus, and Alexander Kheyfits taught courses in analytic geometry, calculus, real and complex analysis, vector spaces, and boundary value problems. Each of these topics was very helpful to me during my graduate career, either in the classroom or in carrying out the work described in this thesis.

I am also thankful to have worked with, and learned from, a number of excellent

scientists within the Molecular Engineering Lab at MIT. Dr. Victor Ovchinnikov; Dr. Erik Santiso; Dr. Geoffrey Wood; and Dr. Diwakar Shukla. Each has kindly shared his thoughts and advice, on matters from the smallest "implementation details" to broad, thematic questions in science.

In particular, Erik Santiso provided detailed advice and technical suggestions for the simulation-based molecular design project described in Chapters 5 and 6, and provided valuable technical help in adapting simulation software and in understanding directional statistics under symmetry for the water evaporation study described in Chapters 3 and 4.

I would like to kindly acknowledge the use of Erik's software libraries for integration, interpolation, and file parsing in the latter project. Molecular graphics in this thesis were produced with VMD.[1]

Most of all, I must thank the members of my family, who collectively were a wellspring of support and encouragement throughout my education. My parents, two younger sisters, and grandparents all were patient listeners, sensible counselors, dependable allies—and even helpful statistical consultants when called upon! Most of all, my parents Larry and Jamie have always encouraged and nutured a love for learning in me, without ever applying undue pressure or pushing me towards a predetermined path, and they will always have my gratitude.

This doctoral thesis has been examined by a Committee of the Department of Chemical Engineering as follows:

Professor Daniel Blankschtein .....
Member, Thesis Committee
Herman P. Meissner '29 Professor of Chemical Engineering

Professor Bernhardt L. Trout .....
Thesis Supervisor
Member, Thesis Committee
Professor of Chemical Engineering

Professor T. Alan Hatton .....
Member, Thesis Committee
Ralph Landau Professor of Chemical Engineering

Contents

List of Figures

List of Tables

Nomenclature

The meaning of symbols and abbreviations used in this thesis are listed here. For physical quantities, typical units are listed in parentheses.

Abbreviations

MFEP	Minimum free energy path through a space defined by order parameters
MFPT	Mean first-passage time, the average time required for a system to reach a final milestone from the given milestone
NVE	A statistical-mechanical ensemble (set of system states and associated probabilities) in which particle number, system volume, and energy are constant
NVT	A statistical-mechanical ensemble which particle number, system volume, and temperature are constant
PMF	Potential of mean force, the non-physical measure of the likelihood of a state in which order parameters are at particular values (kcal/mol)
SMCV	String method in collective variables, a method to identify likely reaction paths in terms of order parameters
VM	Voronoi milestoning, a method to measure free energy changes and kinetic parameters along a reaction path.

Roman letters

C	Condensation coefficient (dimensionless)

$E(x)$	System energy as a function of system coordinates (kcal/mol)
G	Molar flux at interface (mol/cm^2-s)
k_B	Boltzmann's constant (kcal/mol-K)
N	Number of molecules in a simulated system
N_{img}	Number of replicas or "images" simulated along a reaction path.
N_{OP}	Number of order parameters (also called collective variables) used to describe a system
P	Total pressure (bar)
p_B	Reaction committor probability, *i.e.* the probability a reactive system will reach the product basin (labeled B) if assigned Boltzmann velocities
$q_j \equiv q_j(\mathbf{x})$	Order parameter j used to describe state of an aqueous interfacial system
q_j^*	A particular value of order parameter j
Q_l	Partition function of a molecule in state l
q_l^{mode}	Component of partition function corresponding to *mode* (vibrational, rotational, *etc.*) of molecule in state l
r_Q^j	Position of atom Q in molecule labeled j
T	Absolute temperature (K)
$\bar{v}$	Average velocity of gas molecules (m/s)
$w_{hb}(r)$, $w_{den}(r)$	H-bond and density weighting functions, as a function of oxygen–oxygen distance r
$\bar{x}$, $\bar{y}$	Weighted average of vector components, used to calculate an average direction

$\hat{z}$	Unit vector in z-direction, which is direction of interfacial normal
Z	partition function of the canonical ensemble

Greek letters

α	Alternative notation for condensation/evaporation coefficient (dimensionless)
α	Fractional distance along a reaction path
β	Inverse temperature, equal to $(k_B T)^{-1}$ ([kcal/mol]$^{-1}$)
$\delta(\cdot)$	Dirac delta function
$\Delta G^{\ddagger}$	Gibbs free energy of activation for a reaction (kcal/mol)
$\eta_{a,b}$	Angle between dipole vectors of water molecules labeled a and b
γ_E	Evaporation coefficient (dimensionless)
κ	Transmission coefficient (dimensionless)
μ^a	Dipole vector of a water molecule labeled a
ν	Unit vector perpendicular to plane formed by a water molecule's three atoms
τ	Mean first-passage time, *i.e.* the expected time required to reach a final milestone from a particular milestone
θ	Angle between a molecule's dipole vector and interfacial normal
τ'	Mean first-passage time, scaled to increase from 0 (initial state) to 1 (at the final milestone).
ω	Angle between normal ν to water molecule's plane and interfacial normal

Chapter 1

Introduction

The objective of this project is to design, at a molecular level, modifications to an interfacial system that would alter the adsorption/desorption thermodynamics of species in an adjoining fluid. The two technical problems to which this approach is being applied are: the design of a surface-modified adsorption medium for the selective adsorption of impurities from a solution in upstream pharmaceutical manufacturing, and the design of a surface-active additive to aqueous solutions to introduce energetic barriers to evaporation and to retard the evaporation rate at the liquid/vapor interface. Figure 1-1

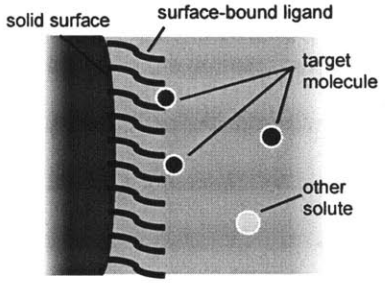


Figure 1-1. Schematic showing selective adsorption of a solute from solution.

This project also provides examples of two paradigms of molecular design using computer simulation. In the first paradigm, practioners employ computer simulations to understand the mecahnisms of physical or chemical processes at a molecular level of detail, a level of detail which may be difficult or impossible to obtain through experimental measurements. Based on this detailed mechanistic understanding, it is

possible to design molecules which, as additives, would alter the physical process in questions.

A second paradigm is to employ computer simulations as fast and cheap screening tools for a group of many possible molecular designs. This approach led to the promise of *in silico* screening of drug libraries for specific interactions, in numbers that would not be possible using physical/experimental screening. Our approach to screening is novel, however, in that we are carrying out directed evolution on molecular architectures. More specifically, our overall method of approach in the second paradigm is to first begin with a set of linear ligands that would be attached to a solid surface (such as gold or silica), to next evaluate their suitability using molecular simulations, for selectively adsorbing a particular pharmaceutical intermediate, designated E2, assign a fitness score to each, and finally to evolve population of such ligands using the fitness scores and genetic information, and repeating this process many times. We would like to compare our computational results to experimental measurements to verify that the former are meaningful.

1.1 Review of experimental and simulation-based studies of evaporation

The evaporation of water at its interface with air has been studied because it plays an important role in atmospheric processes, as well as in technological and analytical applications. In the field of microfluidic technology, for example, excess evaporation leads to crystallization of ink components in inkjet print heads, resulting in efforts to develop additives to address the so-called "inkjet decap problem."[2] Controlling the rate of evaporation from aqueous interfaces is also advantageous in drying operations, to diminish the risk of surface cracking.[3] Likewise, when evaporation is used to create supersaturated solutions in protein crystallization, diminished evaporation rates can favor nucleation of crystallites over the precipitation of aggregates.[4–6]

Despite the importance and ubiquity of the aqueous evaporation process, little is

known about its molecular-level mechanism(s). In fact, there is currently no consensus as to the actual *rate* of evaporation of water into dry air or vacuum. Rates of evaporation are often expressed in terms of the evaporation coefficient γ_E, which is equal to the mass accommodation coefficient α.

This study is an effort to understand, at a molecular level, the process of evaporation, which can be thought of as the inverse of accommodation of water itself. Our motivation for understanding evaporation in this way is to aid in designing soluble solution additives to diminish the rate of evaporation into air.

The evaporation coefficient is the ratio of the actual evaporation rate to the theoretical maximum rate calculated from the Hertz-Knudsen[7,8] equation:

$$G_{max} = \frac{1}{4}\frac{P_v}{k_B T}\left(\frac{8k_B T}{\pi M}\right)^{1/2} = \frac{P_v}{(2\pi M k_B T)^{1/2}}$$

In this equation, G_{max} is molar flux; M the molecular weight; T the temperature of the surface; and P_v the corresponding vapor pressure.

This maximum rate is derived by considering dynamic equilibrium under liquid-vapor coexistence, and neglecting any vapor- or liquid-phase resistance, as will be discussed in more detail in the next chapter. In summary, the evaporation rate is set equal to the rate at which gas molecules condense, assuming that every molecule that strikes the liquid surface enters the liquid. At a surface temperature of 298 K, the vapor pressure is 0.00317 MPa, and this value is, on a mass basis, 0.108 g/(cm^2 · s).[1]

Early measurements of the evaporation rate, performed in the decade 1925–35, obtained values of about 0.4[10] and 0.04 for this coefficient.[11,12] Since that time, practitioners have observed values between about 0.001 and 1.0,[13,14] with most measurements falling between 0.04 and 1.0. The challenging in measuring G and $\gamma_E = G/G_{max}$ is that this rate should be measured under conditions where heat and mass transfer are negligible, and this requires minimizing the resistances in these phases, and compensating for non-zero resistance in either bulk phase. Evaporation rate measure-

[1]As a monograph by Frank E. Jones points out, assuming a constant surface temperature, the water in Lake Mead in Nevada would evaporate in less than a day at this theoretical rate;[9] obviously, heat and mass transfer play a limiting role in macroscopic systems.

ments have been carried out by monitoring an evaporating droplet's size,[11,12] isotope exchange in droplet flow train reactors[15] or in jetted streams,[16] Raman thermometry,[17] and monitoring droplet expansion by Mie scattering.[18] Typically, studies with dynamically renewing surfaces result in values of γ_E close to unity, while those with quasi-static surfaces display values of about 0.1 or less.[13,14,19] Summarizing the state of affairs in their 2011 update[14] to their 2006 review of mass transfer at interfaces,[19] Davidovits and coauthors write (using the notation α for the evaporation coefficient), "The question still remains, why do some studies yield α_{H2O} significantly smaller than 1 while others point to a value of $\alpha = 1$"?

Because of the difficult nature of these experiments, elucidating the molecular-level details of the evaporation process is, in the main, a future goal of experimental work. To date, experimental studies have proposed that evaporation (and condensation) is mediated by the formation of "small clusters or aggregates" of non-bulk liquid water at the interface,[15] and (in separate work) that molecules in such a cluster become "weakly-bound surface species," then finally evaporate to become gas molecules.[17] This stepwise process leads to a free energetic barrier to evaporation, and in light of previous experimental evidence that water leaves the interface with Boltzmann-distributed kinetic energy,[20] the authors identified the barrier as possibly entropic in nature, "due to possible geometric requirements for the evaporation of a water molecule."[17]

In terms of intermolecular interactions, citing the dependence of the empirical evaporation coefficient on isotopic composition, Cappa and coworkers highlighted the "importance of the first solvation shell in controlling evaporation."[16] "Specifically," they wrote, "the nature of acceptor and donor hydrogen bonds, and their influence on librational and hindered translational motions, will determine evaporation rates." The evaporating molecule's accepted hydrogen bonds, they write, would exhibit a strong dependence on isotopic composition, while donated hydrogen bonds would be only indirectly sensitive to composition,[16] which could suggest that accepted hydrogen bonds are more important than donated ones in "determin[ing] evaporation rates."

Molecular simulations are often applied to experimentally challenging physical

problems, because they provide molecular-scale spatial and temporal resolution, and allow precise control over physical conditions like temperature and pressure. Indeed, over the past two and a half decades, computer simulations have been used to study water's interfaces with air, where evaporation takes place, because they allow researchers to observe, at a molecular level of detail, the behavior of this ubiquitous substance outside its well-studied bulk state.

Early molecular dynamics (MD) studies focused on the structural properties of the interface, such as the length scale of density variation, the distribution of surface molecules' orientation, and surface tension.[21–26] Later, first-principles MD simulations examined surface molecules' polarization at interface, in addition to structural properties.[27–30]

Simulations have also be used to obtain a picture of the dual processes of evaporation and condensation or mass accommodation; the latter term denotes the transfer of a water or solute molecule from the gas phase into the solution phase. Mass accommodation of atmospherically relevant solutes, including water vapor itself, has been examined using molecular dynamics simulations.[31–36] These studies examined the potential of mean force (PMF) of such a system as a function of the height of the solute above or below the interface; when a single variable is restrained in this way, the PMF is equal to the free energy. These studies found that (1) as the water molecule in the vapor approaches the interface, the sytem loses free energy, with only a minority of the total FE change occurring inside the Gibbs dividing surface (the plane where time-averaged density is equal to $\frac{1}{2}(\rho_{\text{liq}} + \rho_{\text{vapor}})$); and (2) no activated state was observed between the liquid and vapor states, and no significant minimum in the free energy profile was present on the liquid side of the interface, as for other solutes.

Employing a single coordinate, such as the distance above or below the interface, however, does not provide physical insight into the molecular-level picture of evaporation, as restraining this variable averages over all other physical quantities. In particular, such restraints do not identify the physical conformations (relative to the interface and its neighbors) the water molecule typically passes through during the

evaporation process, or what forces (or entropic considerations) are most influential during this process.

Another method of studying evaporation and accommodation has been to directly observe such events in long MD simulations. Because evaporation is a somewhat rare event on the timescales accessible to molecular simulations, obtaining a representative ensemble of such trajectories can be challenging. For example, at the maximum rate discussed above, an evaporation event would take place, on average, once every 2.8 ns from an interface with area 1000 $Å^2$, which is typical of the systems in MD simulations studies. Gathering a representative ensemble of "reactive" trajectories would therefore require long simulation times and, with frequently-saved coordinate data, concomitantly large trajectory data files.[37]

Accordingly, several MD studies have focused on mass accommodation instead of evaporation, by repeatedly placing water molecules in the vapor region above a liquid slab, and "firing" them at the water surface with Boltzmann-distributed linear and angular momenta. In most cases, very few water molecules are scattered or deflected, so that most remain on the surface or enter the bulk within the time interval of observation (typically 10 to 20 ps), leading to values of the accommodation coefficient near unity.[34,38–40] As practitioners have pointed out, however, the appropriate length of time to monitor the simulated systems for accommodation or desorption back into the vapor phase is not known *a priori*. Other simulation studies monitored the evaporation flux and obtained values of 0.99[41] for TIP3P water at 300 K and 0.3[42] for TIP4P water. (A study by Matsumoto, with few methodological details, reported a value of 0.3.[43])

More recently, Caleman and van der Spoel focused on the structural and energetic results of evaporation,[44,45] and found that evaporated molecules had a surfeit of kinetic energy, compared to the entire system's temperature. Mason observed 74 evaporation events from a 4890-molecule spherical droplet in a non-periodic simulation of TIP3P water.[37] These events occurred after unusually close oxygen–oxygen or hydrogen–hydrogen contacts, suggesting a transfer of van der Waals or electrostatic potential energy into the kinetic energy needed to overcome the energetic barrier to

evaporation. In most cases, the molecules in question had a coordination number of 1 or 2 at the start of the evaporation process.

1.2 Review of work in molecular design for adsorption or binding

1.2.1 Molecular design using genetic algorithm-based approaches

As discussed above, there are (at least) two approaches to employing molecular simulations to carry out molecular design/engineering: The first approach is to carry out computer simulations to gain a detailed, mechanistic understanding of the physical phenomenon of interest, and then to exploit that understanding to design/modify molecules (such as solution additives, adsorption media, or catalysts), and to then use additional simulations and experiments to evaluate the novel designs.

A second approach involves computational high-throughput screening, that is, evaluating many molecules in libraries or databases for desired properties. One example of such work is the BioDrugScreen project,[46] in which about 1600 small molecules were tested for interactions with about 1900 sites in human proteins, and in which the authors cite the use of hundreds of thousands of CPU-hours on a supercomputer to screen the resulting 3 million combinations.

Yet even with ever-increasing hardware capabilities and continuing improvements to simulation algoritms, the "chemical space"—the set of all small molecules which are energetically stable[47]—presents a vast domain. The space of all small, organic molecules has been estimated to contain up to 10^{60} members,[48] while the number of such entries in the CAS Registry reached 60 million in May 2011.

If each application of molecular design can be thought of as a screening (or optimization) problem (*e.g.* to find one or more small molecule ligands to bind strongly to a protein site), then screening/optimization by enumerative search in the chemical space is not a practical possibility. Screening thousands of molecules in a database has the advantage of working with a subset of molecules that may be well-curated:

for example, database members might be known to be "drug-like"[49–51] or synthesizable.[52] But such databases also present a fixed subset of candidate solutions, and to this point have been focused on potential pharmaceutical leads. In addition, screening all members of a database is inefficient, if many unsuitable molecules with similar structures or similar properties are separately evaluated; in this sense, the information gained from identifying high- and low-performing molecules early in a search is not exploited. Trained chemists can be asked to screen large data sets, but a recent study found that their classifications of compounds as promising or not promising can typically be explained by one to two molecular parameters, on a statistically significant basis.[53]

This work is an attempt to overcome these disadvantages, by applying a genetic algorithm (GA) to the broad screening and rough optimization of molecular structures. Genetic algorithms[57] and other optimization approaches[58,59] have previously been used for molecular design; the use of GAs in the context of drug design was reviewed by Gillet[60] and more briefly by Terfloth and Gasteiger.[61] Examples of such work are summarized in Table 1-1.

The approach in this thesis differs from previous work, in that I have employed as an objective function measurements from molecular simulations, rather than similarity to a given molecule or heuristic scores from docking programs. Molecular dynamics (MD) simulations, used in this study, can be used to measure an array of thermodynamic and transport properties. Other evaluation techniques could include properties calculated using DFT or *ab initio* methods,[75] or simpler quantitative structure-activity (or property) relationships (QSPR/QSARs).[76,77] Within the Harvard Clean Energy Project, for example, DFT calculations are used as a screening technique in the evaluation of novel organic photovoltaics.[78,79] Because of a GA optimizer's propensity to broadly explore its underlying state space (in our case, the space of reasonably-constructed organic molecules), our approach would provide a natural means to generate new, yet-unsynthesized compounds.

Such simulations are enabled by automated topological perception and force field parameter assignment methods. Such techniques have been developed[80–82] for molec-

ular mechanics force fields, and have been used with some success to evaluate the binding affinities of small molecules to other small molecules[83] or to proteins.[84–86] (In the cited examples, the GAFF force field[80,81] was applied to a small set of molecules identified *a priori* by researchers.)

1.2.2 Application to design of surfaces for selective adsorption

Our application is the design of a specialized surface, comprising a layer of organic, small-molecule ligands chemically bound to a solid substrate such as gold or silicon.[87] Its purpose is to selectively remove unconverted reactant from a solution also containing a reaction product, which should remain in the solution for further processesing.

Such a material could used to separate undesired solution components (while leaving the desired intermediate in solution) in a continuous fashion using a simulated moving bed (SMB) unit.[88–90] Adsorption-based SMB units simulate the movement of a solid or gel phase countercurrent to the process stream by varying the liquid injection and withdrawal locations along a column.

SMBs have been used in pharmaceutical manufacturing mainly for enatioselective separations,[91–98] as has been reviewed elsewhere,[99–103] although non-enantiomertic separations are also possible.[88,104,105] Preliminary economic evaluations have shown that SMB-based separations[106] and continuous manufacturing more generally[107] have the ability to reduce overall process costs in pharmaceutical manufacturing. For example, one study found that an SMB achieved productivity (mass product purified per mass packing material per unit time) one third higher and solvent use 45% lower than the corresponding batch operation;[108] in economic terms, a 2002 study of optimized batch and SMB chromatography operations, with the same separation medium, showed that the SMB unit reduced separation costs by 13%,[106] with greater cost savings at higher production rates.

The particular separation task in this study arises from the synthesis of a particular active pharmaceutical ingredient, and is the adsorption of 3-[1-(hydroxyl)ethyl]phenol

(called "E2") from an ethyl acetate solution, while 3-[1-(methylamino)ethyl]phenol ("E6") is to remain in solution. The structures of these species are given in Figure 1-2. The surface we aim to design thus must simultaneously satisfy two desgin criteria: to adsorb E2 as strongly as possible, while adsorbing E6 as minimally as possible.

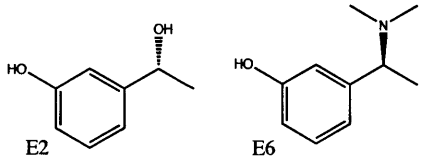


Figure 1-2. Structure of pharmaceutical intermediates designated E2 and E6.

The selection of adsorption media for applications like this is typically guided by heuristic rules, based on "physical property difference in the molecules to be separated, "[88] such as polarity, molecular size, or ease of ionization. After a class of adsorption column (*e.g.* reverse-phase packing) is selected, off-the-shelf packed columns of that type are tested to find the best-performing for the particular separation. In our case, the two solutes exhibited similar polarity (see Section 5.2 below) and overall molecular size, and ion-exchange was not an option in the process solvent.

In previous work designing and synthesizing metal-organic frameworks to separate these species, Centrone *et al.*[?] noted that separating the species chromatographically using a standard C18 reverse-phase HPLC packing is only possible from an aqueous solution. The ability to separate the two species directly in ethyl acetate would eliminate the need for two costly solvent exchanges—from the organic solvent to water, and then from the organic solvent to water after the separation.

Because of the multi-step nature of pharmaceutical syntheses, and the frequent similarty of reactants and byproducts' chemical structures to those of desired products, it is expected that identifying a suitable adsorption medium to effect continuous, SMB-based separations would often present similar challenges. Indeed, as an example of the challenge of such separations (in an analytical, rather than manufacturing, context), MIP Technologies AB of Lund, Sweden, uses molecularly-imprinted polymer matrices to selectively retain one species among several with common chemical moieties.[109–111]

Overall, in this application I sought to (*i*) develop an *in silico* screening and molecular design approach, using molecular dynamics simulations for screening and "molecular evolution" for the design of molecular architectures; and (*ii*) apply this technique to develop solid surface-bound organic ligands, suitable for the selective separation of a particular pharmaceutical intermediate from solution in a process stream.

We see the role of this technique as broadly searching and screening the chemical space, thereby serving as a preliminary design or screening approach. Practitioners could apply this technique, and then choose the most promising designs from advanced generations for further computational or experimental testing. By providing a number of candidate designs, this method allows chemical scientists to consider factors not included in the automated evaluation (like feasibility or cost of synthesis, solubility or stability in a certain solvent, ease of disposal, *etc.*). Promising designs can also serve as examples or templates for modified versions of designs, allowing a chemist to improve upon the identified motif/design, or to use a similar, commercially-available compound.

Table 1-1. Previous approaches to molecular evolution.

program or author	purpose of molecular design	chromosome encoding	fitness function
Weininger[62]	fit to pharmacophore or resemble a given molecule	contemplated: 3D structure, connectivity graph, or SMILES string	contemplated: Tanimoto similarity[63] to a given molecule; presence of features such as rings, cationic sites; steric fit into 3D binding site; or experimental measurement
Chemical Genesis[64]	fit to pharmacophore or mimic a given molecule	3D structure [a]	similarity to desired 2D and 3D QSAR properties, and presence of features at specific interaction locations
PRO_LIGAND[65]	fit to pharmacophore or mimic a given molecule	3D structure	presence of features at specific locations
Nachbar[66]	mimic structure of a given molecule	hierarchical text expressions specifying topology	Atom-pair similarity or Dice similarity[67] to a given molecule
TOPAS[68]	mimic structure of a given molecule	graph of enumerated functional groups	Tanimoto smilarity to target molecule, or 2D topological similarity to pharmacophore
ADAPT[69]	design ligand to bind to a site on a protein target	subset of SMILES strings describing acyclic molecules	binding score produced by DOCK4.0 program,[70] plus penalty functions for violating QSAR constraints
LEA3D[71]	design ligand to bind to a particular site on a protein target	linear string of enumerated functional groups	FlexX 1.13.1[72] docking score
Molecular Evoluator[73]	design pharmaceutically active compounds, *e.g.* a ligand which binds a particular protein	modified version of SMILES with explicit hydrogen atoms	Human input, derived from judgment of purportedly expert user
Dey and Caflisch[74]	design ligand to bind to particular site on a protein	variable linking functional groups between fixed fragments known to dock at particular locations	sum of 2D similarity to known binding molecules, plus 3D similarity to known binding molecules, plus estimated binding energy from grid-based potential at binding pocket (CHARMm force field)
this work	selectively adsorb a molecule for separation	linear string of enumerated functional groups	energetic contribution to $\Delta\Delta F_{ads}$ from MD simulations

[a] "3D structure" means a molecule was manipulated directly in its three-dimensional representation, by altering the identity of atoms/functional groups, altering bonds, performing ring opening/closing operations, and/or by modifying the values of internal coordinates.

Chapter 2

Objectives and overview

2.1 Objectives of this work

The overall goal of the work described in this thesis was to use molecular simulations to design, at a molecular level, interfacial modifications to fluid-adjoining interfaces that could alter the adsorption/desorption equilibria or kinetics of solution components at that interface.

The specific objectives of this work are:

I. To understand the mechanism of water evaporation at a molecular level, and to use this understanding to develop soluble additives to diminsh the rate of evaporation at water's liquid/vapor interface.

II. To design a modified solid surface that could purify a solution of a pharmaceutical intermediate, by selective adsorption of unconverted reactant in reaction effluent solution.

This project also provides examples of two paradigms of molecular design using computer simulation. In the first paradigm, practioners employ computer simulations to understand the mecahnisms of physical or chemical processes at a molecular level of detail, a level of detail which may be difficult or impossible to obtain through experimental measurements. Based on this detailed mechanistic understanding, it is possible to design molecules which, as additives, would alter the physical process in

questions. These designs can then be evaluated using more detailed or more accurate molecular simulations, and experimental testing.

A second paradigm is to employ computer simulations as fast and cheap screening tools for a group of many possible molecular designs. This approach led to the promise of *in silico* screening of drug libraries for specific interactions with drug targets. Such screening techniques typically examine candidate molecules or designs in numbers that would not be possible using physical/experimental screening.

Our approach to screening is novel, however, in that we are carrying out directed evolution on molecular architectures. More specifically, our overall method of approach in the second paradigm is to first begin with a set of linear ligands that would be attached to a solid surface (such as gold or silica), to next evaluate their suitability using molecular simulations, for selectively adsorbing a particular pharmaceutical intermediate, designated E2, assign a fitness score to each, and finally to evolve population of such ligands using the fitness scores and genetic information, and repeating this process many times.

2.2 Overview of this thesis

This document is organized as follows. Further background about evaporation is provided in Chapter 3, along with a description of the simulation methodology used to identify the most likely reaction path for evaporation. Chapter 4 then describes the reaction path found in this study, and details the free energy and kinetic measurements that were made using computer simulations. The three related analyses were performed in order to identify important order parameters for that process, and these are also described in Chapter 4.

The design of surface-bound ligands is described in Chapters 5 and 6. In Chapter 5, I discuss the genetic algorithm approach, as well as the simulation procedures that underlie it. The molecular designs generated by the approach are presented in Chapter 6, along with measures of algorithm performance, and I present why the designs in question are expected to be effective for selective adsorption.

Finally, in Chapter 7, I summarize the conclusions of this thesis, and provide notes about possible future directions for related work.

2.3 Publications originating from this work

This thesis contains material from the following two articles:

"Design of linear ligands for selective separation using a genetic algorithm applied to molecular architecture," Nicholas Musolino, Erik E. Santiso, and Bernhardt L. Trout, submitted to *The Journal of Chemical Information and Modelling.*

"Insight into the molecular mechanism of water evaporation via the finite temperature string method," Nicholas Musolino and Bernhardt L. Trout, submitted to *The Journal of Chemical Physics.*

Chapter 3

Approach to studying water evaporation through molecular simulations

3.1 Physical and chemical background

The thermodynamics and kinetics of water evaporation have been examined as an example of a conceptually simple process occurring in a fluid with complex behavior. The kinetics of water evaporation has been studied as a question of both physical chemistry and transport phenomena. A review can be found in a monograph by Frank E. Jones.[9]

To calculate the rate of evaporation, it is possible to examine quantitatively the state of dynamic equilibrium between liquid water and a saturated water vapor phase in contact with it. The approach leading to the Hertz-Knudsen equation simply states that in this dynamic equilibrium, the number of molecules that enter the liquid phase from the vapor is equal to the number of molecules leaving the liquid, *i.e.* evaporating into the vapor, in a sufficiently long time period.

Since the behavior of gases is much more amenable to quantitative analysis, the derivation of the Hertz-Knudsen equation (adapted from Ref. 9) begins there.

The Knudsen equation[7] gives the flux G at which atoms or molecules in a gas will pass through a plane in one direction:

$$G = \frac{1}{4} N \bar{v}$$

where G is molar flux, N is the molar density of the gas, and $\bar{v}$ is the average velocity of the molecules in the gas. For the Boltzmann-distributed velocities of gas molecules, this latter quantity is:

$$\bar{v} = \left(\frac{8 k_B T}{\pi M} \right)^{1/2}$$

where M is the molecular weight. If the gas in question behaves as a perfect gas, then the number density is $N = P/(kT)$, so that

$$G = \frac{1}{4} \frac{P}{k_B T} \left(\frac{8 k_B T}{\pi M} \right)^{1/2} = \frac{P}{(2 \pi M k_B T)^{1/2}}$$

The mass flux J is simply $J = MG = P \left(\frac{M}{2 \pi k_B T} \right)^{1/2}$, with dimensions of mass per area per time.

To calculate the rate of evaporation from the liquid phase to the vapor at the interface, one can equate this to the amount of water striking the interface from the vapor phase *and entering the liquid phase*:

$$G^{evap} = C G^{HK}_{\text{from vapor}}$$

where J is the flux of vapor-phase molecules striking a plane derived above, and C is the *condensation coefficient*, and represents the fraction of molecules from the vapor phase that impinge upon the interface and actually enter the liquid phase. G^{evap} is the flux that originated from the liquid phase. Because this fraction also represents the portion of mass flux moving away from the interface that evaporated from the liquid, this phenomenological factor is also called an evaporation coefficient, and denoted γ_E.

This coefficient is typically unity for simple, monomolecular liquids, like mercury

and silver studied by Hertz and Knudsen. For water, however, Alty measured the rate of evaporation, and found that the evaporation coefficient was about 0.04 at 24 °C,[11] and in a refined experiment obtained a value of 0.036[12] at 21 °C.

In 1960, Mortensen and Eyring sought to explain the statistical-mechanical, and ultimately molecular, basis for the unexpected behavior of water and other non-simple liquids.[8] They began with the theory of absolute reaction rates, which makes use of a transmission coefficient κ in the Eyring-Polyani equation for a rate constant:

$$k' = \kappa \frac{k_B T}{h} \exp -\Delta G^{\ddagger}/k_B T$$

where $\Delta G^{\ddagger}$ is the Gibbs free energy of activation for the reaction, assuming the activated state is equilibrated. They showed that the transmission coefficient κ is:

$$\kappa = \frac{Q_e}{Q_i}$$

where Q_e is the molecular partition function in the surface state, and Q_i is the molecular partition function in the vapor state (excluding translation; subscript i indicates an initial state).

Mortensen and Eyring, citing spectroscopic data, pointed out that vibrational partition functions are often unchanged between surface and vapor molecules,[1] leaving only rotational degrees of freedom:

$$\kappa = \frac{Q_e}{Q_i} = \frac{f_e^{rot}}{f_i^{rot}}$$

Through thermodynamic arguments, this quotient was linked to the entropy of vaporization ΔS_v by comparison to a reference substance, CCl_4, with a condensation coefficient of unity. This approach had been developed in earlier work in which this quantity (the quotient $\frac{f_e^{rot}}{f_i^{rot}}$) was called the *free angle ratio*,[112] in analogy to the "free volume" approach that Eyring and co-workers used to describe simple molecular

[1] Despite the authors' use of the word *surface* to label the liquid molecules with which a condensing molecule is interacting, their analysis treats those molecules exactly as bulk liquid molecules.

liquids.

Through the approach cited above, the free angle ratio or κ was calculated to be 0.022 and 0.04 at 0 and 100 °C,[8] respectively, in good agreement with then-observed values of the condensation coefficient 0.036 and 0.04 at 20 and 100 °C.

3.2 Choice of intermolecular potential and simulation technique

The three-site SPC/E model[113] of water was used throughout the simulations in this study. This particular model was chosen because it exhibits an enthalpy of vaporization, self-diffusion coefficient, and dielectric constant in close agreement with water's actual values,[36] and for its previous success in reproducing an interfacial thermodynamic property, namely the surface tension of water.[24,114] A recent comparison of six water potentials[114] found that the SPC/E and TIP6P potentials "provide the best agreement [of surface tension] with experiment at all temperatures." As Alejandre *et al.* write,[24] such simulations truly challenge potentials, as they are extensions of a potential beyond the bulk-liquid conditions to which it was parameterized.

The general procedure to simulate an interface was to begin with a bulk-liquid simulation, and then extend the simulation cell size in the z direction, thereby creating a "slab" of water in the primary cell (see Figure 1 of Ref. 36). When periodic boundary conditions are applied, a lamella with thickness $\approx$ 30 Å was formed, with about 95 Å of vacuum separating it in either direction from the next periodic image of the lamella. The primary unit cell, along with its boundaries, is shown in Figure 3-1, and its time-averaged density profile is shown in Figure 3-2.

In order to validate our simulation procedure, the surface tension exhibited by the system was measured. To do so, the components of the pressure tensor were calculated at each recorded timestep: at each position z (sampled at 1 Å intervals), intermolecular forces from a pair of atoms contributed to the pressure tensor when the line segment joining those atoms crossed through the z-plane in question. Then,

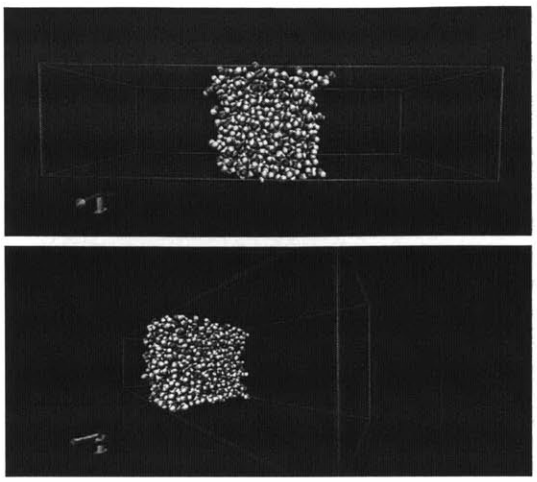


Figure 3-1. Rendering of 1025 water molecules in the $31 \times 31 \times (4 \times 31$ Å) unit cell. The z-axis is in the horizontal direction.

the surface tension was calculated from the components of the pressure tensor at each point in space:[26,115]

$$\gamma = \frac{1}{2} \int_{-\infty}^{\infty} (P_N(z) - P_T(z))\, dz \tag{3.1}$$

where $P_N(z)$ and $P_T(z)$ are the normal (in this case, zz) and transverse (xx and yy) components of the pressure tensor, respectively, at z.

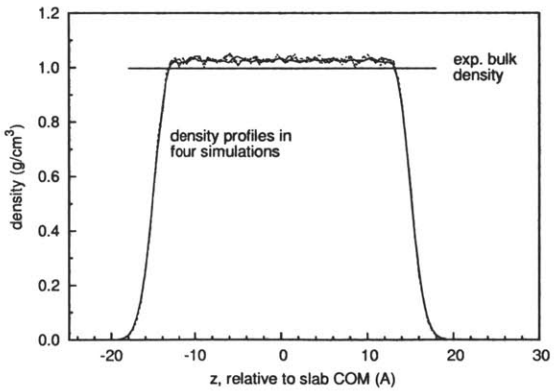


Figure 3-2. Time-averaged density profiles from four 2.0-ns simulations, with frames recorded every 5 ps.

These simulations and those described below were carried out in the canonical ensemble, using a timestep of 1.0 fs and rigid bond lengths. During equilibration and production, temperature was controlled using Langevin dynamics (298 K, damping coefficient 4 ps^{-1}) in NAMD.[116] Electrostatics were treated with the particle mesh

Ewald procedure (PME),[117,118] with grid size $32 \times 32 \times 128$. The use of PME has been shown to be important for obtaining accurate values of the surface tension in such systems.[24,114] Simulations of bulk liquid water were carried out in the NPT ensemble at 1.0 atm, with pressure controlled with by the Langevin piston approach implemented in NAMD, with oscillation period 200 fs and decay time 100 fs.

The surface tension measured in this way from a 1025-molecule, 2-ns simulation was $\gamma = 61 \pm 2$ dyn/cm; the statistical error was taken to be one standard deviation of the value of γ over eight block averages. This was in good agreement with Chen and Smith's "final value" of 61.3 dyne/cm for the same potential,[114] and the values of 61 to 62 dyne/cm obtained in other studies.[119–121]

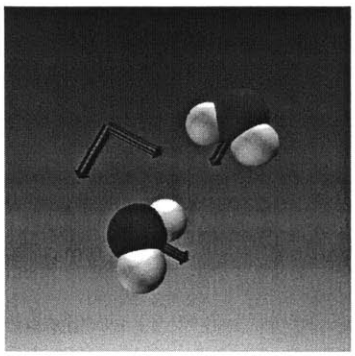
3.3 Identifying reaction mechanisms through use of order parameters

Pathways and mechanisms for physical or reactive processes can be identified quantitatively as a series of order parameter values. *Order parameters*, also called collective variables, are functions of the simulated system's atomic coordinates, and aim to quantitatively characterize the state of system, in this case as either liquid, vapor, or some intermediate state.

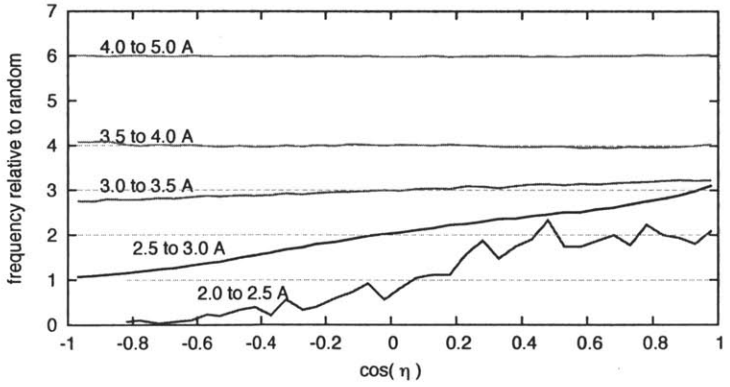
The order parameters (OPs) used for this study (listed in Table 3-1) will be introduced here, and detailed definitions can be found in the following section. Order parameter zero described the z-position of the evaporating molecule relative to the center of mass of the "slab" of other molecules; for reference, the Gibbs dividing surface (GDS) was located at $z = 15.1$ Å, using the same datum. The GDS was identified by finding the location at which the time-averaged density of the slab first reached $\rho_{\text{GDS}} = \frac{1}{2}(\rho_{\text{liq}} + \rho_{\text{vapor}}) \approx \frac{1}{2}\rho_{\text{liq}}$, since $\rho_{\text{vapor}} \approx 0.03\%\rho_{\text{liq}}$ at standard temperature and pressure. This location and the density profile itself were reproducible from simulation to simulation.

Order parameter 1 described the local density in the vicinity of the selected water

molecule, using a smooth weighting function to count molecules within about 3.5 Å. In essence, this order parameter includes contributions from the oxygen–oxygen and oxygen–hydrogen radial distribution functions, both measured at each frame from the selected molecule's oxygen atom. Order parameters 2 and 3 summarize the distribution of the angles between the evaporating molecules dipole vector and the dipole vectors of nearby molecules, as depicted in Figure (a). Figure (b) shows that the decay length for dipole correlation is about 3 to 3.5 Å, and this informed the weighting function used to define these variables.



(a) The angle η between dipole vectors (blue cylinders) of two water molecules; for clarity, the dipole vectors are translated and reproduced.

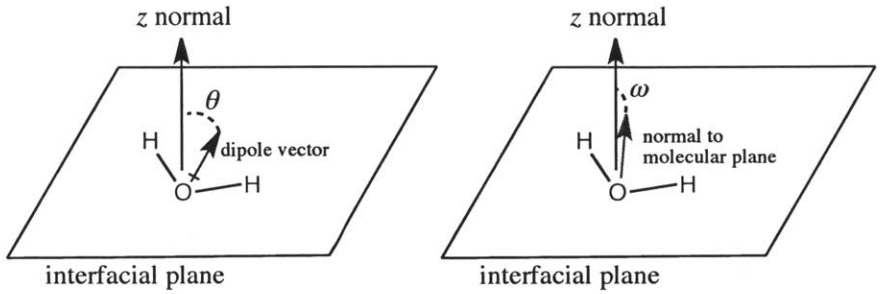


(b) Distribution of cosine of dipole-dipole angle η for water molecules with indicated O–O separation distances in a 1.0-ns bulk SPC/E water simulation.

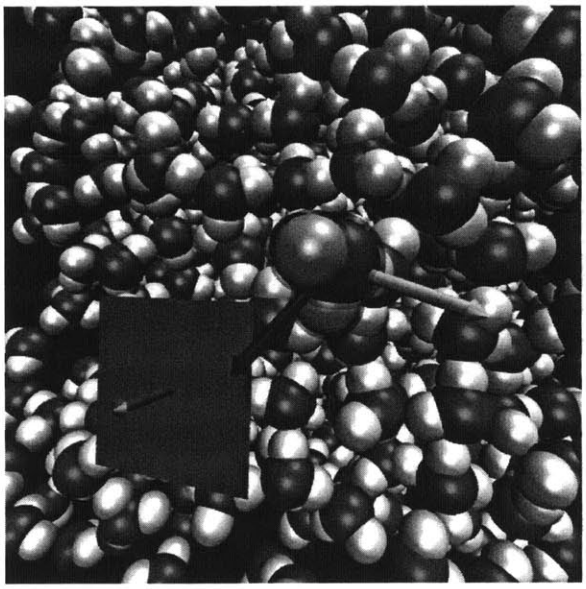
Figure 3-3. The dipole-dipole angle η and its distribution in bulk water.

Order parameters 4 and 5 summarize the "absolute" orientation of the evaporating

molecule, that is, its orientation relative to the interfacial normal $\hat{z}$. OP 4 measures the direction of the molecule's dipole vector as outward or inward facing, while OP 5 measures how outward-facing is the vector ν, perpendicular to the H−O−H plane. Graphical definitions of these angles are shown in Figure 3-4(a), and the dipole vector and molecular normal ν are shown in Figure 3-4(b).



(a) Schematic illustration of the two "absolute orientation" angles.



(b) Rendering of the dipole vector μ (blue cylinder), the molecular normal vector ν (yellow cylinder), and the interfacial normal $\hat{z}$ (gray cylinder).

Figure 3-4. The two "absolute" orientation variables θ and ω, used to define $q_4 = \cos(\theta)$ and $q_5 = \cos^2\omega$. The two angles are defined in relation to the interfacial normal vector and the evaporating molecule's dipole vector and the molecular normal, respectively.

Order parameters 6 and 7 count the number of hydrogen bonds the evaporating molecule is donating and accepting, respectively. Contributions are counted in a con-

tinuous manner using a weighting function, based on O $\cdots$ H distance. Finally, order parameters 8 and 9 were the distance- and angle-based tetrahedrality measures of Chau and Hardwick.[122] These order parameters can take values between zero, representing a perfectly tetrahedral arrangement of water's oxygen atoms, to a maximum value of 1 in a disordered state.

These order parameters, along with their derivatives with respect to atomic coordinates, were implemented in non-parallel C++ code in NAMD version 2.6.[123] Their derivatives with respect to atomic coordinates were also implemented, and allowed me to restrain the values of the order parameters.

Table 3-1. Description of order parameters used to describe state of water molecule near interface. Order parameters 8 and 9 have definitions that do not permit the imposition of forces, but listed force constants were used to calculate distances in order parameter-space.

OP	quantity measured for evaporating molecule	force constant range (kcal/mol/[OP]^2)
q_0^z	z-position of COM relative to slab COM	2.5–5.0
q_1^{lden}	local density	2.5–20
q_2^{ravg}	average of relative orientation to neighbors	2.0–10
q_3^{rstd}	standard deviation of relative orientation to neighbors	6.0–10
q_4^{thet}	orientation of dipole relative to interface normal	10–40
q_5^{omeg}	orientation of molecular normal rel. to interface normal	15–75
q_6^{don}	number of H-bonds donated	10–20
q_7^{acc}	number of H-bonds accepted	10–20
q_8^{tdist}	homogeneity of distance of four nearest neighbors	20–100
q_9^{tang}	angular tetrahedrality of four nearest neighbors	20–100

3.4 Interfacial order parameters and their definitions

All order parameters are measured for a specific, pre-selected molecule, which is denoted with superscript "a", and the z-axis is normal to the interfacial plane. The first order parameter is the distance between the center-of-mass of the evaporating molecule and the center-of-mass of all the other molecules, which collectively are

designated the "slab":

$$q_0 = r^a_{\mathrm{COM},z} - r^{\mathrm{slab}}_{\mathrm{COM},z} \qquad \text{z pos.}$$

The local density order parameters involves a sum of the masses of all atoms, weighted by their distance to the oxygen atom the evaporating molecule:

$$q_1 = \frac{1}{V_w} \sum_{\text{atoms } i} m_i w_{lcl}(r_i - r^a_{\mathrm{O}}) \qquad \text{local density}$$

where the normalization factor V_w is the bulk density integrated using the weighting factor describing the local vicinity:

$$V_w = \int_0^\infty \rho_{bulk} w_{lcl}(r) 4\pi r^2 dr$$

The w_{den} function is one of two smoothing functions used to define what neighbors are local, and what pairs of atoms are hydrogen bonding (see below). These smoothing functions, rather than sharp distance cutoffs, are used to make the order parameters differentiable functions of atomic coordinates, which is required to apply conservative restraint forces during the restrained simulations (see Section **??**).

$$w_{lcl}(\mathbf{u}) = \frac{1}{1 + \exp\left(\kappa(|\mathbf{u}| - R_{lcl})\right)}$$
$$w_{hb}(\mathbf{u}) = \frac{1}{1 + \exp\left(\kappa(|\mathbf{u}| - R_{hb})\right)}$$

where $\kappa^{-1} = 0.2$ Å; $R_{lcl} = 3.25$ Å; and $R_{hb} = 2.3$ Å. These functions are graphed in Figure 3-5 below. The local density order parameter includes the mass of the evaporating molecule itself, and thus has a minimum value of M_a/V_w.

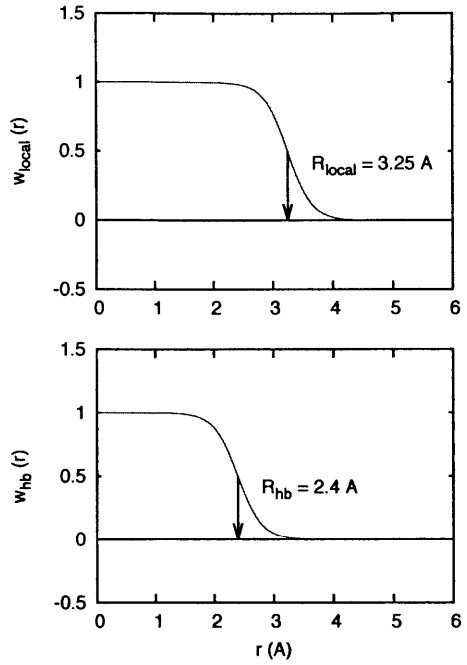


Figure 3-5. Smooth weighting functions used for calculating local density and local relative orientational order (top), and number of hydrogen bonds donated and accepted (bottom).

The next order parameters involve the relative orientation of the evaporating molecule a to its neighbors. For any two water molecules, this relative orientation is described by the angle between their dipole moments:

$$\eta_{a,i} = \arccos \frac{\mu^a \cdot \mu^i}{|\mu^a||\mu^i|}$$

where μ^j is the geometric dipole given by $\mu^j = r^j_{\mathrm{H1}} - r^j_{\mathrm{O}} + r^j_{\mathrm{H2}} - r^j_{\mathrm{O}}$. Once a set $\{\eta_{a,j}\}_{j=1}^{N-1}$ of relative orientations has been generated, their average $\langle 2\eta \rangle$ and variance

var (2η) are calculated using special approaches for angular random variables:

$$\bar{x} = \frac{1}{W_{tot}} \sum_{\text{mol. } j \neq a} w_{lcl}(r_{\text{O}}^{j} - r_{\text{O}}^{a}) \cos 2\eta_j$$

$$\bar{y} = \frac{1}{W_{tot}} \sum_{\text{mol. } j \neq a} w_{lcl}(r_{\text{O}}^{j} - r_{\text{O}}^{a}) \sin 2\eta_j$$

$$W_{tot} = \sum_{\text{mol. } j \neq a} w_{lcl}(r_{\text{O}}^{j} - r_{\text{O}}^{a}) \qquad \text{normalization factor}$$

$$q_2 = \arctan \frac{\bar{y}}{\bar{x}} \qquad \text{average rel. orient.}$$

$$q_3 = \left(-2 \ln \left(\bar{x}^2 + \bar{y}^2\right)^{1/2}\right)^{1/2} \qquad \text{std. dev. of rel. orient}$$

where the arctan function is calculated in the interval from $[0, 2\pi)$ based on the signs of both $\bar{x}$ and $\bar{y}$, using the `atan2` function of the C++ standard library.

The orientation of the water molecule relative to the interfacial normal $\hat{z}$ can be completely specified using two angular variables: the angle θ between the evaporating molecule's dipole and the normal, and the angle ϕ between the interfacial normal and the normal to the plane formed by the molecule's three atoms.

$$q_4 = \cos\theta = \frac{\mu \cdot \hat{z}}{|\mu|} \qquad \text{angular orient.}$$

$$\mu = (r_{\text{H1}} - r_{\text{O}}) + (r_{\text{H2}} - r_{\text{O}})$$

$$q_5 = (\cos\omega)^2 = \left(\frac{\nu \cdot \hat{z}}{|\nu|}\right)^2 \qquad \text{angular orient.}$$

$$\nu = (r_{\text{H1}} - r_{\text{O}}) \times (r_{\text{H2}} - r_{\text{O}})$$

These definitions are illustrated schematically in Figure 3-6.

The reason that the order parameter q_5 is defined with the square of the cosine is to account for symmetry. Because of this, the proper approach is to take the angle between the *directed* z-normal and the *directionless* normal to the molecular plane. This OP gives the same value, whether the computationally used normal is facing

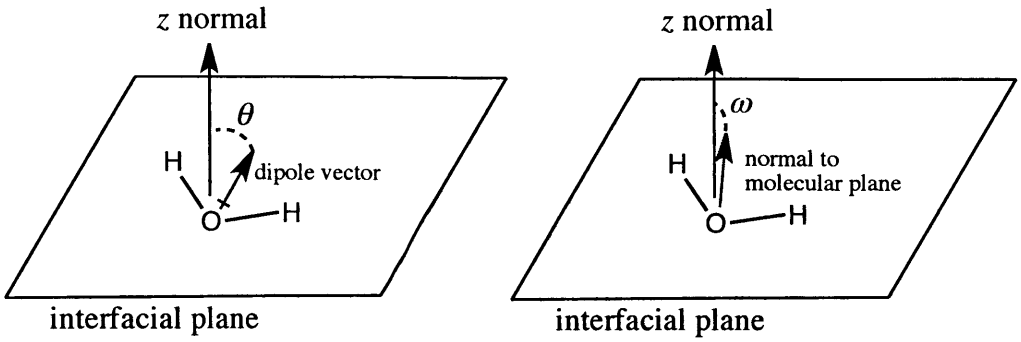


Figure 3-6. Schematic illustrating definitions of two orientation variables.

inwards or outwards. Using the square of the cosine is equivalent (in information content) to using the cosine of twice the measured angle.

Order parameters q_6 and q_7 use a smoothing function to count the number of hydrogen bonds. The H-bond weighting function w_{hb} (defined above) smoothly changes from a value of 1 to 0 at a cutoff of $r_{hb} = 2.3$ Å. The value of this cutoff was chosen based on the O–H RDF of water.

$$q_6 = \sum_{\text{H1, H2}\in a} \sum_{\text{O}\notin a} w_{hb}\left(r_{\text{H}}^a - r_{\text{O}}\right) \qquad \text{H-bonds donated}$$

$$q_7 = \sum_{\text{H}\notin a} w_{hb}\left(r_{\text{H}} - r_{\text{O}}^a\right) \qquad \text{H-bonds accepted}$$

Finally, two order parameters measure the tetrahedrality of the evaporating molecule's local environment. For purposes of these OP measurements, the local environment is defined as the four nearest neighbors, as measured by oxygen-oxygen distance. These neighbors are shown at two snapshots in Figure 3-7

The first tetrahedrality OP, designated q_8, measures the variance of the oxygen-oxygen distances from their mean value. For a perfect tetrahedral arrangement, this order parameter would be zero. The second tetrahedrality OP, q_9, measures the deviation of the neighbor-central molecule-neighbor angles (denoted $\psi_{j,k}$ between neighbors

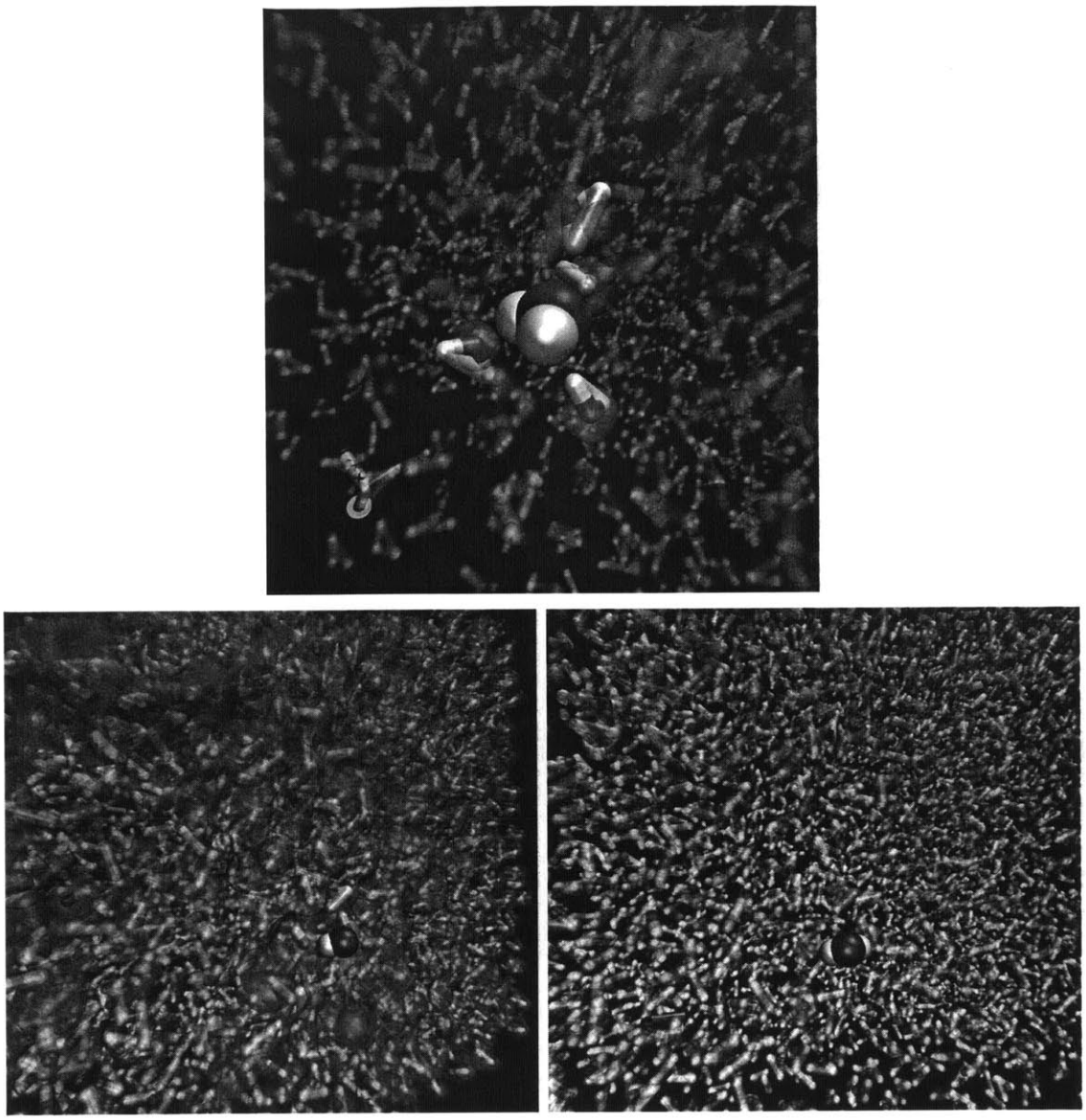


Figure 3-7. Selection of four nearest neighbors for use in tetrahedrality order parameters in bulk water (*top*) and at an interface (*bottom*). The four nearest neighbors (as measured by O–O distance) are highlighted in yellow. In general, the arrangement of the four nearest neighbors is much more tetrahedral in bulk water.

j and k) from the value of 109.5° they would take in a tetrahedral arrangement.

$$q_8 = \frac{1}{3}\sum_{k=1}^{4} \frac{\left(|r_{\mathrm{O}}^{j} - r_{\mathrm{O}}^{a}| - \bar{r}\right)^2}{4\bar{r}^2} \quad \text{with} \quad \bar{r} = \frac{1}{4}\sum_{k=1}^{4} |r_{\mathrm{O}}^{j} - r_{\mathrm{O}}^{a}| \qquad \text{dist. tetrahed. meas.}$$

$$q_9 = \frac{3}{32}\sum_{j=1}^{3}\sum_{k=j+1}^{4} \left(\cos\psi_{j,k} + \frac{1}{3}\right)^2 \quad \text{with} \quad \cos\psi_{j,k} = \frac{(r_{\mathrm{O}}^{j} - r_{\mathrm{O}}^{a}) \cdot (r_{\mathrm{O}}^{k} - r_{\mathrm{O}}^{a})}{|r_{\mathrm{O}}^{j} - r_{\mathrm{O}}^{a}|\,|r_{\mathrm{O}}^{k} - r_{\mathrm{O}}^{a}|} \qquad \text{angular tetrahed. meas.}$$

Simulations in the microcanonical ensemble were performed to demonstrate the conservative nature of constraint forces associated with order parameters 0–7, indicating that each was correctly implemented in NAMD's[116] C++ code.

3.4.1 Demonstration of conservative nature of restraint forces

Before employing restraints on the order parameters, it was necessary to calculate their gradients with respect to all atomic coordinates. Then, restraints of the order parameter to a particular target value q_j^* were imposed as follows:

$$U_{res}(q_j) = \frac{1}{2} k_j \left(q_j - q_j^*\right)^2$$
$$F_i = -\nabla_{r_i} U_{res} = -k_j \left(q_j - q_j^*\right) \nabla_{r_i} q_j$$

where j is an index for order parameters, i is an index for atoms, and F_i is the restraint force on atom i.

In order to test whether these restraint forces were correctly derived and implemented in the modified C++ code of NAMD(version 2.6), I performed microcanonical (constant-NVE) simulations in which only one order parameter (or none) was restrained.

For each order parameter, four microcanonical simulations were conducted, using both a 0.5- and 1.0-fs timestep. Figure 3-8 and the subsequent figures show that a 1.0-fs timestep is acceptable for the restraints on order parameters 0 (not shown), 1, 4, 5, 6, 7.

For each OP test, four simulations were carried out for different values of the restrained order parameter; for example, when testing the gradients/forces for OP 1, the local density around a water molecule, the target value was set to 1.05, 1.05, 0.85, and 0.40 in the four simulations, to test the application of forces in different configurations representing bulk water, an interfacial configuration, and a near-vapor configuration.

Figures 3-8 (bottom panel), 3-9, and 3-10 show representative results for five order parameters. In these simulations, the total system energy (red lines) varied in a range

of 0.3 kcal/mol, consistent with the energy conservation observed in a pure, restraint-free NVE simulation (Figure 3-8, top panel). For brevity, results from this series of 32 simulations (4 particular values for each of 8 order parameters) are omitted here; all results were similarly conservative of energy, indicating that the restraint forces calculated using the gradients of the order parameters defined in Section 3.4 are correct.

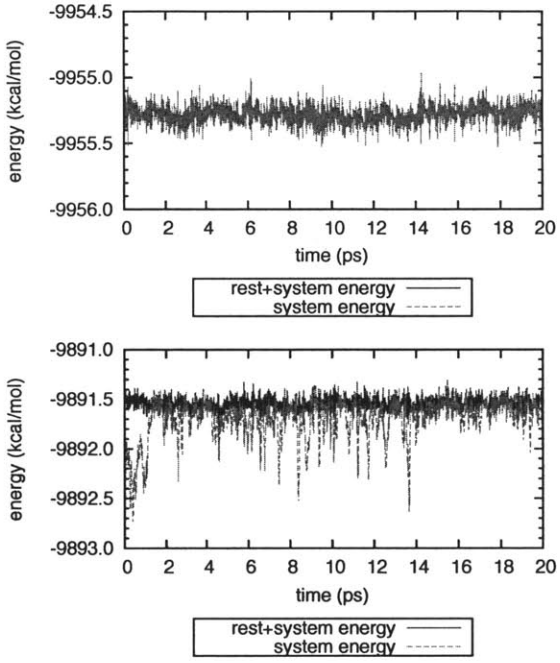


Figure 3-8. Accuracy of energy conservation in microcanonical simulation in which no order parameters were restrained (*top*), in which the range of variation due to imperfect numerical integration is about 0.3 kcal/mol (*i.e.* per mol of simulated systems). Demonstration of energy conservation in microcanonical simulation in which order parameter 1 (local density) is restrained (*bottom*). The lower line is the configurational energy calculated using the potential; the upper line is that configurational energy plus the restraint energy.

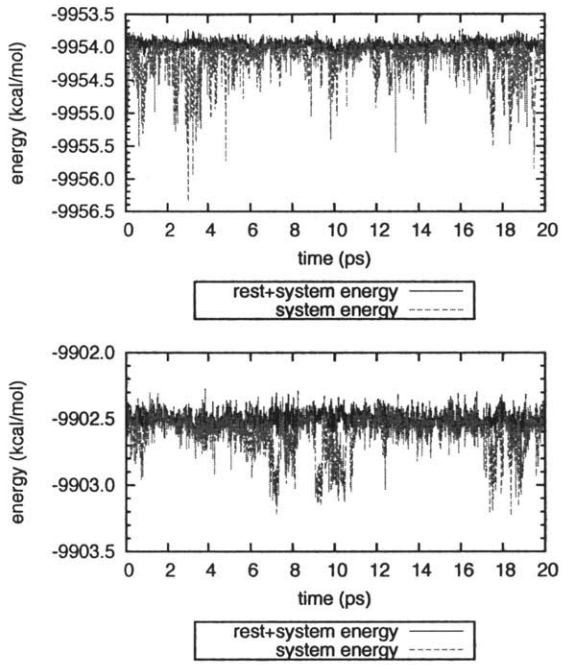


Figure 3-9. Demonstration of energy conservation in microcanonical simulation in which order parameter 4 (orientation of dipole, *top*) and order parameter 5 (orientation of water's molecular plane, *bottom*) is restrained. The lower line is the configurational energy calculated using the potential; the upper line is that configurational energy plus the restraint energy.

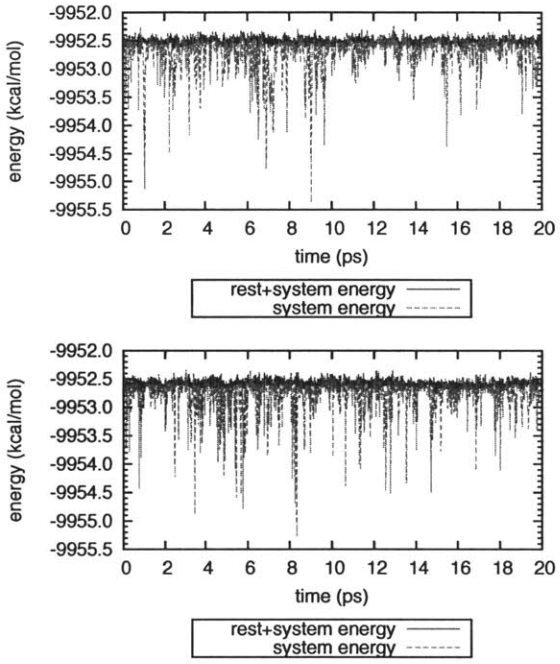


Figure 3-10. Demonstration of energy conservation in microcanonical simulation in which order parameter 6 (number of hydrogen bonds accepted, *top*) and order parameter 7 (number of hydrogen bonds accepted, *bottom*) is restrained. The lower line is the configurational energy calculated using the potential; the upper line is that configurational energy plus the restraint energy.

3.5 Procedure for identifying most likely reaction pathway

To study the evaporation process, we employed the string method in collective variables[124] (SMCV), which is based on the finite-temperature string method.[125–128] Our goal was to identify the minimum free energy path (MFEP) from a single molecule's bulk liquid state to its evaporated state; this path is the most likely path for transitions from the former state to the latter.

For a physical process, the minimum free energy path can be determined by performing restrained dynamics simulations with several copies of the system (called "replicas" or "images") along a pathway constituting a transition, using the SMCV procedure. In these simulations, the order parameters are restrained to target values using a restraint potential of the form $U_{rest} = \sum_i \frac{1}{2} k_i \left(q_i - q_i^*\right)^2$, where q_i^* is each OP's target value, and where gradients $\nabla_{\mathbf{x}} \left(q_i\right)$ of the order parameters with respect to atomic coordinates used to calculate forces on individual atoms. The restraint forces in all images[2] along the string are then used to calculate the next iteration of the string, which should be closer than its predecessor to the MFEP. In this study, the initial string was created using OP measurements from a series of previous simulations in which the single order parameter q_0 was restrained.

Once the evolving string converged to a final MFEP, the free energy profile and the mean first-passage time for the evaporation process was computed using milestoning, carried out using the boundaries of Voronoi cells in order parameter-space,[129–132] since the boundaries of Voronoi cells supported by the MFEP points in order parameter-space are optimal milestones.[130]

Further details about the MFEP identification and Voronoi milestoning are provided in Sections 4.1 and 4.2 in the next chapter.

[2]In this thesis, the word "image" typically refers to the several copies of the system, which are subjected to independent MD simulations, rather than a spatially translated set of coordinates under periodic boundary conditions.

Chapter 4

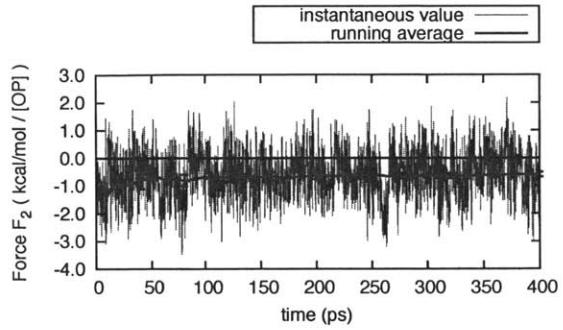
Elucidation of mechanism, reaction thermodynamics, and kinetics of evaporation

4.1 Most likely path of evaporation, as quantified by order parameters describing local physico-chemical environment

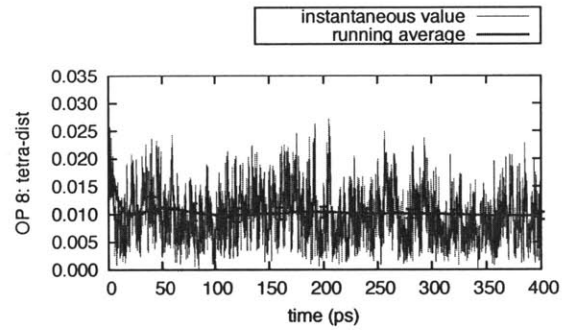
To identify the minimum free energy path, a string comprising $N_{img} = 16$ images was constructed. The initial target values for the order parameters in each image were chosen based on previous simulations in which only the relative z-coordinate q_0^z was restrained. Restrained molecular dynamics was performed, with production times of 125 to 500 ps in each iteration.

In accordance with the SMCV procedure,[124] values of the order parameters, restraint forces, and metric tensor were recorded every 100 fs. In general, restraint forces reached a steady value after 10 to 20 ps; examples are shown in Figure 4-1. After each iteration, recorded data were used to compute the potential of mean force and the target OP values for the next iteration of the string, placing images at equal

arc-length intervals along the string. This process was repeated until the new string was not far from its predecessor, as measured by Frechét distance.[133]



(a) Restraint force for order parameter $q_2^{ror-avg}$.



(b) Value of order parameter $q_8^{tetra-ang}$.

Figure 4-1. Examples of restraint force and order parameter convergence from image 9 of string 4.

Two changes were implemented over the course of the string evolution procedure. First, after string 17, the definition of $q_2^{ror-avg}$ was modified to return values in the range $[0, 2\pi)$, rather than in the range$(-\pi, \pi]$, as it originally did. Target values in string 18 were shifted to match the new definition. Second, after string 29, the recorded values of $q_8^{tetra-dist}$ and $q_9^{tetra-ang}$ in the string were set to their average values in the last simulation; before that, their values were simply the result of the movement and parameterization of the string in order parameter space. The values of $q_8^{tetra-dist}$ and $q_9^{tetra-ang}$ along the final string, measured in non-restrained simulations (described in the next section), are shown in Figure 4-2.

The distance of the evolving string from the initial string is shown in the upper panel of Figure 4-3. Because the string took a large step from its initial state in the first SMCV iteration, we also measured the distance from the second or third string,

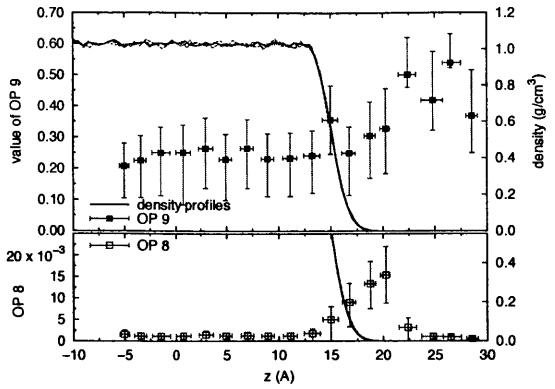


Figure 4-2. Values of tetrahedrality order parameters in each Voronoi dynamics image. Bars indicate one semi-standard deviation over simulation trajectory.

and confirmed that these distances were not evolving when iteration was stopped. The two changes in string evolution mentioned above explain the positive deviations in the lower panel of Figure 4-3.

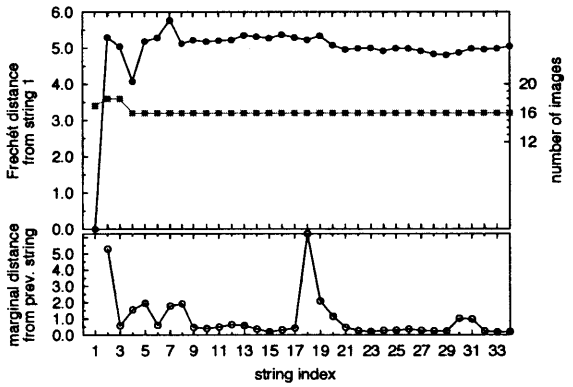


Figure 4-3. Frechét distance from initial string, and Frechét distance from each string's predecessor. See text for notes about methodological adjustments at string 18 and string 29.

The MFEP obtained from SMCV is depicted in Figure 4-4, and the transition from a bulk liquid to a vapor state can be described as follows:

1. From the bulk, the water molecule diffuses toward the interface, with increasing q_0 values. During this time, the values of q_2, q_3 describing the orientation of nearby molecules are approximately constant, as are the hydrogen bond counts and the tetrahedrality OPs. The absolute orientation OPs change, although their values are not physically important in the bulk phase. This diffusion is

represented by a gradually decreasing mean first-passage time, although it is difficult to see in Figure 4-10 below.

2. The water molecule enters "inner interface" region, (inside the Gibbs dividing plane), and its dipole vector gradually shifts from being somewhat in-plane to become outward-facing. This is *unlike* the typical interfacial molecule, which has its dipole in-plane.[25] The local density is not significantly lower than the bulk, suggesting first solvation shell still surrounds the water molecule; accordingly, its hydrogen bond values are bulk-like as well, with $\mathrm{D} + \mathrm{A} \approx 2 + 2$.

3. Next, the water molecule loses one of the hydrogen bonds it is donating, and it rotates around its normal vector ν to becomes more outward-directed, with a dipole directed about 40° from the normal $\hat{z}$. At this point, $\cos\omega \approx 0.3$, about half of its maximum possible value given the value of θ, indicating one O–H vector is more nearly in plane than the other, outward-facing hydrogen.[1] In this position, the molecule necessarily stops donating a second H-bond with its outward-facing O–H, and on average accepts one H-bond.

4. As the water molecule moves to the outer fringes of the interface, it rotates (again, about its molecular normal axis) so that its dipole is more outward-facing, about 20° from $\hat{z}$, and no longer makes the H-bond it had been donating, leaving only one accepted H-bond, At this point, the time-averaged density is about 0.05 g/cm^3; and there are few atoms within the 3.25- Ådensity averaging radius, except for the donor hydrogen.

5. With both O–H bonds facing outward, the single hydrogen bond from a neighbor, which had been holding the molecule in place, can break, and at this point, the molecule is free.

The recorded frames in each image in which the system was closest to its target OP values, as measured by minimum restraint energy, were used to generate the snapshots shown in Figures 4-5(a) and 4-5(b).

[1]Geometrically, the maximum value of $\cos\omega$, and thus q_5^{omeg} given a certain value of q_4^{thet} or θ, is $\cos\omega^{max} = \cos(\pi/2 - \theta)$.

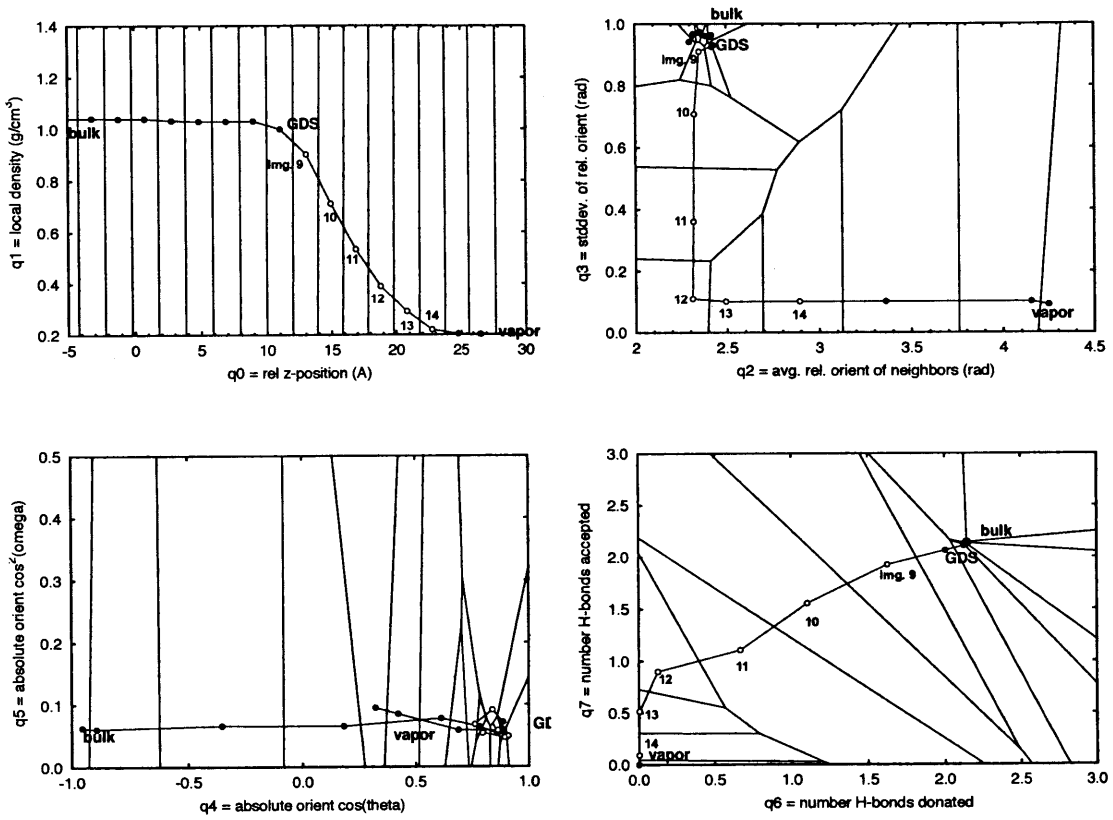
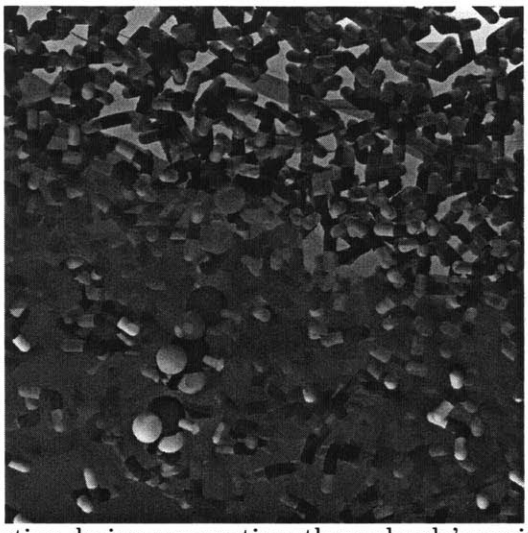
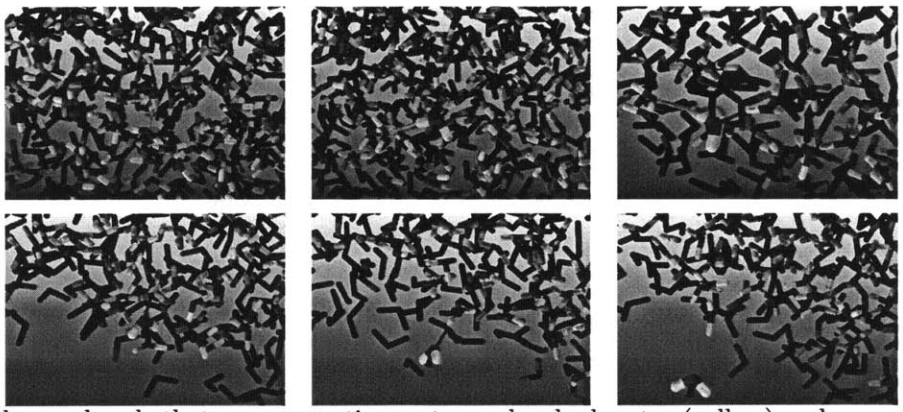


Figure 4-4. Minimum free energy path for evaporation, along with Voronoi cell boundaries between images, projected onto two order parameter dimensions at a time. The point labeled "GDS" is image 9, which contains the plane $q_0 = z_{GDS}$, the Gibbs dividing surface. The free energy changed most dramatically over the images (numbers 10–13), highlighted with white centers. Note that Voronoi cell boundaries do not necessarily appear normal to the string because they respect the scaling of order parameters (see text), and because of the plots' axis scaling.



(a) Water molecule's orientation during evaporation; the molecule's position and orientation (subject to translation) from images 9–13 is shown in a single figure. The other molecules' configuration is from image 9. The blue and yellow vectors are the dipole and molecular normal vectors, respectively. The transparent surface is the water surface.



(b) Hydrogen bonds that an evaporating water molecule donates (yellow) and accepts (purple) in images 8–13. The blue arrow is the dipole vector of the evaporating molecule.

Figure 4-5. Snapshots from the frames in images 8–13 in which the system was closest to its OP target values, as measured by minimal restraint energy. In these images, $q_0 = z$ varied from 13 to 21 Å.

4.2 Free energy and kinetics of evaporation along most likely reaction path

The potential of mean force (PMF) calculated with this procedure is not equivalent to the physical free energy along the reaction path. The PMF is a function of N_{OP} variables, namely all the order parameters, while the free energy is a function of a single variable, namely the fractional distance α along the path between reactant and product states. More specifically, the PMF is a free energy value calculated as an integral over all microstates which take a particular value of order parameters:

$$PMF(\mathbf{q}^*) = -\frac{1}{\beta} \ln Z^{-1} \int d\mathbf{x} \exp\left(-\beta E(x)\right) \delta\left(\mathbf{q}(x) - \mathbf{q}*\right)$$

where q^* is the particular value of order parameters, and δ is the Dirac delta function, which is zero when any elements of its vector argument are non-zero.

Once the MFEP was obtained, the image points were used as the support points for Voronoi dynamics, which can be used to measure free energy differences and mean reaction times. The MFEP is well suited for this taks since the boundaries between such Voronoi cells are expected to be, in general, optimal milestones.[129] Two additional images were added, one at each end of the string, to ensure that the final milestone, which separates image N_{img} from image $N_{img} - 1$, was outside the region of free energy change. In these Voronoi dynamics simulations, all entries and departures to and from Voronoi cells were recorded, along with the number of steps during which the simulated system was within its home cell, Instead of reversing the system's velocities at cell boundaries, half-pseudoharmonic soft-wall restraints were used, as described in Ref. 134, with force constant $k_w = 14.0$ kcal/mol.

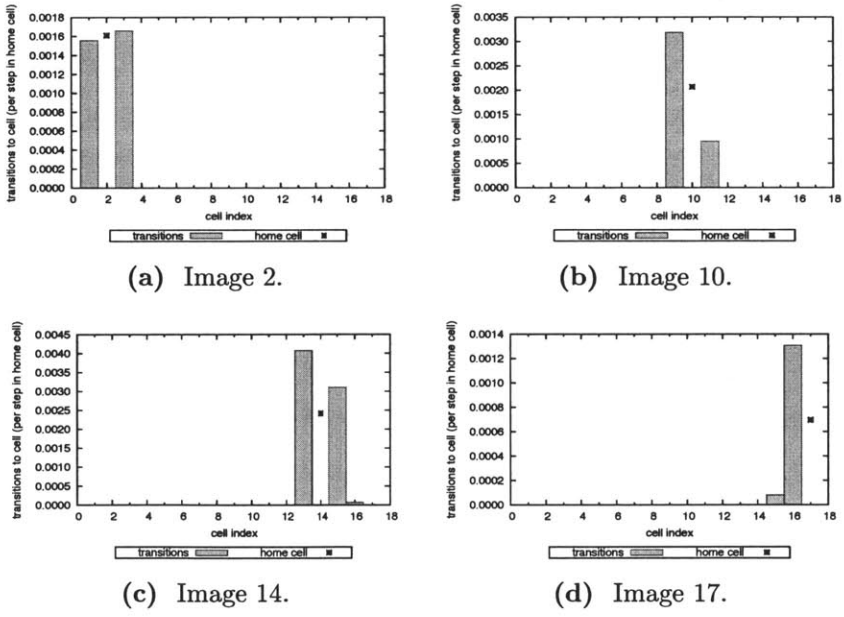
In dividing any space into Voronoi cells, it is necessary to establish a distance metric, because each cell is defined as the set of points in the space closer to one central point than to any other points. The ten different order parameters used in this study had different natural ranges of variation—for example, the number of Hydrogen bonds donated by a molecule might vary from 0.0 to 2.5, while the distance-

based tetrahedrality measure varies from 0.000 to 0.002. Because of this, a scaling factor was used in each dimension in order parameter-space. This scaling factor was taken to be the inverse of the force constant:

$$d(\mathbf{q},\mathbf{r}) = \left(\sum_{\text{OPs } i} \frac{1}{k_i^2} (q_i - r_i)^2 \right)^{1/2}$$

The value of each force constant are listed in Table 3-1 above.

Production MD was carried out for 2.0 ns. Transition events to neighboring cells were counted, and transition rates from four simulations are shown in Figure 4-6. In most images, almost all transitions took place to sequential cells, i.e. from cell j to cells $j-1$ and $j+1$.



(a) Image 2. **(b)** Image 10.

(c) Image 14. **(d)** Image 17.

Figure 4-6. Transition frequencies from home cell to other cells for four selected images during Voronoi dynamics simulations. Simulations for other images exhibited transitions to sequential cells, as in panels (a) and (b).

Using milestoning analysis, the free energy (FE) of the system at each milestone was determined, as shown in Figure 4-7, along with mean first-passage time to the final milestone, shown in Figure 4-10. The free energy profile is shown in Figure 4-7, and contains a flat region in the bulk phase, a change in the interfacial region of changing density, and then levels off once the molecule has broken free into the vapor phase,

with no FE maximum. The character of the profile is similar to others computed using similar simulations, with only one parameter (z-position) restrained.[32,33,35] The total free energy change is 7.4 kcal/mol, which is in good agreement with the same value measured by Taylor and Garrett,[32] and slightly larger in magnitude than the value of 6.8 kcal/mol observed by Vácha *et al.*,[33] both for SPC/E water.

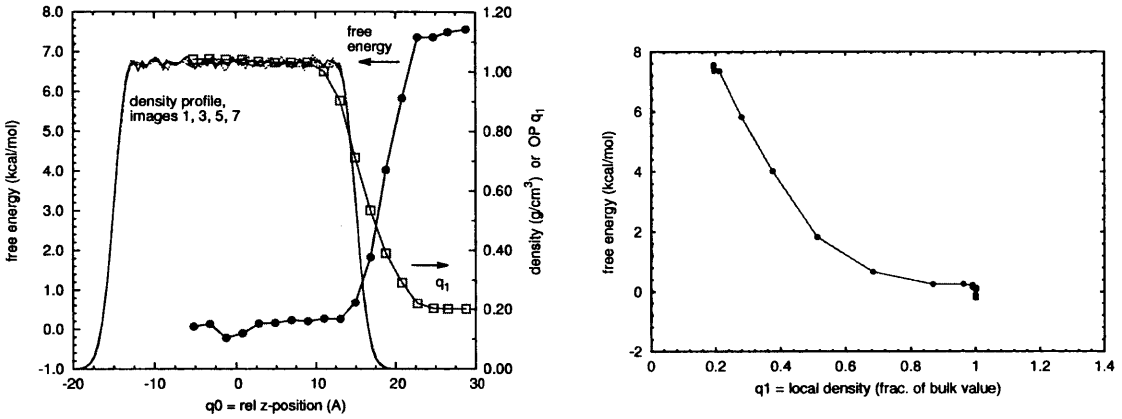


Figure 4-7. Free energy measured through Voronoi milestoning, as a function of order parameter q_0 = relative z-position (*left*) and order parameter q_1 = local density (*right*). The local density achieves a minimum value in the vapor phase when only the evaporating molecule itself is contributing to the local density.

The free energy profile is reproduced in Figures 4-8 and 4-9. The majority of the free energy change takes place after the evaporating molecule has reached the surface, where the number of donor and acceptor hydrogen bonds is $(D, A) = (1, 1)$. In fact, about 2 kcal/mol of FE change occurs as the molecule transitions from $(1, 1)$ to $(0, 1)$, and about 1.5 kcal/mol between $(0, 1)$ and $(0, 0)$.

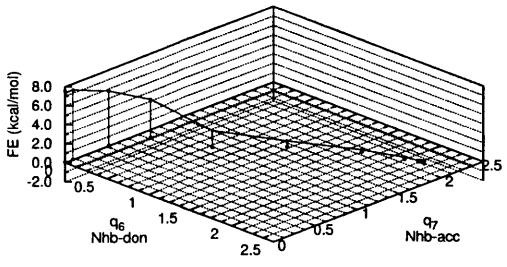


Figure 4-8. Free energy as a function of order parameters 6 and 7.

Figure 4-9 shows that the average energy, which was measured only when each

simulation was inside its respective home cell and free of restraint energy, increases by about 11.5 kcal/molas the water molecule evaporates. This energy penalty, then, must be offset by a corresponding increase in entropy upon evaporation: in the liquid phase, the water molecule is part of a tetrahedral network which extends throughout the bulk, and therefore is severely restricted in its rotational degrees of freedom. As the water molecule leaves the bulk, these restrictions are loosened, although even with two hydrogen bonds, the molecule may have only one or zero unrestricted rotational degrees of freedom.

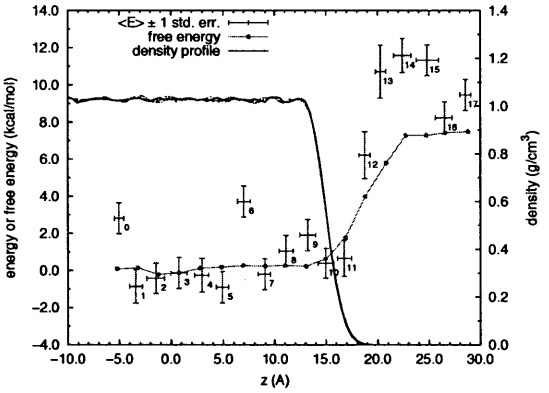


Figure 4-9. Free energy profile, along with average system energy values. Error bars are 1.5 standard errors.

There also appears to be a peak in the energy profile at around $z = 20$ to 24 Å, and lower energy values at 27 and 29 Å. It appears that these higher energy values occur because as the evaporating restrained water is restrained a few angstroms above the interface, other water molecules continue to solvate the evaporating molecule, making a total of 3, 2, or 1 hydrogen bonds. The other molecules in this shell extend beyond the GDS, and have *their own* local hydrogen bond networks disrupted. Once the evaporating molecule loses all its hydrogen bonds, the "protrusion" of solvating shell can reform into the flat interface, thereby minimizing the number of molecules with fewer than a full complement of hydrogen bonds. This is analogous to the role of surface tension effects in the separation of a macroscopic droplet from bulk liquid phase.

The mean first-passage time (MFPT) from each milestone is plotted in Figure 4-

10. The overall MFPT from the first milestone, in the bulk region, is 1375 ns. While the evaporating molecule is in the bulk liquid portion of the slab, the MFPT slowly decreases, although this behavior is difficult to see with the scale of Figure 4-10. This portion of the MFPT profile corresponds to diffusion in the z-direction. Then, beginning at the milestone where $(D, A) = (1.4, 1.8)$, the MFPT starts to decrease more dramatically; the greatest change in the MFPT, indicating the *slowest* part of the evaporation process, occurs when the water molecule loses its *final*, accepted hydrogen bond, *i.e.* the transition from $(0, 1)$ to $(0, 0)$. This corresponds to one of the larger (not the largest) changes in free energy discussed above.

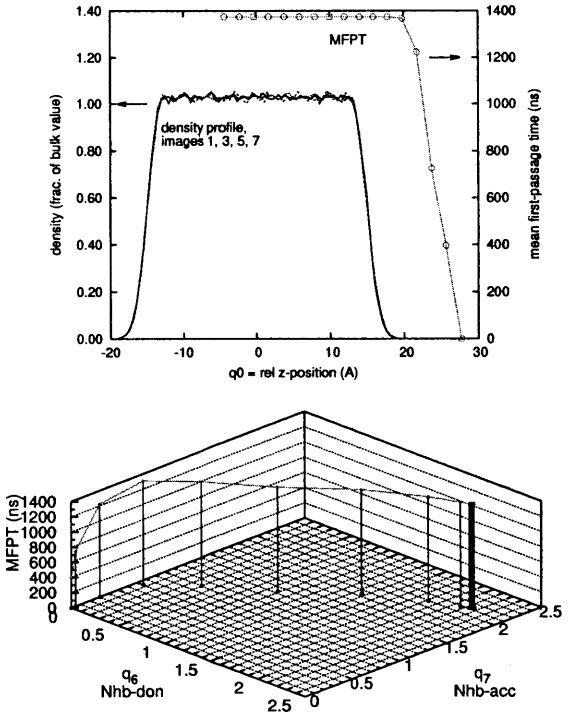


Figure 4-10. Mean first passage time to the final milestone as a function of order parameter q_0 = relative z-position (*top*) and as a function of q_6 and q_7, the number of hydrogen bonds accepted and donated (*bottom*).

4.3 Comparison to experimental results

The free energy difference corresponding to the transfer of a solute molecule from a vapor into solution has been termed the "free energy of solvation" by Ben Naim and

Marcus,[135] who showed that values can be obtained from vapor-liquid equilibrium and other data, interpreted through thermodynamic arguments. The evaporation process is the reverse of the self-solvation process for water, with $\Delta G_{evap} = -\Delta G_{solv}$. The free energy change measured in this study, and its enthalpic and entropic components, are compared with experimental values in Table 4-1.

In these simulations, the formal system size was fixed, which would suggest the simulations were carried out in the canonical ensemble, leading to measured FE values that are Helmholtz free energy differences. However, because the volume physically occupied by the system of molecules could fluctuate, practitioners have argued that the systems behave as if in the NPT ensemble, so that the free energies measured should be directly compared to experimental Gibbs free energy values.[32,40]

Table 4-1. Comparison of simulation measurements to experimental values for the evaporation or "desolvation" process at 298 K. All values are in kcal/mol.

	ΔG_{evap}	ΔH_{evap}	$-\Delta S_{evap}T$
SPC/E water	7.4 $\pm$0.4	11.5 $\pm$1.0	$-$4.2 $\pm$1.4
actual[135]	6.23	9.97	$-$3.64

The error bars reported in Table 4-1 come from examiming the free energy profile in the bulk-liquid region of the system, where it is expected to be constant. Overall, the results obtained show good agreement with the actual values for water, considering the simiplicity of the water model used, and in particular its lack of polarizability.

The evaporation flux implied by these simulation measurements can also be calculated, using the mean first-passage time. We chose a particular water molecule to evaporate, so the mass flux corresponding to our MFPT is $G = M\frac{1}{\tau}\frac{1}{a}$, where a is the specific area occupied by a water molecule. Counting the water molecules in the bulk liquid phase intersected by the plane $z = 0$ (within the SPC/E molecule's van der Waals radius of 1.76 Å) at each frame in 2-ns slab simulations, a was 8.28 $Å^2$. This leads to a mass flux of $G = 0.026$ g/(cm^2 $\cdot$ s), and an evaporation coefficient $\gamma_E = 0.24$. The MFPT measured in this series of simulations therefore corresponds to an evaporation rate within the (broad) range of measured values.

4.4 Identification of most important order parameters in evaporation

4.4.1 Principal component analysis

The objective of this analysis was to determine what order Parameter(s) varied the most over the critical part of the MFEP. Because evaporation is *not* an activated process with a transition state, we examined images 10 through 13 (where the entire string comprised images 0–16). This region of the string accounted for half the free energy change and about half the change in MFPT values.

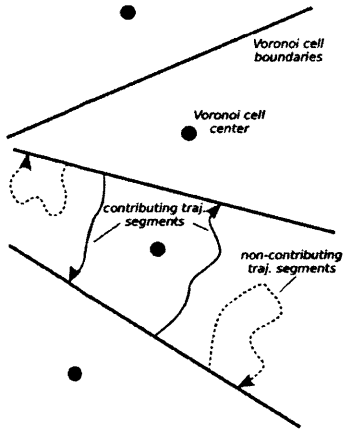


Figure 4-11. Schematic showing forward (reactant-to-product) and backward (product-to-reactant) contributing trajectories (solid curves) and non-contributing trajectories (dashed curves).

The trajectories analyzed were "contributing trajectories," as shown in Figure 4-11. These contributing trajectories are defined as those that contribute to the forward or backward reaction rate in the milestoning scheme, by, for example, starting at one milestone in a Voronoi cell, and reaching the opposite milestone before intersecting the original milestone again. The label "forward" indicates the direction from reactant to product along the string or within an image, *i.e.* in the direction of evaporation, from the liquid to vapor state, while the label "backward" indicates the reverse direction.

For example, in a 2.0-ns simulation of image 10, there were 47 forward contributing and 47 backward contributing trajectory segments observed, with average length

1.6 ps for both. During these simulations, order parameter values were recorded every 5 fs, to provide greater resolution in time; the simulation code also printed OP values whenever a system entered or left its home cell. The union of all these contributing trajectories from the four simulation cells was analyzed with principal component analysis (PCA).

To normalize differently-scaled order parameters, each OP was scaled by $\frac{1}{\sqrt{k_i}}$, as in the simulations themselves, after subtracting OPs' mean values from all recorded points. This scaling approach was used to reflect the original dynamics used to create the trajectory points.

The points from all forward contributing trajectories, after being projected onto the first three principle components, are shown in Figure 4-12. This shows that the first principle component (PC) is aligned along the length of the string. In addition, the string (represented by the image centers, which serve as Voronoi support points) lies in the middle of the "tube" of reactive trajectories, as would be expected under the SMCV methodology. The eigenvalues from PCA, which represent (after normalization so that their sum is unity) the amount of variance captured by each principal component are shown in Figure 4-13(a) (page 76).

Figure 4-13(b) shows projections of the first two principal components onto the original order parameters. The first principal component is aligned most closely with OPs q_0^z, q_6^{don}, and q_7^{acc}. Order parameter q_2^{ravg} is directly nearly parallel with PC2, although PC2 explains only about one third as much variation in the trajectory points' OP values. All order parameters' projections along PC1 and PC2 are listed in Table 4-2 (page 76), which shows that similar results were obtained by examining backward trajectories.

This analysis suggests that the order parameters can be divided into a "first tier" of importance, containing the z-position and the hydrogen bond counts. However, the case of OP q_2^{ravg} is less clear: PC2 is, by construction, orthogonal to PC1. Because all reactive trajectories were aggregated together, it is not clear whether the presence of q_2^{ravg} as the main component of PC2 reflects its importance. That is, q_2^{ravg} could appear in PC2 because (1) its value changes over the course of the reactive trajectories,

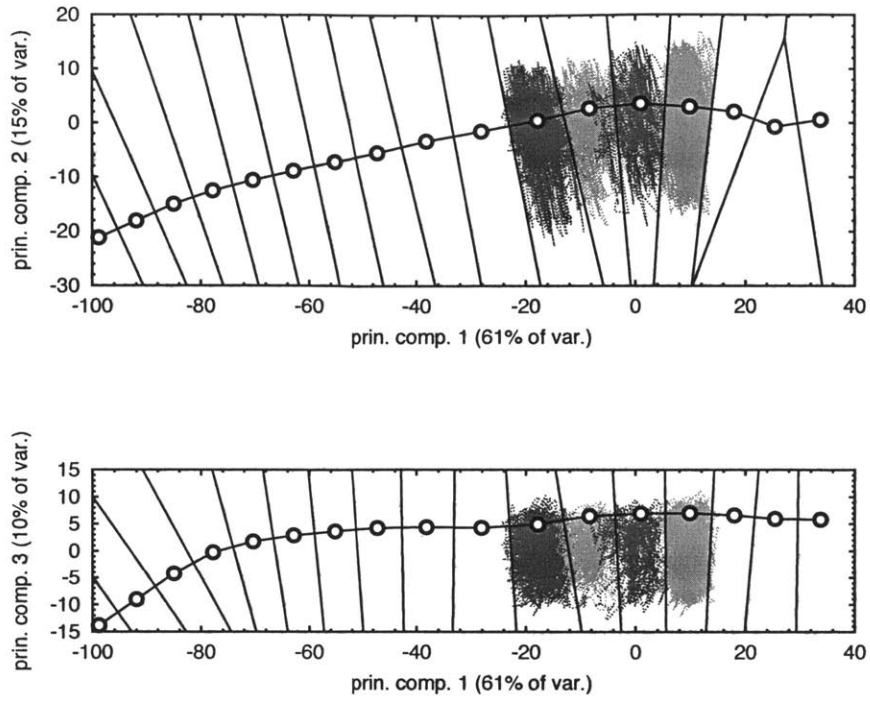


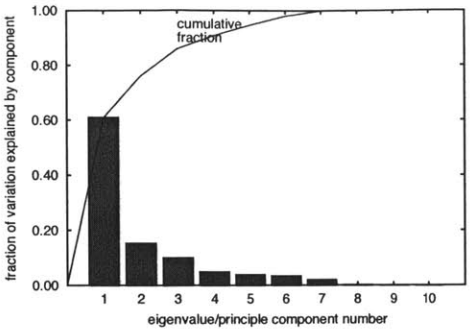
Figure 4-12. Projection of contributing trajectory segments in the forward (evaporating) direction onto principle components. Image centers (Voronoi support points) are the black points, while the trajectories from images 10–13 are shown in alternating shades of gray. The rightmost point represents the final, vapor-phase image.

which themselves cover a significant region of FE/MFPT changes, or (2) the order parameter does not change much along the trajectories, and the reaction tube is a collection of many parallel trajectories with many different values of q_2^{ravg}, unchanging along each trajectory. Possibility (1) would suggest that q_2^{ravg} is important, while possibility (2) would suggest that it is not.

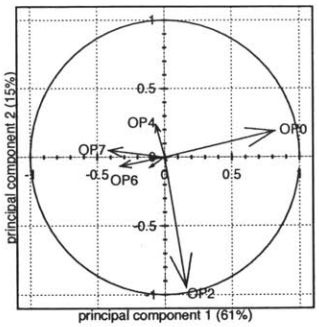
To address this question, I applied two other analyses, which are described in the next two subsections.

4.4.2 Directional analysis

To disaggregate the collections of OP values in many trajectory segments, this analysis focused on the trajectories one at a time. Initially, I attempted to apply PCA to each individual trajectory, but because of their typically non-linear behavior in



(a) Amount of variance explained by principal components.



(b) Contributions of each order parameter to principal components 1 and 2.

Figure 4-13. Summary of PCA results.

Table 4-2. Order parameter components of first and second principle components in analysis of contributing trajectories in Images 10–13. The three largest components in PC1 and the two largest in PC2 are highlighted.

	forward (evaporating) dir.		backward (opposite) dir.	
OP	PC1	PC2	PC1	PC2
0	**0.81**	**0.20**	**0.82**	**0.19**
1	−0.11	−0.0084	−0.11	−0.0098
2	0.18	**−0.96**	0.18	**−0.97**
3	−0.11	−0.068	−0.11	−0.045
4	−0.065	0.16	−0.045	0.12
5	0.015	−0.014	−0.018	−0.0058
6	**−0.33**	−0.068	**−0.33**	−0.04
7	**−0.41**	0.017	**−0.40**	−0.027
8	−0.0015	0.0004	−0.0015	0.0001
9	0.038	0.0047	0.037	0.0024
$\left(\lambda_i / \sum_j \lambda_j\right)^a$	0.61	0.15	0.61	0.15

[a]Percent of data variance explained by this component.

the 10-dimensional space, projections of the trajectory points onto their principal components often appeared unsatisfactory.

Instead, to understand the nature of reactive trajectory segments, and what order parameters were changing in this most interesting region of the string, we looked at the vector $\Delta\mathbf{q}'$ for each of the forward- or backward-contributing segments. This

vector is the overall direction from the point in OP space where the trajectory *enters* the Voronoi cell, to the point where the trajectory *leaves* the cell:

$$\Delta \mathbf{q}' = \mathbf{q}'_{f,segment} - \mathbf{q}'_{0,segment}$$

The prime mark (') indicates that OP scaling was applied, in the same manner discussed above.

These directions were then normalized, and the mean direction in each cell was calculated, using techniques for directional variables.[136] These directions are listed in Table 4-3, which also shows that these vectors were closely grouped around the mean in each image, as the "circular variance" listed in the last column of Table 4-3 was typically ~ 0.2.

Table 4-3. Results of local direction analysis for forward-directed transitions in Images 8–14. The two largest components of the mean vector in each image are highlighted.

					— components of normalized mean direction $\hat{\mu}$ along OPs —											
img.	N_{trans}	q_6^{don}	q_7^{acc}	F^a	0	1	2	3	4	5	6	7	8	9	$\delta(\hat{\mu})^b$	$(1-R)^c$
8	74	2.0	2.1	-0.1	0.50	-0.18	-0.16	-0.13	-0.10	-0.01	**-0.60**	**-0.55**	0.00	0.00	(0.99)	0.216
9	76	1.6	1.9	0.3	0.42	-0.18	-0.05	-0.32	0.03	-0.04	**-0.61**	**-0.55**	0.00	-0.00	(0.99)	0.202
10	47	1.1	1.6	1.5	**0.49**	-0.16	-0.08	-0.31	0.15	-0.12	**-0.63**	-0.44	0.00	-0.01	(0.98)	0.209
11	34	0.7	1.1	3.7	**0.64**	-0.13	0.37	0.04	0.12	-0.18	**-0.48**	-0.39	0.00	0.01	(0.96)	0.374
12	32	0.1	0.9	5.7	**0.57**	-0.05	**0.77**	0.09	-0.08	0.12	-0.09	-0.22	-0.00	-0.00	(0.97)	0.312
13	213	0.0	0.5	7.0	**0.47**	-0.00	**0.84**	0.02	-0.28	0.02	0.00	-0.01	-0.00	-0.00	(0.997)	0.254
14	438	0.0	0.1	7.0	0.11	-0.00	**0.94**	0.04	**-0.26**	0.01	0.00	-0.00	-0.00	0.00	(0.999)	0.126

[a] Free energy, in kcal/mol.
[b] 95% confidence interval or "cap" for mean; in image 8, for example, the 95% confidence interval is given by $(\mu_{sample} \cdot \hat{\mu}) > 0.99$.
[c] Measure of variance around mean direction, which takes values between 0 (no variance) and 1 (random distribution).

This shows that the OPs which changed most during the forward trajectories were different in each image: initially, the molecule loses its first hydrogen bond (images 8 and 9), and the trajectories are directed along the two H-bond OPs; next, the molecule continues moving outward, and drops it remaining donated bond (images 10 and 11); next, the alignment of nearby molecules undergoes a shift, as the value of q_2^{r-avg} increases, indicating *decreasing* alignment with neighbors. Examining the OP values, the average dipole-dipole angle η increases from about 65° to 100° (images 11 through 13). Once again, by the time this point along the string is reached, less than one H-bond is being accepted, and the molecule is ready to evaporate.

While the underlying data examined in this trajectory direction analysis and the PCA approach described above, they appear to paint a consistent picture, in which the z-position, average relative orientation, and the hydrogen bond numbers are the most important order parameters.

4.4.3 Examining MFPT as a function of order parameters

A common goal in characterizing reactive systems with collective variables is to identify how the reaction committor probability, p_B (sometimes written p_{fold} in the protein simulation literature) can be related to those collective variables. The Voronoi boundaries between images points along the reaction path (string) identified through SMCV,[2] and which serve as ideal milestones for measuring kinetic properties, are isocommitor surfaces, at least in the local neighborhood of the string.[129,130]

After performing milestoning calculations, the mean first-passage time to the final milestone—in this case, the evaporated state—can be examined as a function of collective variables, as the MFPT is monotonically related to the committor probability p_B. While the analyses above described how the evaporating water molecule's state changed during evaporation events, it did not include information from the MFPT,

[2]In Ref. 129, the commitor function, denoted $q(\mathbf{x})$, is a function of system coordinates, not collective variables, and has a single value at each point $\mathbf{x}$ which is conceptually measurable. The committor function at a particular value of *collective variables* has a range of values in a distribution, relating to the set of microstates that exhibit those particular CV values. For purposes of discussion in this subsection, we will consider the mean value of the committor probability p_B at collective variable values.

which reflects a water molecule's likelihood of evaporation at its different collective variable states along the reaction path.

To understand the relationship between MFPT and the order parameters, the MFPT values at each milestone were used for linear regression. The order parameters at each milestone were calculated as the midpoint between the image centers (support points) on either side of the milestone. (While an ideal approach might be to use the point of maximal hitting point density[129] on the milestone for the $\{q_i\}$ values at each milestone, Figure 4-12 shows that the string and its constiutent image points are within the main reaction channel identified.)

The order parameter values were then centered and scaled by their standard deviations, and the MFPT was transformed by taking $\tau' = 1 - \tau/\tau_{bulk}$, where τ_{bulk} is the MFPT value in the at the milestone farthest from the evaporated state. Subtracting this ratio from one simply allowed tau' to increase from 0.0 (bulk state) to 1.0 (evaporated state), so that evaporation is in a "positive" direction, consistent with the rest of this paper. All milestone points were equally weighted, although the final four points, where the MFPT changed the most, typically had a relatively large influence on regressions, with values of Cook's distance[137,138] of approximately 1.

As in any multivariate regression, identifying a model requires a compromise between model simplicity (parsimony) and goodness of fit. We identified the two best-fitting combinations of $1, 2, \ldots, 10$ order parameters using an exhaustive search; these are listed in Table 4-4, where they are sorted by the values of the Bayes information criterion (BIC). The coefficients on each OP for the top five models are shown in Figure 4-14.

In Table 4-4, the two most frequently appearing variables are the average relative orientation q_2^{ravg} and the number of hydrogen bonds accepted q_7^{acc}, and model B, the second-best model, contains these two order parameters, along with q_0^z and q_9^{t-ang}. This is consistent with the results obtained above, while it should be noted that this analysis, instead of looking at local, trajectory-based data, identifies these order parameters using the MFEP points themselves, along with the quantitative values of the MFPT.

Table 4-4. Best models of $\tau' = (1 - \tau/\tau_{bulk})$ with different numbers of order parameters used. The combinations are sorted by BIC value.

OPs	order parameters in linear model	BIC	designation[a]
1	q_2^{ravg}	−71.5	model A
4	$q_0^z + q_2^{ravg} + q_6^{acc} + q_9^{tang}$	−70.5	model B
7	$q_1^{lden} + q_3^{rstd} + q_4^{thet} + q_6^{don} + q_6^{acc} + q_8^{tdist} + q_9^{tang}$	−69.8	model C
6	$q_1^{lden} + q_2^{ravg} + q_3^{rstd} + q_6^{don} + q_6^{acc} + q_8^{tdist}$	−69.5	model D
2	$q_2^{ravg} + q_9^{tang}$	−69.2	model E
2	$q_2^{ravg} + q_5^{omeg}$	−69.1	
7	$q_1^{lden} + q_2^{ravg} + q_3^{rstd} + q_4^{thet} + q_6^{don} + q_6^{acc} + q_8^{tdist}$	−68.9	
8	$q_0^z + q_1^{lden} + q_3^{rstd} + q_4^{thet} + q_5^{omeg} + q_6^{don} + q_6^{acc} + q_8^{tdist}$	−68.9	
4	$q_2^{ravg} + q_4^{thet} + q_6^{acc} + q_9^{tang}$	−68.8	
3	$q_2^{ravg} + q_5^{omeg} + q_9^{tang}$	−68.7	
6	$q_1^{lden} + q_3^{rstd} + q_4^{thet} + q_6^{don} + q_6^{acc} + q_8^{tdist}$	−68.7	
8	$q_1^{lden} + q_3^{rstd} + q_4^{thet} + q_5^{omeg} + q_6^{don} + q_6^{acc} + q_8^{tdist} + q_9^{tang}$	−68.2	
5	$q_1^{lden} + q_2^{ravg} + q_3^{rstd} + q_6^{acc} + q_8^{tdist}$	−68.0	
3	$q_2^{ravg} + q_6^{acc} + q_9^{tang}$	−68.0	
5	$q_0^z + q_2^{ravg} + q_4^{thet} + q_6^{acc} + q_9^{tang}$	−67.9	
9	$q_0^z + q_1^{lden} + q_3^{rstd} + q_4^{thet} + q_5^{omeg} + q_6^{don} + q_6^{acc} + q_8^{tdist} + q_9^{tang}$	−66.5	
9	$q_0^z + q_1^{lden} + q_2^{ravg} + q_3^{rstd} + q_4^{thet} + q_5^{omeg} + q_6^{don} + q_6^{acc} + q_8^{tdist}$	−66.1	
10	$q_0^z + q_1^{lden} + q_2^{ravg} + q_3^{rstd} + q_4^{thet} + q_5^{omeg} + q_6^{don} + q_6^{acc} + q_8^{tdist} + q_9^{tang}$	−63.9	all OPs

[a] These names are used in a comparison of the models' coefficients in Figure 4-14.

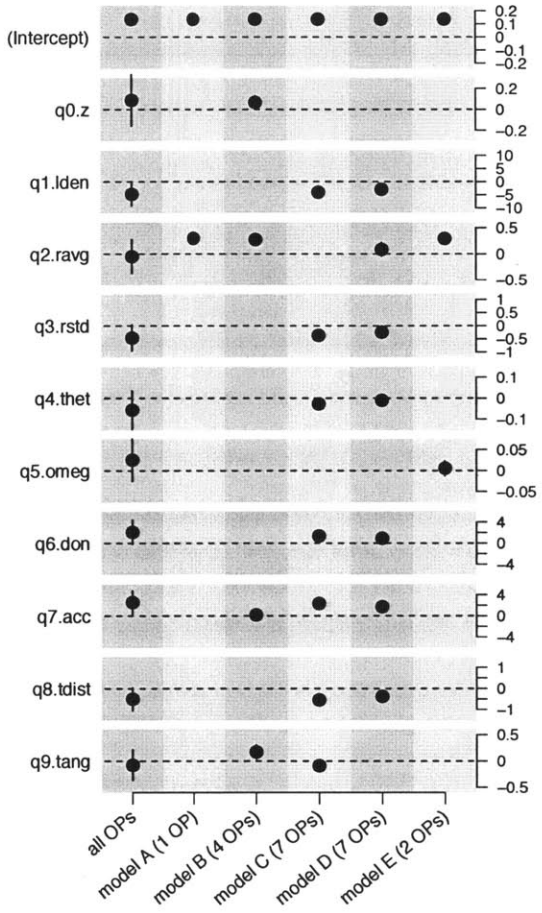


Figure 4-14. Coefficients of order parameters for the five models A–E with best BIC values, and the linear model containing all order parameters. The coefficients are for the normalized order parameters, and the error bars are 95% confidence intervals.

The purpose of this regression analysis is not necessarily to construct a quantitative model for the MFPT profile, but rather to identify order parameters that may be important in determining the value of the mean first-passage time, which is related to p_B. Nonetheless, the coefficients for the top five models listed in Table 4-4, as well as the model containing all 10 OP terms, are given in Figure 4-14, and Figure 4-15 shows that even a simple linear model in four terms can reasonably reproduce the shape of the MFPT profile.

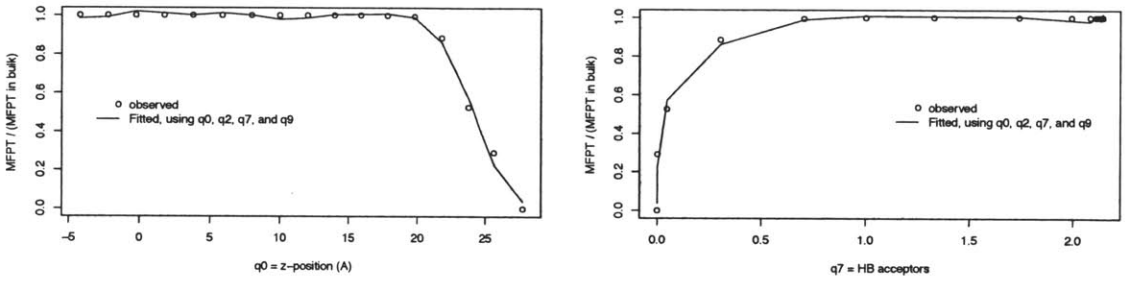


Figure 4-15. Observed and fitted values of the MFPT values at milestones, plotted against two order parameters. The fitted data were from "model B" of Table 4-4 and Figure 4-14.

4.5 Features of soluble additives suggested by this work

4.5.1 Implications for additive design

Based on the mechanistic understanding of the evaporation process described above, I believe that a successful additive to impede evaporation would possess the following features. Note that the following features are based on the reasoning that the inhibiting additive would target the existing, natural kinetic bottleneck in the evaporation process, in order to induce an energetic barrier there.

i. The additive should adsorb at the liquid-vapor interface, *i.e.* be surface active, in order to impede evaporation there.

ii. The additive would exhibit a strong propensity to form a "second" hydrogen bond with water, *i.e.* to *donate* a hydrogen bond to a water molecule which has only one "natural" (accepted) H-bond from the liquid phase.

iii. The feature above may require that the additive's donor group exhibit a certain orientation relative to the interface, such as facing generally outward.

iv. The hydrogen-bond donating feature would also ideally be placed well into the outer half of the interfacial region, beyond the Gibbs dividing surface, since most of the evaporating molecule's free energy change (and passage time) takes place there.

v. In designing a soluble additive, the hope is that it will be possible to "tune" surface activity separately from the particular evaporation-inhibiting features suggested above, by adjusting the number or degree of hydrophobicity of those functional groups which do not participate in the interactions with interfacial water, but instead lead to the surface active or amphiphilic nature the additive. by adjusting its amphiphilic nature, or making other similar adjustments to the additive's design.

Chapter 5

Molecular evolution using automated evalaution of molecular designs and a genetic algorithm

5.1 Overview of screening and evolution approach

In general, the key steps to employ an evolutionary approach for a screening or optimization task are (*i*) to define a genomic representation of objects in the problem domain; (*ii*) to formulate an objective function to evaluate those objects; and (*iii*) to implement reproductive steps (*e.g.* mutation and crossover) to generate new objects from parents' genomes.

In this problem, we seek to optimize the design of organic ligands which could be chemically attached in a close-packed manner to a solid surface of silica or gold.[87] To optimize such a material, we have chosen to focus on the chemical architecture of the attached organic ligand. Since quasi-linear ligands are well-suited for self-assembly on such surfaces, we have represented such molecules as chains of functional groups, from the enumerated sets listed in Table 5-1. For example, a ligand with structure H−CH(CH_3)−CH(OH)−CH_2−OH would be represented by the genome `0 1015 1012 1000 2`. By convention, the end of the molecule attached to the solid

surface is the first group listed.

Table 5-1. Terminal and intermediate functional groups used in design of linear ligands.

— terminal groups —			— intermediate groups —		
codon	name	structure	codon	name	structure
0	hydrogen	$-H$	1000	methylene	$-CH_2-$
1	methyl	$-CH_3$	1001	ether	$-O-$
2	hydroxyl	$-OH$	1002	carbonyl	$-(CO)-$
3	aldehyde	$-CHO$	1003	ester	$-COO-$
4	carboxyl	$-COOH$	1004	secondary amino	$-NH-$
5	primary amino	$-NH_2$	1005	*o*-didehydrobenzene	$-$(o)Ph$-$
6	phenyl	$-Ph$	1006	*m*-didehydrobenzene	$-$(m)Ph$-$
7	vinyl	$-CH=CH_2$	1007	*p*-didehydrobenzene	$-$(p)Ph$-$
8	acetylenyl	$-C\equiv CH$	1008	*cis*-ethylene-1,2-diyl	$-$(cis)CH=CH$-$
9	allenyl	$-CH=C=CH_2$	1009	*trans*-ethylene-1,2-diyl	$-$(trans)CH=CH$-$
10	isopropyl	$-CH(CH_3)_2$	1010	acetylene-1,2-diyl	$-C\equiv C-$
11	terbutyl	$-C(CH_3)_3$	1011	allene-1,3-diyl	$-CH=C=CH-$
12	amide	$-CONH_2$	1012	methanol-1,1-diyl	$-CHOH-$
13	thiol	$-SH$	1013	thioether	$-SH-$
14	fluoride	$-F$	1014	isopropyl-methylene	$-$CH(iPr)$-$
15	chloride	$-Cl$	1015	methyl-methylene	$-CH(CH_3)-$
16	bromide	$-Br$	1016	ethyl-methylene	$-CH(CH_2CH_3)-$
			1017	dimethyl-methylene	$-C(CH_3)_2-$
			1018	phenyl-methylene	$-CHPh-$
			1019	carboxyl-methylene	$-CHCOOH-$
			1020	amine-methylene	$-CHNH_2-$
			1021	1,5-didehydronapthalene	$-C_{10}H_6-$
			1022	2,6-didehydronapthalene	$-C_{10}H_6-$

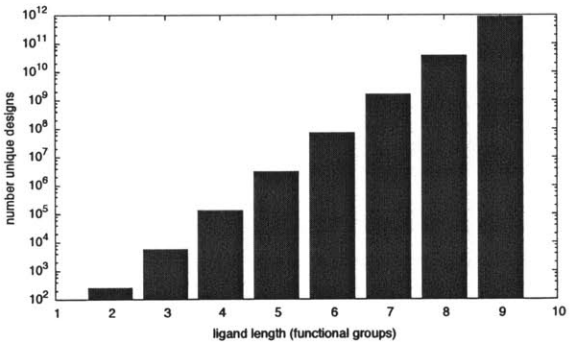


Figure 5-1. Number of ligand designs that can be created with functional groups used in this study, neglecting functional group combinations that are prohibited (see Table 5-2 in Supplementary Information).

In describing molecules in this way, it is useful to designate certain combinations of functional groups as *forbidden.* For example, if two ether groups were placed adjacent to one another in a linear molecule, the result would be a chemically unstable peroxide

group. The forbidden combinations used are listed in Table 5-2, and also include combinations that would lead to incorrect atomtype designations under the GAFF force field (see below). In practice, entries could be added to this list for other reasons, such as to prevent the exploration of part of the chemical space which is already well-characterized, or which is unattractive due to commercial/IP restrictions.

While it is not strictly necessary to designate GAFF atom types within each functional group, we found that doing so—and obviating the need for atomtype perception by ANTECHAMBER—increased the robustness of the simulation setup.

Table 5-2. Forbidden functional group combinations.

gene codes		chemical groups	notes
2	2	HO–OH	peroxide
1001	2	–O–OH	peroxide
2	1001	HO–O–	peroxide
1001	1001	–O–O–	peroxide
1003	2	–COO–OH	peroxide
1003	1001	–COO–O–	peroxide
5	5	$H_2N–NH_2$	hydrazine
1004	5	–NH–NH_2	hydrazine-1-yl
5	1004	H_2N–NH–	hydrazine-1-yl
2	1013	HO–SH	thio-peroxide anologue
1001	13	–O–SH	thio-peroxide anologue
1003	1013	–COO–SH	thio-peroxide anologue
1001	1013	–O–S–	thio-peroxide anologue
1013	1001	–S–O–	thio-peroxide anologue
1001	0	–O–H	would create hydroxyl group with incorrect GAFF atomtypes
1003	0	–COO–H	would create carboxyl group with incorrect GAFF atomtypes

To evaluate potential ligand designs, we have employed molecular dynamics simulations. As noted in Table 1-1, other studies employing similar techniques have used as objective functions properties calculated from molecules' 2D or 3D structures. This approach has the advantage of speed and ease of calculation, but also presupposes a particular solution to the molecular design problem, such as similarity to a given ligand or satisfaction of certain property criteria.

Molecular dynamics simulations (or electronic structure calculations, in other possible applications), in contrast, make no *a priori* assumptions about the mecha-

nism(s) or molecular features that would lead to desired performance. The challenge in using MD-based evaluation, however, is that the steps preparatory to running a simulation—creating topology files, identifying force field parameters, and finding reasonable initial structures—are often performed "by hand" by practitioners, and are not trivially automated. The approach we have taken is to build each candidate molecule's 3D structure from its constituent fragments using custom software named FORM2GEOM, and then to employ the GAFF force field,[80] as described in greater detail below.

Finally, in order to generate new designs, genetic operators were implemented, in order to explore the chemical space in a broad yet efficient manner. The genetic operations employed are gene deletion, gene addition, gene mutation, and two-parent crossover exchange. These operations are described in greater detail below. An overview of the iterative process of molecular evolution is shown in Figure 5-2.

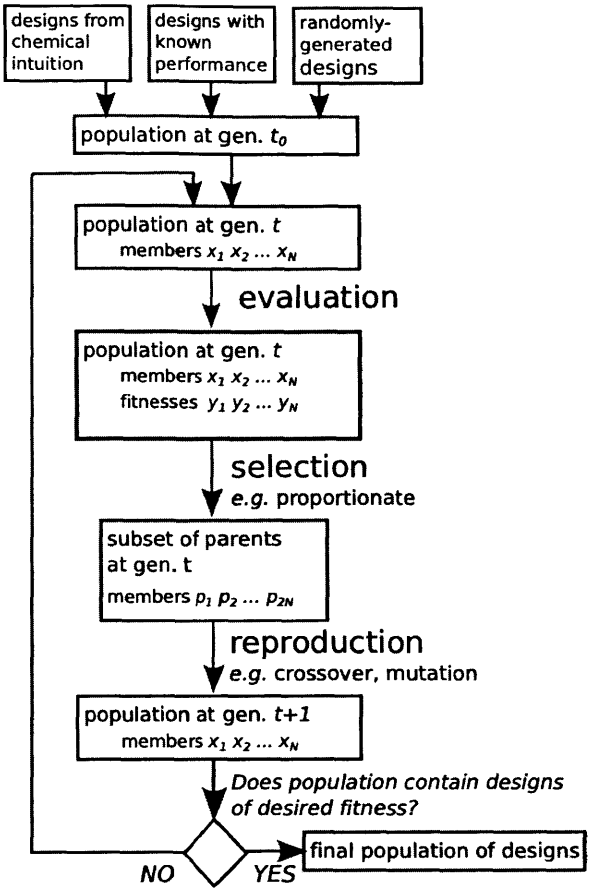


Figure 5-2. Schematic of iterative evaluation/evolution process. The $\{x_i\}$ and $\{p_i\}$ represent sets of genomes, while $\{y_i\}$ are sets of fitness values.

5.2 Evaluation of ligand candidates through molecular dynamics simulation

Formulating an objective function requires an understanding of how the adsorption medium would be used in a chemical process for the manufacturing of a pharmaceutical API. As stated in the previous section, the goal here is to design a system that could separate the unreacted intermediate E2 from a solution, while allowing the desired intermediate E6 to remain in solution.

Thus the ideal objective function would provide information about the selectivity of the ligand system for adsorption of E2 as against adsorption of E6, at liquid-phase concentrations equal to process conditions.

$$\begin{pmatrix}\text{species } j\\ \text{in solution}\end{pmatrix} \rightleftharpoons \begin{pmatrix}\text{species } j\\ \text{adsorbed on surface}\end{pmatrix}$$

$$K_{ads,j} = \frac{a_{j,ads}}{a_{j,sol}} = \frac{\phi_j y_j}{\gamma_j x_j}$$

where $a_{j,m}$ activity of species j in phase m

x_j, y_j mole fraction of species j in solution phase, adsorbed phase

γ_j, ϕ_j activity coefficient of species j in solution, adsorbed phases

$$\alpha_{\mathrm{E2,E6}} = \left(\frac{\theta_{\mathrm{E2}}}{x_{\mathrm{E2}}}\right)\left(\frac{\theta_{\mathrm{E6}}}{x_{\mathrm{E6}}}\right)^{-1} = K_{ads,\mathrm{E2}} K_{ads,\mathrm{E6}}^{-1} = \frac{\exp\left(-\frac{\Delta G_{ads,\mathrm{E2}}}{RT}\right)}{\exp\left(-\frac{\Delta G_{ads,\mathrm{E6}}}{RT}\right)}$$

$$\text{maximize} \qquad \Delta\left(\Delta G_{ads}\right)_S = -\Delta G_{ads,\mathrm{E2}} - \left(-\Delta G_{ads,\mathrm{E6}}\right)$$

This $\Delta\Delta G$ for a particular ligand represents the best possible objective function we could use in our simulation-based evaluations. The problem is that free energy measurements are computationally expensive. So, the objective functions used in the molecular dynamics-based evaluation was the energetic component of this quantity:

$$\text{maximize} \qquad \Delta\left(\Delta E_{ads}\right)_S = -\Delta E_{ads,\mathrm{E2}} - \left(-\Delta E_{ads,\mathrm{E6}}\right)$$

The energy, rather than the enthalpy, was measured, because in simulating a two-dimensional system such as this one by extending vacuum in the z-direction, it is not possible to impose a specified pressure, and thus it is not feasible to measure enthalpy.

To evaluate ligand designs, the ideal objective function would be the free-energetic selectivity, which is related to the logarithm of the selectivity factor:

$$\text{maximize} \quad \Delta\left(\Delta G_{ads}\right)_S = -\Delta G_{ads,\mathrm{E2}} - \left(-\Delta G_{ads,\mathrm{E6}}\right)$$

But free energy differences are expensive to measure computationally, so we have

employed an energetic-only fitness score:

$$\text{maximize} \quad \Delta\left(\Delta E_{ads}\right)_S = -\Delta E_{ads,\mathrm{E2}} - \left(-\Delta E_{ads,\mathrm{E6}}\right)$$

The adsorption energy of each species is calculated from the three simulations: a simulation of the surface-bound ligand alone, of the surface-bound ligand with E2 adsorbed, and of the surface-bound ligand with E6 adsorbed.

$$\Delta E_{ads,\mathrm{E2}} = \langle E_{\mathrm{lig+E2}} \rangle - \langle E_{\mathrm{lig}} \rangle - \langle E_{\mathrm{E2}} \rangle$$
$$\Delta E_{ads,\mathrm{E6}} = \langle E_{\mathrm{lig+E6}} \rangle - \langle E_{\mathrm{lig}} \rangle - \langle E_{\mathrm{E6}} \rangle$$

where $\langle \cdots \rangle$ indicates ensemble averaging. $\langle E_{\mathrm{E2}} \rangle$ and $\langle E_{\mathrm{E6}} \rangle$ are the average energies of E2-only and E6-only simulations, which were performed one time. In separate quantum calculations,[139] the gas-phase dipole moments of the E2 and E6 molecules were measured as 2.6 D and 1.7 D (each averaged from two different configurations), respectively, indidcating there is not a significant difference in polarity.

In addition to the selectivity function, a quadratic penalty function is applied to overly long (greater than 7 functional groups) linear ligand designs; their sum was the fitness score.

As noted above, each potential ligand candidate in our scheme is a linear arrangement of functional groups, represented by a string of integers listed in Table 5-1. The FORM2GEOM software first translates such a string into a 3-dimensional structure. The software contains a library of functional groups' 3D structures, excerpted from molecules in the NIST Standard Reference Database Number 69.[140] Each functional group fragment contains one (terminal groups) or two (intermediate groups) "bond vectors," which extend from designated atoms along the axis of a chemical bond to preceding/succeeding functional groups.

5.3 Molecular structure and simulation setup

To construct the molecule's 3D structure, the software first places the initial functional group fragment, then aligns the bond axis of the second functional group fragment parallel to that of the first, and places their two "bonding atoms" an appropriate distance apart. If the fragment has an important rotational degree of freedom in the dihedral angle about the bond (*e.g.* for the fragment –CH(iPr)–CH(CH_3)–), the functional group is rotated until a target dihedral value is met. This process is repeated for subsequent functional group fragments until the molecule is complete.

At that point, the molecule is subjected to energy minimization, using a simplified force field in which atoms experience a Lennard-Jones interaction, and in which each linked fragment has a direction and associated dihedral energies. The purpose of this minimization step is to eliminate any close overlaps of atoms that would render simulations with the full molecular force field unstable. More specifically, after the constituent fragments are joined together using the information about connecting atoms and the "bond vectors" along which chemical bonds are created, the molecule's energy is minimized using an *ad hoc* force field. In this force field, each atom constitutes a Lennard-Jones center, and in addition linked functional groups are described with their orientations. These are used to calculate the second term in the force field, the dihedral energy, using the dihedral angle between two adjacent fragments' orientation, calculated according to a standard form:

$$E_{\text{dihed}} = \frac{V_d}{2I_{div}} \left(1 + \cos\left(n * \phi - \phi_0\right)\right)$$

where V_d is a force constant; I_{div} is the number of dihedrals quartets that can be formed using the two central atoms; ϕ is the dihedral angle; n is an integer reflecting the rotational symmetry of the dihedral; and ϕ_0 is the phase.

After a reasonable 3D structure for the molecule is obtained from the FORM2GEOM program, a topology file is prepared using the ANTECHAMBER suite[81] and the GAFF force field.[80] Because atoms in the FORM2GEOM fragment library are already described by their GAFF atomtype, and bond types are likewise pre-specified, no atom

or bond-type perception needs to be carried out in this step, although in other work, these capabilities of ANTECHAMBER could be used. Partial charges are estimated using the AM1-BCC semi-empirical technique[141,142] within ANTECHAMBER.

Two arrangements of the ligand candidate molecule could be used to simulate ligands bound to a solid surface. A single bound ligand molecule could be used to evaluate interactions with the E2 and (separately) E6 molecules. We also developed multi-ligand simulations, in which a two-dimensional array of ligands was generated, to reflect arrangement in a self-assembled monolayer (described in detail in the Supplemental Information). The results in this study were obtained using the single-ligand procedure.

To begin that evaluation procedure, the ligand was rotated so that the vector separating its first and last atoms (by index) was parallel to the z-axis. Then, its initial atom (with lowest index) was fixed in place for the later molecular dynamics simulations, to represent the ligand's attachment to a planar solid surface (*e.g.* a gold surface, with attachment through thiol chemistry), or to a fixed point in a sol-gel polymer network. A quadratic half-well potential was imposed, with its minimum at a position z_{wall} equal to the z-coordinate of the fixed atom and a force constant of $k_{wall} = 1.0$ kcal/mol (with a 1/2 prefactor), as depicted in Figure 5-3.

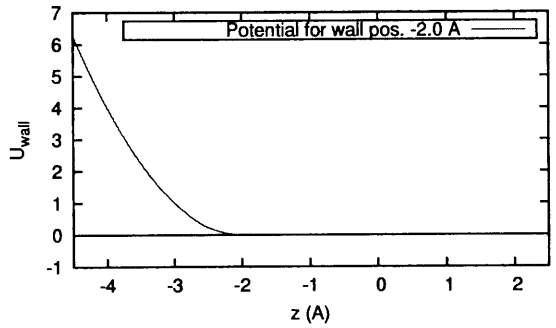


Figure 5-3. Illustration of half-well potential representing the solid surface to which ligands are attached. In this example, $z_{wall} = -2.0$ Å.

In all cases, the initial geometries of single ligands produced initially by FORM2GEOM were suitable to begin molecular dynamics simulations. The ligand system was subjected to 10,000 steps of minimization, using NAMD's[116] congujate gradient minimizer.

Next, simulated annealing was used, to allow the ligand to escape from a local minimum in configuration space, and to increase its probability of starting production in a relatively low-energy conformation. This step consisted of running 50 ps of MD at 450, 600, 450, and then 300 K. Langevin temperature control[116] was used with a damping coefficient of 50 ps^{-1}. The ethyl acetate solvent was represented by setting the dielectric constant to its experimental value of 6.0.[143]

A dielectric medium (vacuum) was used, rather than explicit solvent molecules, to speed up the computational evaluation of ligand designs; in an ideal evaluation, the system would include a layer of ligands with a layer of explicitly-modeled solvent over it, to capture solvent effects. In future work, refined simulations could be used as an objective function in a second optimization phase, after a population of candidates is obtained from less accurate, but faster, simulations. Additionally, in other applications involving adsorption, the solvent could be separately chosen or optimized after a surface/medium is designed.

Next, the ligand system was equilibrated for 250 ps in the canonical ensemble (at 300 K) for single-ligand simulations; the limited duration of equilibration, in relation to the production time, was deemed suitable because of the limited number of degrees of freedom of the translationally-restricted single ligand molecule. Next, production MD was carried out for either 3.0, 4.5 or 6.0 ns at 300 K, as listed in Table 5-3. Depending on system size, these simulations took 3 to 6 hours on a single CPU core, so that each generation could be evaluated in about one day on an 8-core computer.

The final structure of the surface-bound ligand in its production run was used to begin the E2–ligand and E6–ligand simulations. The E2 or E6 molecule was placed so that its minimal z-coordinate is 1.5 Å from the maximal z-coordinate of the ligand or ligand layer, to establish the initial configuration for the each of these simulations, which were performed for the same amount of equilibration/production time as the ligand-only simulations. Statistical standard errors were calculated using a customary approach.[144]

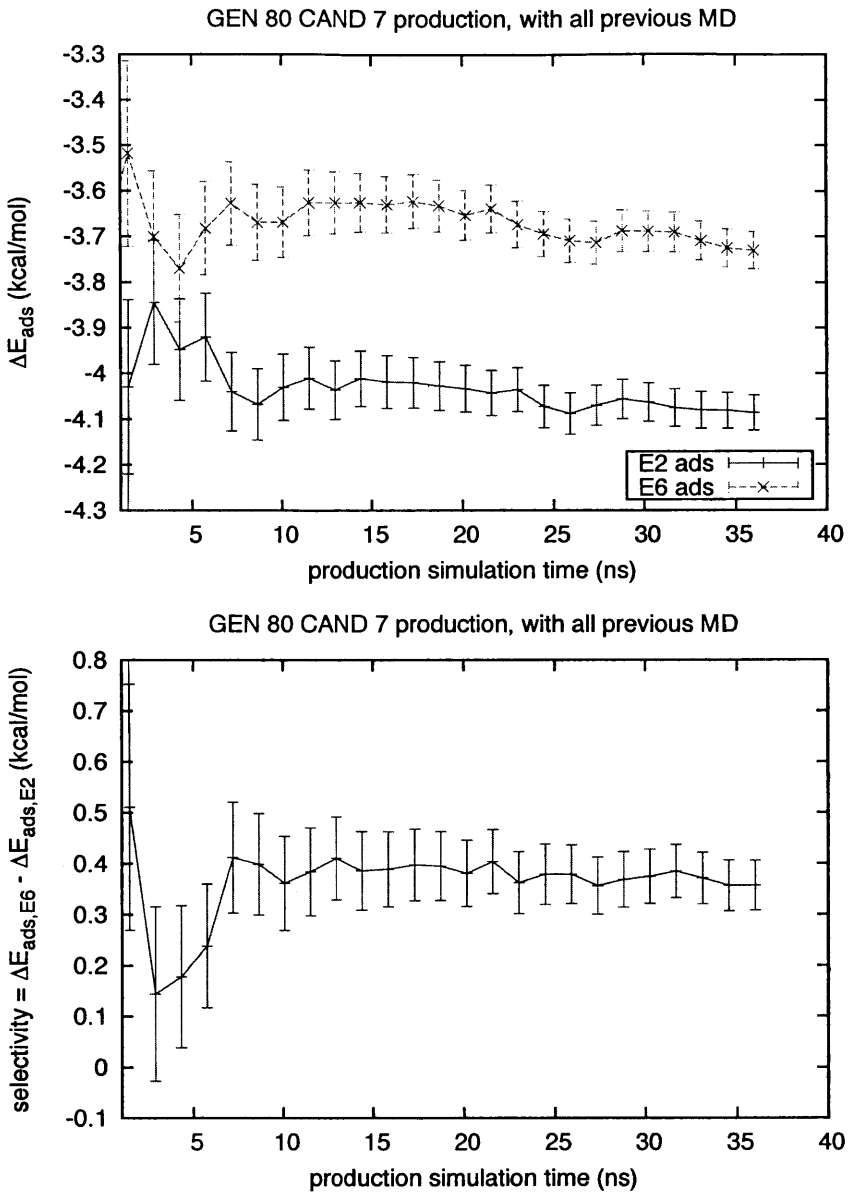


Figure 5-4. Average adsorption energy values (*top*) and selectivity score (*bottom*) over repeated evaluations of a popular ligand design in Experiment III: these data are taken from evaluation with a ligand having structure $H_2NC(=O)C(CH_3)_2C(=O)H$ in generation 80 of that experiment.

5.4 Molecular evolution procedure

As depicted in Figure 5-2, the key steps in genetic optimization are (*i*) evaluation of each member of a population; (*ii*) selection of a set of parents from the population as a whole, based on members' fitness scores; and (*iii*) the establishment of the

subsequent generation's member sequences based on parents' genomes. Step *i* was described above, and steps *ii* and *iii* will be discussed below.

Several techniques have been developed for selecting members of the parental subset from among the whole population of evaluated members; the optimal selection technique for a given problem has been shown to depend on the underlying fitness landscape and the accuracy of fitness function evaluations.[145,146] In addition to the selection/reproduction schemes described below, our molecular evolution process employed elitism; that is, the highest-scoring member from each generation was automatically propagated to the subsequent generation.

As noted in Table 5-3, two computational experiments were carried out with roulette-wheel selection. In this selection scheme, each member of the population is randomly selected to be a parent with probability $P_{sel}(x_i) = f_i^s / \sum_j f_j^s$ proportional to its scaled fitness value. In this work, the fitness value is *scaled* to accommodate members with negative fitness, or with fitnesses that are closely grouped in value away from zero.[145] The scaled fitness value f^s calculated by window scaling: the scaled fitness score is the raw fitness score of that member minus the fitness value of the minimal-scoring member: $f_i^s = f_i + \min_j f_j$.

In the N_t-member tournament selection scheme, N_t designs are randomly chosen at a time from the population, with all N members having equal probability. Then, the highest-scoring member among the N_t chosen members is designated a parent. This process is repeated (with replacement) to generate the entire parental subset. It should be noted that for either selection technique the parental subset may contain multiple copies of certain members of the current population.

In general, tournament selection (with a small value of N_t, say 2 or 3) is often recommended over roulette wheel section:[145] it obviates the need to re-scale raw fitness scores to obtain the uniformly positive scores required by proportionate selection, and in general, tournament selection has been shown to achieve convergence faster than proportionate selection in simple demonstration problems.

In this work, performing molecular evolution using automated molecular dynamics simulations is complicated by the fact that thermodynamic measurements made

from such simulations include statistical errors, due to limited sampling. To address the statistical error in the measurement of each ligand's adsorption selectivity, we developed a selection scheme that accounts for these uncertainties, and which selects tournament winners stochastically, which we denoted "fuzzy tournament selection." In this scheme, two members were selected at random from the population, as in traditional tournament selection. Then, their scores and the standard errors of those scores were used to calculate a scaled score difference:

$$z = \frac{y_1 - y_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

Where y_i and σ_i are the measured score and standard error of member i in the tournament. Then member 1 is chosen with probability $p_1 = \Phi(z)$, where $\Phi(\cdot)$ is the standard normal CDF. This approach, which is based on the statistical method for estimating the difference of sample means, ensures that two members with very similar score values (as compared to the statistical error) are chosen with roughly equal probability, and thus was intended to diminish the effect of the fitness function's statistical noise on the evolution process.

Using the selection scheme described above, $(N-1)$ pairs of parent sequences were selected based on fitness scores from the population, with the -1 term accounting for the member selected by elitism. Then, for each pair of parent designs, a crossover operation was applied with probability $p_{\text{crossover}} = 0.40$ or 0.80, as listed in Table 5-3. In this case, each parent's genome was split into two parts at a random location, and the corresponding portions from the two parents were interchanged. In cases where crossover was not applied, one member of the pair was subjected, with equal probability $\frac{1}{3}(1 - p_{\text{crossover}})$, to either gene deletion (at a random position), gene insertion (of a random functional group at a random position), or gene mutation (to a random functional group at a random position).

To better understand the effects of selection scheme and evaluation details on the evolution process, we performed four *in silico* experiments with different procedural parameters, as detailed in Table 5-3.

The initial population in Experiment I consisted of eight ligand designs (of length 5–13) taken from evolution experiments using the surrogate objective, plus 9 randomly-generated ligands of each length from 4 to 7 functional groups, for a total of 45 ligands. The initial population in subsequent experiments consisted of 45 randomly-generated ligands, constructed to have a uniform length distribution from 4 to 7 (Exp. II) or 4 to 8 functional groups (Exp. III and IV).

Table 5-3. Summary of genetic algorithm and evaluation function parameters in four *in silico* evolution experiments. The population size in each experiment was 45.

exp.	GA selection technique	fitness score scaling	crossover prob.	number generations	MD prod. length	accumulative scoring	source of init. coordinates
I	roul. wheel	window scaling	0.40	75	3 ns	no	newly generated for each eval.
II	roul. wheel	window scaling	0.40	45	6 ns	no	newly generated for each eval.
III	2-mem. tourn	N/A	0.80	88	4.5 ns	yes	copied from previous eval., if available
IV	2-mem. fuzzy tourn.	N/A	0.80	68	4.5 ns	yes	copied from previous eval., if available

5.5 Measuring diversity of a population of molecules

Finally, when carrying out molecular evolution, it is helpful to understand the degree of homogeneity within the ligand population as it evolves. There are ways to measure the difference between molecules' so-called "2D structures," like the Tanimoto similarity,[63,147] and such a metric can be applied in a pairwise fashion to produce an overall diversity measure. In our problem, we implemented a phenotypic diversity metric, based on estimated values of several properties (number of H-bond donors, number of H-bond acceptors, molecular volume, hydrophobicity, *etc.*) for each ligand, using the ligand's structure and QSAR relationships. After scaling all such measurements by their standard deviations in a reference population, as is done in the ChemGPS system,[148,149] the diversity metric was defined as the sum of pairwise differences in property space. Details can be found in Section 5.5.1 of the Supplemental Information.

5.5.1 Property estimation and phenotypic diversity

Once ligands are formed using a sequence of functional groups, various physico-chemical properties can be evaluated, as summarized in Table 5-4 and explained in greater detail below.

Table 5-4. QSPR measurements available for fast phenotypic characterization of ligands.

$\|p\|$	magnitude of dipole moment
N_{don}	number hydrogen bond donors
N_{hb}	number hydrogen bond acceptors
N_{hb}	total number hydrogen bond donors/acceptors
$\log P$	logarithm of octanol/water partition coefficient
V	estimated molecular volume
a	ligand spacing on 2D square lattice, from 3D config.

Polarity. The dipole moment of each candidate ligand can be caluclated using the definition for net-neutral systems: $\mathbf{p} = \sum_i q_i r_i$. The charge q_i on each atom is the partial charge assigned to each atom using an AM1-BCC semi-empirical QM calculation[150] on an isolated ligand molecule. This calcuation is performed routinely for each ligand as part of the GAFF parameterization process. The reported value of the dipole moment is the magnitude of this vector.

Hydrogen bonding properties. The number of hydrogen bond donors and the number of hydrogen bond acceptors can in the molecule can be used as two discrete descriptors. The proposed list of atom types that would be considered hydrogen bond donors and acceptors is provided in Table 5-5 below.

Lipophilicity/Hydrophobicity. The hydrophobicity of each candidate ligand is quantified using an estimated octanol-water partition coefficient, P. Several quantitative structure-property relationships (QSPRs) exist for $\log P$, the logarithm of this partition coefficient.

For this work, we have access to both the structure/topology as well as reasonable 3D configurations of the ligands, and the approach to $\log P$ estimation we used, by Bodor and Buchwald,[151] employs both kinds of information. The partition coefficient is estimated using the volume V contained within the van der Waals radii of all its

Table 5-5. GAFF atom types of hydrogen bond donor and acceptors counted in ligand descriptions.

hydrogen bond donors	
ho	hydrogen in hydroxyl group
hn	hydrogen in amine group
hs	hydrogen on sulfur
hydrogen bond acceptors	
oh	oxygen in hydroxyl group
o	oxygen in carbonyl group
n3, nh	nitrogen in amine group
s2	sulfur in −C=S carbonyl analogue
sh	sulfur in thiol group
f	fluorine atom

atoms, along with an additive parameter N that is the sum of contributions from various functional groups, and which represents hydrogen bond acceptor basicity:

$$\log P = 0.032V + (0.010V)\, I_{alk} - 0.723N$$

In their work, Bodor and Buchwald calculated V using adjusted vdW radii, then accounted for any regions of space in which two (or three) vdW spheres overlap. The variable I_{alk} is an indicator variable that is 1 for a saturated alkane and 0 for all other molecules.

This equation is applied to newly-generated structures by estimating the volume V by using a group-contribution adaption[152] of the Bodor-Buchwald estimation, which uses only topological information. Under this approach, the contributions of each possible functional group "bead" to the estimated volume V and total oparameter N are given in Tables A-2 and A-1 (pages 178 and 178).

Molecular size and shape. The molecular volume calculated in the step above gives one measure of the gross size of the molecule. In order to characterize the gross shape of the molecule, it would be possible to calculate the moment of gyration tensor, which is analogous to the moment of inertia tensor, but is not weighted by

mass. Its elements are:

$$\mathbf{S}_{nm} = \frac{1}{N} \sum_i r_n^{(i)} r_m^{(i)}$$

where the coordinates are measured relative to the center of mass $\mathbf{r}_{com}$, and where N is the number of atoms.

This tensor can then be diagonalized so that $\mathbf{S}' = \mathrm{diag}\,(l_x, l_y, l_z)$, and these $\{l_i\}$ values give an idea of the gross shape of the molecule. These three independent quantities can be thought of as the axes of an ellipsoid that represents the distribution of atoms in space, if all atoms are considered equivalent (i.e. no weighting).

5.5.2 Pairwise diversity measurement and property distributions of populations

The diversity measurement is defined below. It is the average, over all pairs within a population, of a distance $d(\cdot,\cdot)$, which is the sum of the difference in properties between a pair, with each property measurement normalized by a standard deviation. The standard deviation of each property, listed in Table 5-6, was evalauted using a random population of 2,000 members, having properties shown in Figure 5-5 (page 103).

$$\text{diversity} = \frac{1}{N_{pairs}} \sum_{\substack{\text{members} \\ i<j}}^{i,j} d(y_i, y_j)$$

$$d(r,s) = \left(\sum_{\text{properties } l} \frac{1}{\sigma_l^2} \left[p_l(r) - p_l(s) \right]^2 \right)^{1/2}$$

In order to understand in what way a population is diverse, *i.e.* which properties differed most or least among population members, the relative diversity of a population in a certain property l can be computed: it is the variance of the property l in the population, divided by variance of the property in the reference population.

$$\text{rel. div. of property } l \text{ in set A} = \frac{\left\langle \frac{1}{\sigma_l^2} \left(\Delta p_l\right)^2 \right\rangle_A}{\left\langle \frac{1}{\sigma_l^2} \left(\Delta p_l\right)^2 \right\rangle_{\substack{\text{ref.} \\ \text{pop.}}}}$$

where $\langle \cdots \rangle_S$ denotes a pairwise average over all pairs in a population S.

Under this metric, the diversity of a random population was found to be 2.90.

Table 5-6. Scaling factors used for distance measurements in property space.

PROPERTY	index l	scaling factor σ_l
number rings	1	1.192
number H-bond donors	2	1.647
number H-bond acceptors	3	2.045
molecular volume	4	85.01
$\log(P)$ = log of part. coeff.	5	2.940

After developing this diversity metric, a second population of 2,000 ligands was randomly generated in the same way as the first (properties not shown). Its measured diversity was 2.87, very close to that of the reference population, and the relative diversity of each property between 96 and 99%, indicating that the scaling factors calculated from the reference population were obtained in a consistent fashion from this other, large sample.

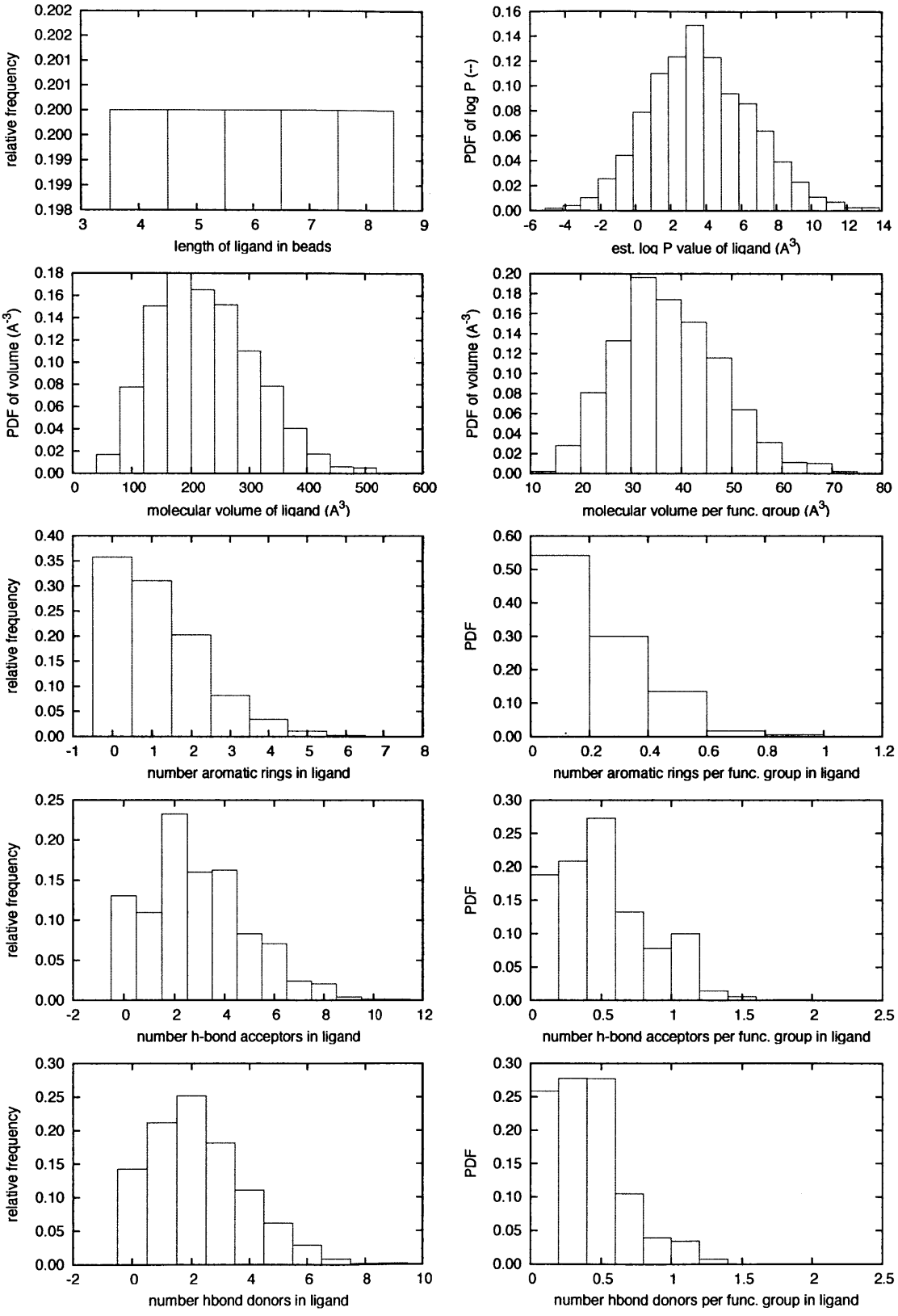


Figure 5-5. Distribution of properties of reference population of 2,000 random ligands having uniform distribution of length.

5.6 Testing molecular evolution with a surrogate objective function

In carrying out molecular evolution to optimize adsorption-based selectivity, it was not clear at first what evolution parameters (such as selection technique, population size, crossover rates, and so on) would work best for this problem. To better understand this question, we developed a surrogate objective function, which could evaluate a ligand's fitness based on its structure alone, in the manner of a (deterministic) quantitative structure-activity relationship (QSAR). This work is described in Section B.4 of the Supplemental Information, and the main conclusions can be summarized as follows:

- The genetic algorithm was able to optimize features of the ligand design that were minor components of the surrogate objective function, such as the hydrophobicity of the ligand design.

- For the deterministic surrogate objective function, a population size of 30 converged fastest (when measured in function evaluations), compared to sizes of 60 and 90.

- When unbiased noise was added to the objective function at levels of 7, 14, or 20% of the maximum value, convergence was delayed and the distribution of fitess scores in the evolving population is broader, compared to a base case without added noise.

- In a study of different score scaling methods preceding roulette wheel selection, linear scaling (with scaling coefficient 2.0) and rank scaling (with coefficient 2.0) both worked very well, but each technique was highly sensitive to the coefficient used. Window scaling worked nearly as well as those methods' optimal performance, and better than roulette wheel selection with no scaling.

Details about the objective function used and the data leading to these conclusions can be found in SectionB.4 of the Supplemental Information.

Chapter 6

Molecular designs for adsorption-based purification of a pharmaceutical intermediate

Results from applying molecular evolution to the design of a surface-bound ligand for the separation of E2 and E6 are presented and analyzed in this chapter. Because Experiment II began with a randomly-generated set of ligand designs, it will be used to illustrate results obtained from evaluation (in Section 6) and molecular evolution (in Section 6.2).

6.1 Ligand population and evaluation outcomes

To understand the variety of ligand designs that could emerge from our linear fragment-based construction approach, we generated a number of random ligand designs, using the FORM2GEOM software's random-sequence feature. The properties of a reference population of 2,000 ligands having uniform distribution of length from four to eight functional groups are shown in Figure 5-5 in the previous chapter; a second randomly-generated sample of the same size had equal values of properties' means and standard deviations, within about 1%. The initial population used in Experiment II was also randomly generated, with near-uniform distribution of length between 4 to 7 func-

tional groups.

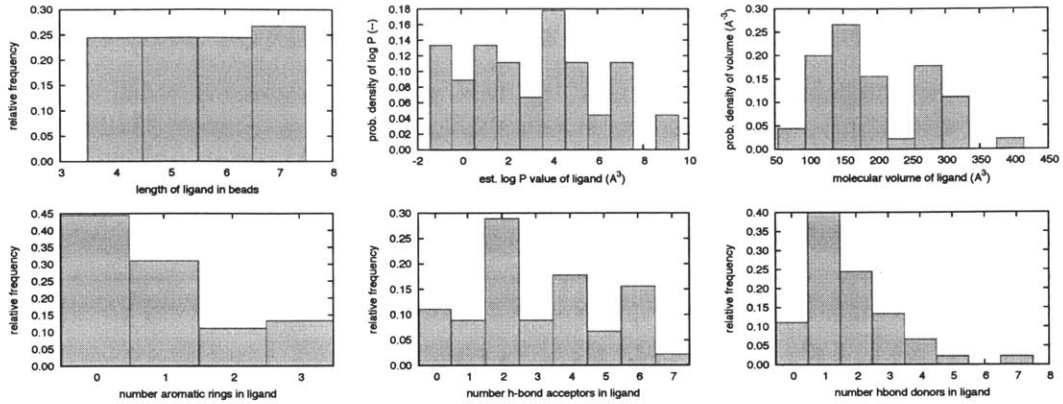
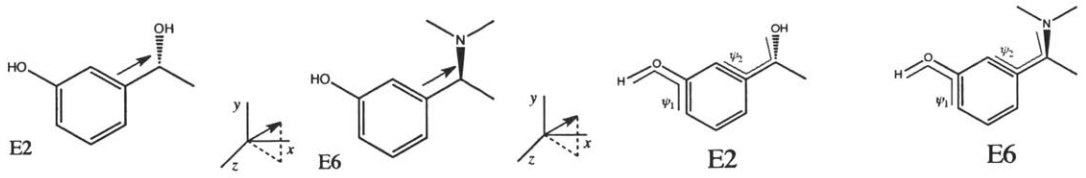


Figure 6-1. Properties of constituent ligand designs in generation 1 of Experiment II. These distributions can be compared to those shown in Figure 5-5 for a large, randomly-generated population.

As noted in the Methods section, evaluation of each ligand design was carried out using several nanoseconds of production MD, after subjecting the initial ligand structure to minimization, annealing, and equilibration, and then simulating the binding of the E2 and E6 molecules in separate simulations.

We sought to confirm that these simulations of adsorption/binding broadly sampled an energetically-relevant set of ligand-target conformations. In these simulations, the bound atom in the ligand is anchored to the surface, restricting the ligand's translational freedom; additionally the soft potential partially limits the rotational and internal degrees of freedom. The adsorbing (E2 or E6) molecule's conformation was represented by its two internal dihedral angles (designated ψ_1 and ψ_2 in Figure 6-2), and its absolute orientation (measured by direction cosines of the vector in Figure 6-2(a)).

Distributions of the E2 and E6 molecules' absolute orientations and dihedral angles in the evaluation of Experiment II's generation 1, candidate 25 are shown in Figure 6-3. This ligand candidate was chosen because it had the median fitness score (0.21 ± 0.14 kcal/mol) of generation 1 in that experiment. Figure 6-3(a) shows that in each simulation, the E2 and E6 molecules were in close contact with the surface-bound ligand molecule. This was confirmed by visualizing the trajectory, and by measuring the distance along the z axis between the two molecules centers of mass (data not



(a) Bond vector used to define absolute orientation.

(b) Dihedral angles.

Figure 6-2. Definitions of absolute orientation and internal degrees of freedom for E2 and E6.

shown).

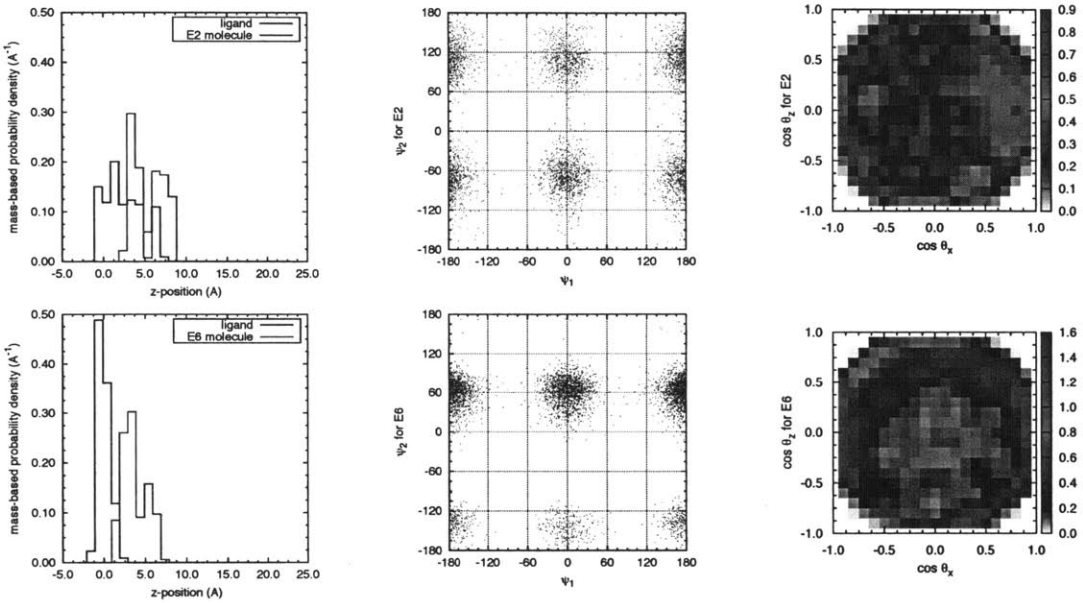
In examining the sampling that took place in evaluation simulations, Figure 6-3(b) shows that both the E2 and E6 molecules explored their dihedral angle space, and did so independently, as the joint distribution of (ψ_1, ψ_2) could be separated into a product of distributions of ψ_1 and ψ_2. The E2 and E6 molecules also sampled many different absolute orientations (with respect to the lab frame suggested by the "wall" in the xy-plane), as shown in Figure 6-3(c). Finally, the convergence of $\Delta\Delta E_{ads}$ and its two component adsorption energies are shown in Figures 6-3(d) and 6-3(e). In this case, the fitness score converged to a stable value after about 3.0 ns of production MD.

Similar measurements were made for many other ligand evaluations, and similar results were observed.

To understand the reproducibility of fitness score evaluations, seven ligand designs were evaluated five times each using the procedure described in the previous section. The production MD was extended to 20 ns in each case, and in the majority of these cases, score consistency was obtained within about 10 to 15 ns, equivalent to two to three 4.5- or 6.0-ns evaluations. Figure 6-4 is one such example, and others can be found in the Supplemental Information. Because successful candidates tend to be re-evaluated in successive generations, these ligand candidates will quickly undergo several dozen nanoseconds of production MD in those experiments (III and IV) with cumulative scoring. Even when evaluations are independent, as in Experiment II, the scores appeared consistent from run to run. An example of the distribution of fitness scores in multiple evaluations is shown in Figure 6-5, and others can be found

in Figures B-18 and B-19 of the Supplementary Information.

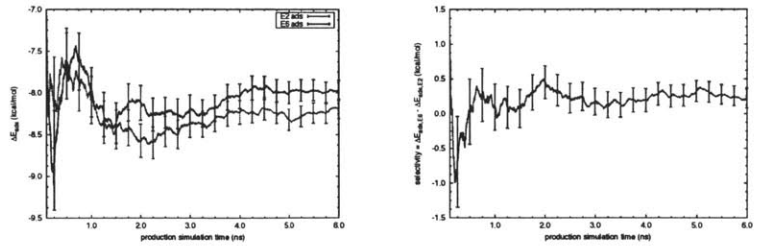
After evaluating all 45 randomly-generated members of the initial population in Experiment II, fitness scores ranged from -0.68 to $+1.6$ kcal/mol, with standard errors of about 0.15 kcal/mol, as shown in Figure 6-6. The magnitude of the standard errors, calculated using the method of Allen and Tildesley,[144] were confirmed using bootstrap sampling.



(a) Mass distribution profiles in ligand–E2 and ligand–E6 simulations.

(b) Dihedral angles of E2 and E6 molecules in simulations with ligand.

(c) Absolute orientation of E2 and E6 molecules in lab frame, as measured by direction cosines of defined orientation vectors.



(d) Values of $\Delta E_{ads,\mathrm{E2}}$ and $\Delta E_{ads,\mathrm{E6}}$ over course of simulations.

(e) Value of selectivity score $\Delta\Delta E_{ads}$ over course of simulations.

Figure 6-3. Evaluation of ligand candidate 25 of generation 1 in Experiment II, having sequence $\mathrm{HO{-}CH_2{-}COO{-}CH(C_6H_5){-}NH{-}CH{=}C{=}CH_2}$. This ligand was chosen because its fitness score was the median in generation 1 of Experiment II.

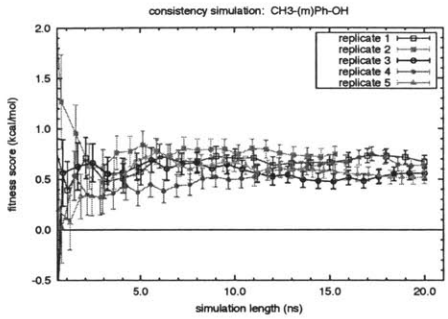


Figure 6-4. Convergence of fitness score of ligands with structure $CH_3-(m)Ph-OH$. Error bars are ± 1 standard error. Similar results for six other ligand designs are shown in Figure B-3 in the Supplementary Information.

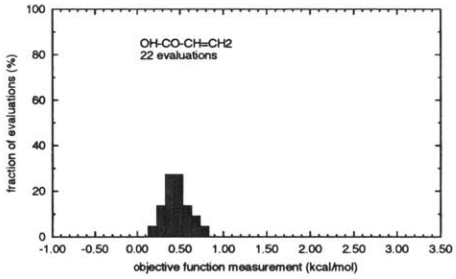


Figure 6-5. Histogram of fitness score evaluations for the indicated ligand design in Experiment II. In that experiment, all fitness evaluations were independent.

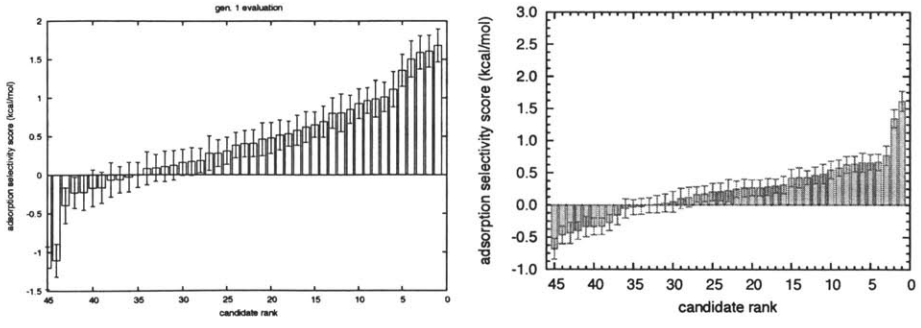


Figure 6-6. Fitness scores of members of initial populations ($N = 45$) in Experiment I (*left*) and Experiment II (*right*), in rank order.

6.2 Molecular evolution outcomes

The molecular evolution processes observed in Experiments I through IV are summarized in Figures 6-7 through 6-10. These figures show that as evolution takes place, the populations become less diverse, and at the same time, members' fitness scores, as measured by their median and 80^{th} percentile, increased. This process does not occur in a smooth way, because in Experiments I and II, when a ligand design is re-evaluated, it takes a new fitness score independently of its previous performance.[153]

As the GA was applied to the ligand population, the score distribution generally shifted toward higher scores, as suggested by Figures 6-7 through 6-10. However, even in later generations, the distributions generally contained a left tail—that is, they typically contained poorly-performing offspring of the previous generations' parental subset, which tends to contain better-than-average ligand designs. This illustrates that because of the molecular genome's discrete nature, crossover or changes in a single gene can lead to a significantly different phenotype (*i.e.* physicochemical properties) *and* fitness.

In all four experiments, the number of unique ligands evaluated (shown in the top panels of Figures 6-7 through 6-10) increases at a rate less than 45 per generation, especially towards the end of each experiment, because each generation contains designs that have already been evaluated. As noted above, the elitism feature of the GA automatically propagates the top-scoring design from one generation to the the succeeding generation. In addition, the genetic algorithm allows multiple copies of a single design to exist within a single generation. This feature was included in the algorithm to allow designs with favorable performance to "win out" by generating replicates within each generation. These repeated designs, which are treated as independently-evaluated members, would then increase the likelihood of reproduction and propagation of the successful design to the succesive generation. Each experiment was halted after the diversity sharply dropped; in Experiments III and IV, in order to test the "consolidation" process, it was applied after the population's first steep decline in diversity (Exp. III) and after a moderate decline in diversity (Exp. IV).

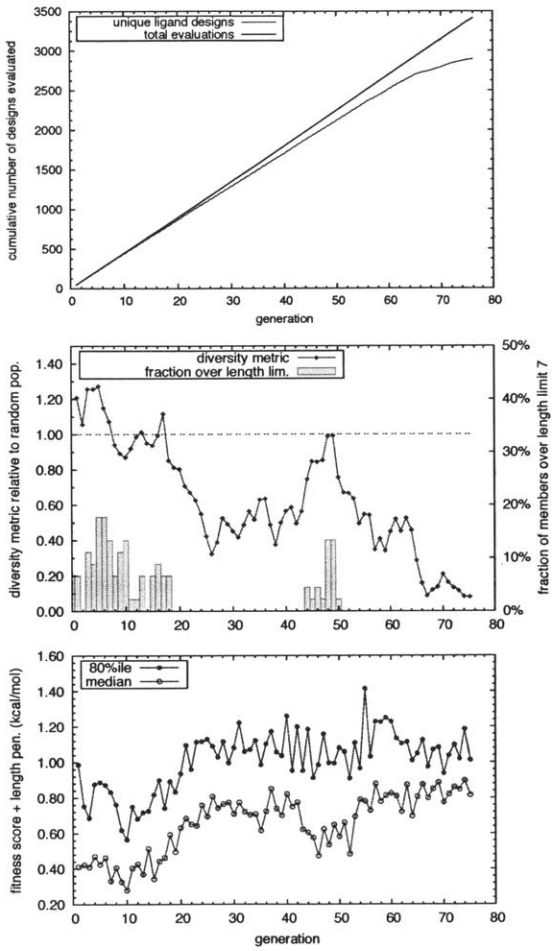


Figure 6-7. Characterization of evolution over generations 1 to 76 in Experiment I, which featured $N = 45$ ligands, roulette wheel selection after window-based scaling, and 3.0-ns production MD in each evaluation.

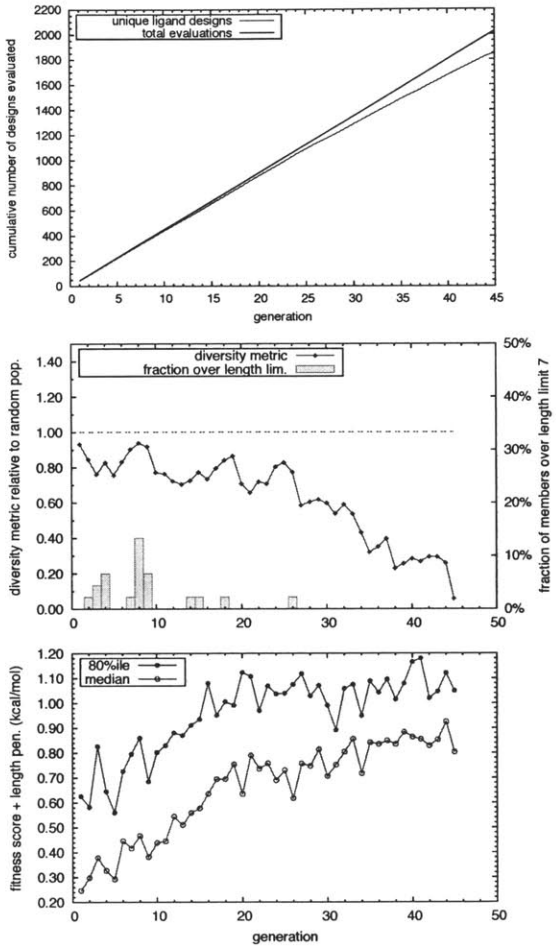


Figure 6-8. Characterization of evolution over generations 1 to 45 in Experiment II, which featured $N = 45$ ligands, roulette wheel selection after window-based scaling, and 4.5-ns production MD in each evaluation.

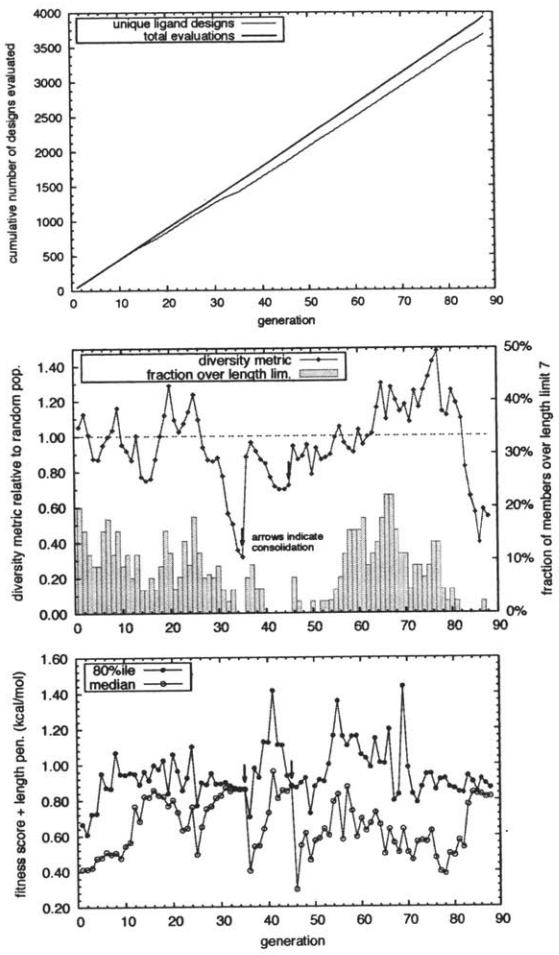


Figure 6-9. Characterization of evolution over generations 1 to 88 in Experiment III, which featured $N = 45$ ligands, 2-member tournament selection and 4.5-ns ligand evaluation. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

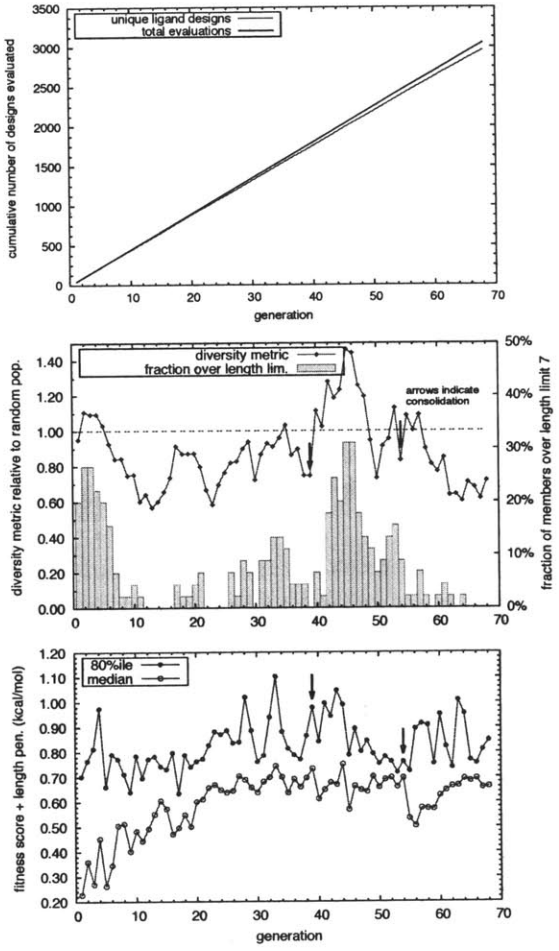


Figure 6-10. Characterization of evolution over generations 1 to 68 in Experiment IV, which featured $N = 45$ ligands, 2-member fuzzy tournament selection and 4.5-ns ligand evaluation. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

Ligand design

The top-scoring ligand designs from each experiment are listed in Table 6-1. As noted in the caption, only ligand desings with multiple evaluations are included, to lessen the chance of identifying a ligand design with a high score that was a statistical fluctuation, *i.e.* a departure from its long-run, mean fitness score.

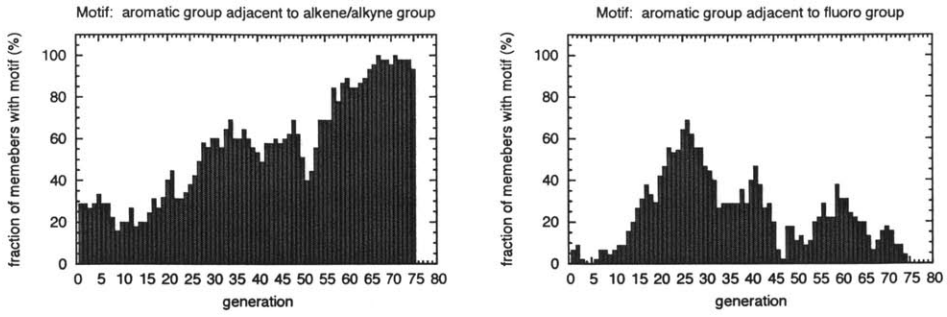
In the first section of Table 6-1, showing results from Exp. I, eight out of 10 ligands have evaluation times of 9.0 ns or less, corresponding to 2 or 3 evaluations. This suggests that, in Experiments I and II, promising ligand candidates like the eight listed can be eliminated from the population by non-selection, after a one-time fluctation of its fitness score in the negative direction, because ligands were evaluated independently in each generation.

This potential to drop promising candidates was considered a drawback, and motivated us to implement "accumulative scoring" in Experiments III and IV. In the accumulative scoring scheme, the energy values sampled in the current evaluation are averaged with all production MD in previous generations' evaluations of the same design. This led to more consistent evaluation results, as shown in fitness score histograms from the two methods in Figures B-18, B-19 and B-29 in the Supplemental Information, and would tend to mitigate this problem.

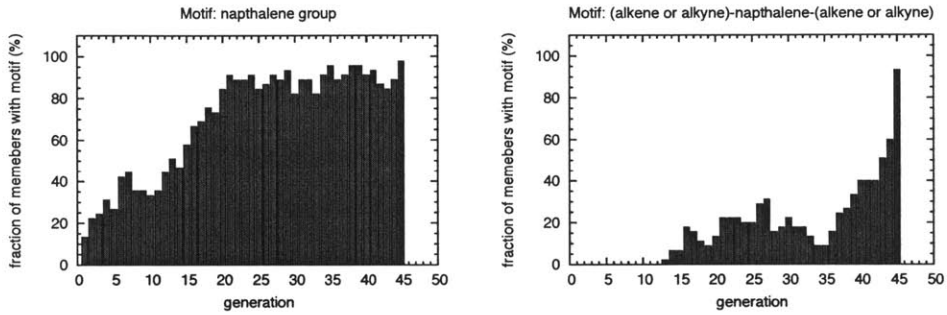
In the ligand designs in Table 6-1, certain functional groups appear with greater-than-random frequency, namely phenyl, napthalene, sp^2 groups (ethene, ethyne, allenene, and amino), and hydrogen bond-accepting groups (hydroxyl, aldehyde, carbonyl, carboxyl). Before carrying out molecular evolution, we had identified H-bond acceptors as a chemical motif that could contribute to selectivity in adsorption, because the E2 molecule contains a hydrogen bond donor in the hydroxyl group that differentiates it from E6. Phenyl, naphthyl, and the other groups listed above had not been identified as potentially contributing to selectivity by the chemists and engineers who initially examined the separation problem. The reason their inclusion in ligands leads to selectivity is discussed in Section 6.3 below.

A selection of relevant chemical motifs from each experiment is shown in Fig-

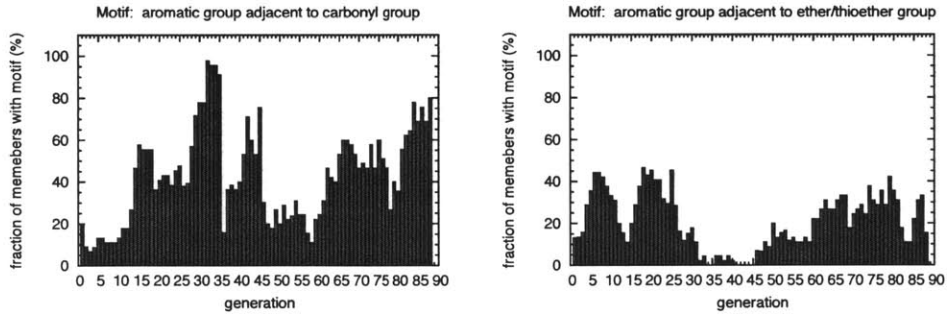
ure 6-11. In all four experiments, phenyl and naphthyl groups grew to be present in a majority of members, so motifs containing those groups adjacent to others are shown. Figures 6-11(a) through 6-11(d) show that motifs could grow popular somewhat quickly, expanding from approximately 10% of the population to a majority or near-takeover of the population within about 20 generations. In particular, Figures 6-11(b) and 6-11(c) show that a successful motif can be germinated during the evolution process, and then successfully emerge to be present in a significant fraction of the population's 45 members.



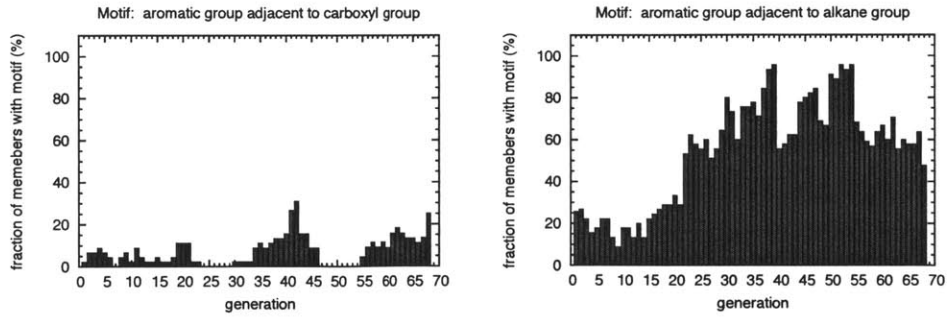
(a) Experiment I



(b) Experiment II



(c) Experiment III



(d) Experiment IV

Figure 6-11. Prevalence of motifs in Experiments I through IV. Motif descriptions are listed in the title of each graph. An "aromatic group" is a phenyl or napthalene group.

Table 6-1. Top-scoring ligand designs in each experiment, for which at least two ligand evaluations were performed (except Experiment IV, which has a threshold of 4 evaluations). Listed scores are averages of all evaluations for each design, and inlcude the length penalty each ligand. All score and ΔE_{ads} values are in kcal/mol.

	AVG. SCORE	PEN.	$\Delta E_{ads,\text{E2}}$	$\Delta E_{ads,\text{E6}}$	PROD. (ns)	SEQUENCE
Exp. I	1.60 ± 0.13	0.0	-7.2	-5.6	6.0	F–$C_{10}H_8$–CH=C=CH–(trans)CH=CH–H
	1.50 ± 0.14		-10.3	-8.8	9.0	Cl–C(Ph)H–$C_{10}H_8$–CH=C=CH–CH(COOH)–CH(NH_2)–C(CH_3)$_3$
	1.44 ± 0.15	-0.1	-7.4	-5.9	6.0	CH_2=C=CH–(trans)CH=CH–CH(NH_2)–(p)Ph–C≡C–(trans)CH=CH–CF_2–CH_3
	1.39 ± 0.08		-6.8	-5.4	15.0	COOH–CH(COOH)–CH(COOH)–Cl
	1.32 ± 0.13		-6.6	-5.3	6.0	SH–$C_{10}H_8$–CH_3
	1.17 ± 0.17		-8.0	-6.8	6.0	CH_2=CH–CH(NH_2)–(p)Ph–(p)Ph–$C_{10}H_8$–CH(CH_3)–CH_3
	1.17 ± 0.13		-8.9	-7.7	9.0	Cl–C(Ph)H–$C_{10}H_8$–C(Ph)H–CH_3
	1.16 ± 0.11		-7.3	-6.2	9.0	F–$C_{10}H_8$–CH=C=CH–CH=C=CH–CH_3
	1.12 ± 0.12		-8.1	-7.0	9.0	CH_2=CH–$C_{10}H_8$–C≡C–$C_{10}H_8$–H
	1.10 ± 0.13		-6.9	-5.8	6.0	NH_2–$C_{10}H_8$–C≡C–CHO
	1.10 ± 0.05		-7.1	-6.0	54.0	CH_2=CH–$C_{10}H_8$–CH=C=CH–H
Exp. II	1.04 ± 0.10		-7.8	-6.8	12.0	CH_2=CH–$C_{10}H_6$–NH–CO–NH–CH=CH_2
	0.99 ± 0.05		-7.3	-6.3	48.0	OH–CH=C=CH–$C_{10}H_6$–CH=C=CH_2
	0.97 ± 0.08		-7.5	-6.5	18.0	CH_2=CH–$C_{10}H_6$–CO–CH=CH_2
	0.95 ± 0.11		-7.7	-6.7	12.0	CH_2=CH–(p)Ph–(m)Ph–CH(CH_3)–NH–CH=CH_2
	0.94 ± 0.03		-7.8	-6.9	138.0	CH_2=CH–$C_{10}H_6$–CO–NH–CH=CH_2
	0.93 ± 0.11		-9.1	-8.2	12.0	CH_2=CH–$C_{10}H_6$–CHOH–$C_{10}H_6$–CH=CH_2
	0.93 ± 0.07		-7.7	-6.8	30.0	CH_2=CH–$C_{10}H_6$–CH=C=CH–CO–NH–CH=CH_2
	0.93 ± 0.06		-7.3	-6.4	24.0	OH–CO–$C_{10}H_6$–C≡CH
	0.92 ± 0.10		-7.3	-6.4	12.0	CH≡C–CH(COOH)–(p)Ph–(o)Ph–F
	0.92 ± 0.02		-7.1	-6.2	432.0	CH_2=CH–$C_{10}H_6$–CH=C=CH_2
Exp. III	3.02 ± 0.10		-7.9	-4.9	22.5	$CONH_2$–C(CH_3)$_2$–$C_{10}H_6$–$C_{10}H_6$–C(CH_3)$_2$–H
	2.57 ± 0.05		-9.4	-6.8	81.0	$CONH_2$–C(CH_3)$_2$–$C_{10}H_6$–$C_{10}H_6$–$C_{10}H_6$–CH_3
	2.48 ± 0.05		-7.8	-5.3	49.5	CH_3–$C_{10}H_6$–(trans)CH=CH–COOH
	2.40 ± 0.09		-12.3	-9.9	22.5	$CONH_2$–$C_{10}H_6$–CF_2–O–$C_{10}H_6$–Ph
	2.35 ± 0.07				22.5	Ph–CO–C(CH_3)$_2$–CO–CH_3
	2.25 ± 0.06	-0.4	-11.0	-8.3	49.5	Ph–CH(iBut)–(m)Ph–O–NH–C(CH_3)$_2$–$C_{10}H_6$–O–CH_3
	2.03 ± 0.08		-9.9	-7.9	31.5	Ph–CH(iBut)–$C_{10}H_6$–CH(COOH)–H
	1.81 ± 0.06	-0.1	-10.0	-8.0	49.5	$CONH_2$–O–CO–$C_{10}H_6$–C(CH_3)$_2$–$C_{10}H_6$–O–CH_3
	1.77 ± 0.05		-7.5	-5.8	63.0	$CONH_2$–O–$C_{10}H_6$–O–CH_3
	1.68 ± 0.08		-7.2	-5.5	22.5	CH_2=C=CH–$C_{10}H_6$–CF_2–O–C(Ph)H–CHO
	1.58 ± 0.07		-9.0	-7.5	27.0	$CONH_2$–$C_{10}H_6$–CO–C(CH_3)$_2$–CO–H
	1.28 ± 0.10		-10.2	-8.9	18.0	Ph–(p)Ph–CH(iBut)–CH(COOH)–C(CH_3)$_2$–H
	1.27 ± 0.09		-9.2	-7.9	22.5	$CONH_2$–$C_{10}H_6$–CH(CH_3)–(p)Ph–CH_3
	1.18 ± 0.07		-3.2	-2.1	18.0	F–CH=C=CH–(trans)CH=CH–COOH
Exp. IV	3.67 ± 0.04		-10.1	-6.4	72.0	COOH–(m)Ph–(m)Ph–Ph
	2.00 ± 0.10	-0.4	-11.1	-8.7	18.0	CH_3–(m)Ph–CH(COOH)–(m)Ph–(m)Ph–(m)Ph–O–CH(CH_2CH_3)–COOH
	1.42 ± 0.09		-8.4	-7.0	18.0	CH_3–(m)Ph–CF_2–O–(m)Ph–Ph
	1.42 ± 0.05		-9.1	-7.7	58.5	CH_3–(m)Ph–CH(COOH)–(m)Ph–(m)Ph–CH_3
	1.12 ± 0.06		-7.0	-5.8	31.5	CH_3–(m)Ph–CH(CH_2CH_3)–Ph
	1.03 ± 0.10		-9.0	-8.0	18.0	CH_3–CO–(m)Ph–(m)Ph–(m)Ph–CH_2–Ph
	1.01 ± 0.06		-6.4	-5.4	31.5	F–CHOH–CO–CH(iBut)–COOH
	0.96 ± 0.07		-8.0	-7.0	27.0	CH_3–(m)Ph–(m)Ph–COO–CH(CH_3)$_2$
	0.95 ± 0.09		-8.3	-7.3	22.5	CH_2=CH–(m)Ph–(m)Ph–(m)Ph–Ph
	0.95 ± 0.08		-7.2	-6.3	18.0	CH_3–(m)Ph–(m)Ph–CF_2–CH_3

6.3 Mechanism of selectivity for E2 adsorption over E6 adsorption

One advantage of employing our approach to accomplish *in silico* screening is that it presupposes no particular mechanism of selectivity. However, in contrast to design approaches that first develop physical insights into the underlying physical process, the reasons that a particular design is successful are not necessarily clear, and may require *a posteriori* investigation.

As shown in Figure 6-11, ligand designs that emerged from molecular evolution, including many of the highest-scoring designs in Table 6-1, contained aryl (phenyl or naphthalene) groups, alkenyl/alkynyl groups, or carbonyl groups, especially in close proximity to each other. These functional groups are all planar, although we did not at first understand the significance of this fact, nor why they had emerged from the GA.

To understand why such ligands would achieve selectivity, we examined their partial charge profiles, to see if the charge assignment process was generating concentrations of negative charge, which could interact favorably with E2's differentiating hydroxyl group. We also checked to see whether the presence of a sp^2 terminal group (*e.g.* CH_2=CH–) at the simulated wall kept the ligand's principal axis more perpendicular to the wall, which might influence the adsorption of E2 or E6. Neither of these possibilities were supported by the simulation trajectories (data not shown).

Instead, trajectory visualization suggested that that planar or mostly-planar ligand molecules achieve selectivity by allowing E2's phenyl ring to lie flat against an aryl core in the ligand. The E6 molecule is prevented from doing so by steric interference of its tertiary amine group. Several snapshots are presented in Figure 6-12, in which each subfigure includes the minimum-potential-energy frames from all three simulations.

To quantify this difference, and compare selective and non-selective ligand designs, we have measured aryl-aryl alignment between the adsorbing molecule (E2 or E6) and the ligand molecule. This approach does have limitations: first, it relies on

molecular mechanics (and the AMBER force field in particular), which makes no special provision for π–π stacking interactions, other than ordinary van der Waals forces, and second, there is no natural way to apply analysis to negative control ligands that do *not* contain an aryl ring.

To measure the alignment of the ligand and adsorbing molecule's aryl cores, the following variables were measured at every recorded frame in a simulation: h, the distance between the target (E2 or E6) center of mass and ligand aryl plane; ϕ_c, the bond orientation angle, between the ligand's aryl normal vector and joining the centers of mass; and ϕ_q, the relative orientation angle between the two aryl rings' normal vectors. The two angles are depicted in Figure 6-13 below.

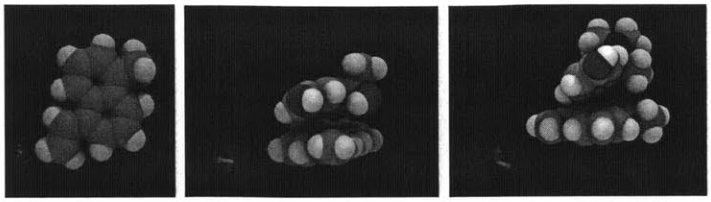
The values of ϕ_q and h were then recorded and histogrammed for all configurations in which $\phi_c < \phi_{\text{cutoff}} = 60°$. This cutoff was imposed to avoid counting occurences of false aryl alignment, in which the target molecule is aligned with the ligand, but not "above" the ligand's aryl core.

This process is depicted in the renderings in Figures 6-14(a) and 6-14(b) below. Visually, it appears that the adsorbing molecule's aryl direction vector (depicted in purple for each configuration) are loosely aligned with the ligand's direction in the case of E6 (*right*), but more strongly aligned in the case of E2 (*left*).

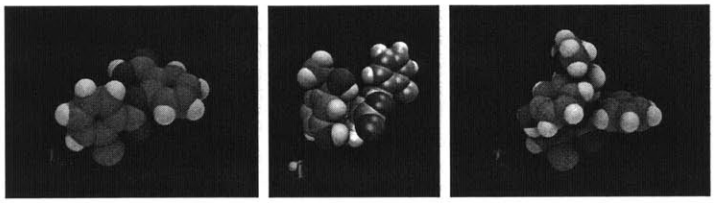
Histograms of the aryl normal/aryl normal angle ϕ_q (Figure 6-14, left panels) and of the same quantity as a function of height (right panels) show distinct peaks for E2 adsorption at low values of ϕ_q, even though these angles are entropically disfavored (compare to the random distribution), and almost all recorded frames had the ligand–E2 inter-ring distance between 3.25 and 4.25 Å. That state, with minimal values of h and low values of ϕ_q, corresponds to aligned, approximately stacked rings. In contrast, the ligand–E6 inter-ring distances were spread over a greater range of values, while the ϕ_q distribution was much less peaked at low values.

As noted above, a ligand design with an aromatic group must be chosen to serve as a non-positive control. In this case, the design in generation 3, cand. 20 in Experiment II had a poor selectivity score of −0.36 kcal/mol over 9 ns of production. The non-selective control does not exhibit such noticeable differences between E2–ligand and

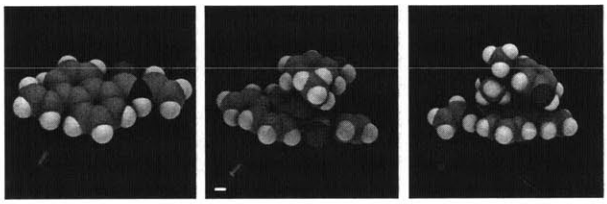
E6–ligand adsorption configurations, as shown in Figure 6-15(c). Physically, a review of trajectories shows that the allenyl groups on either side of the *m*-phenyl ring, and especially the $-O-CH=C=CH_2$ subsequence on the ligand's free end, constituted inflexible "arms" that prevented either the E2 or E6 from laying flat against the ligand's phenyl core.



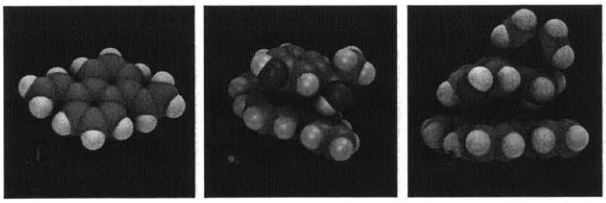
(a) Ligand candidate with design 1-(CH2=CH-),5-(CH3-)-napthalene (exp. I, gen. 76, cand. 1), with fitness score 0.87 ± 0.15 kcal/mol.



(b) Ligand candidate with design Cl-(*o*)Ph-COO-(*o*)Ph-COOH (exp. I, gen. 7, cand. 45), with fitness score 3.1 ± 0.2 kcal/mol. Note that terminal chlorine atom is rendered in cyan (like carbon atoms) with a larger radius.



(c) Ligand candidate with design 1-(CH2=CH-),5-(-C(=O)NH-CH=CH2)-napthalene (exp. II, gen. 23, cand. 31), with fitness score 0.79 ± 0.14 kcal/mol.



(d) Ligand candidate with design 1-(CH2=CH-),5-(CH2=CH-)-napthalene (exp. II, gen. 45, cand. 1), with fitness score 0.88 ± 0.13 kcal/mol.

Figure 6-12. Minimum-energy configurations from simulations of four successful ligand candidates from experiments I and II. Ligand is pictured alone (*left*), with E2 (*center*), and with E6 (*right*).

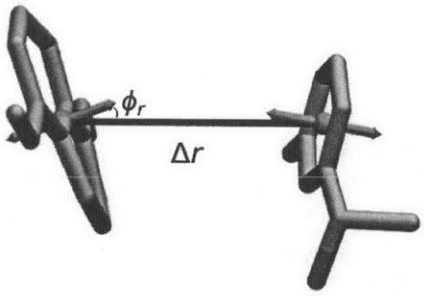
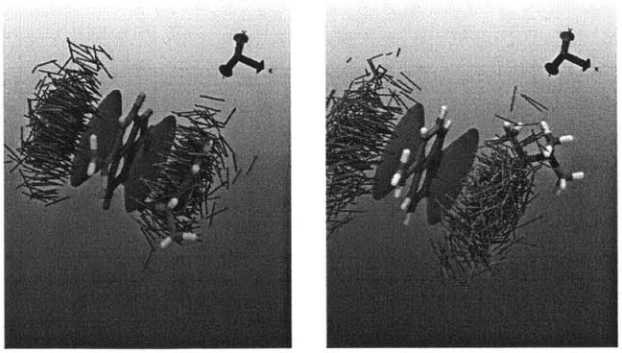
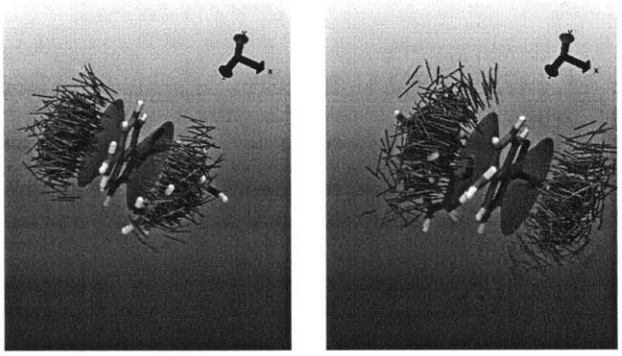


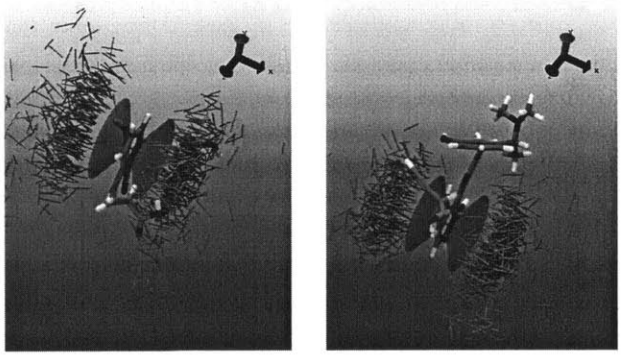
Figure 6-13. Illustration of bond orientation and relative orientation angles for two phenyl rings. The scheme was adapted from Ref 154.



(a) Ligand design CH_2=CH−($C_{10}H_6$)−CH=CH_2 (exp. II, gen. 45, cand. 1, with fitness score 0.84 ± 0.13 over 1.1 μs evaluation.

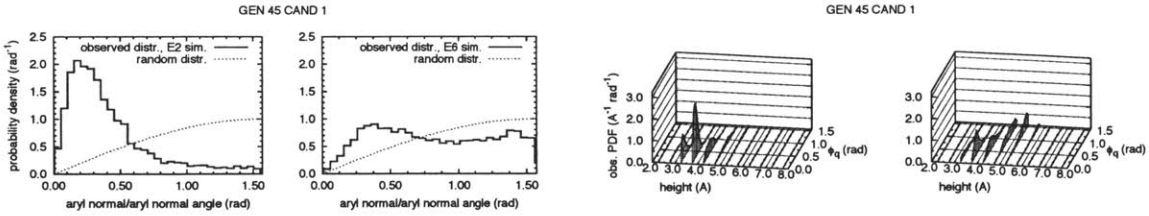


(b) Ligand design CH_2=CH−($C_{10}H_6$)−CO−NH−CH=CH_2 (exp. II, gen. 23, cand. 31), with fitness score 0.94 ± 0.14 over 168ns evaluation.

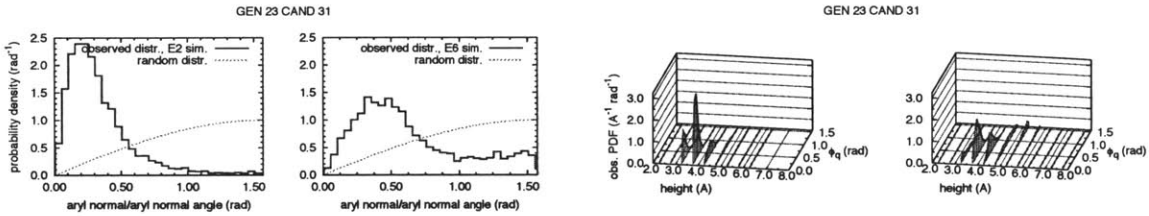


(c) Ligand design CH#C−CH(COOH)−CH=C=CH−(o)Ph−CO−CH=C=CH (exp. II, gen. 3, cand. 20), with fitness score -0.36 over 9 ns of production.

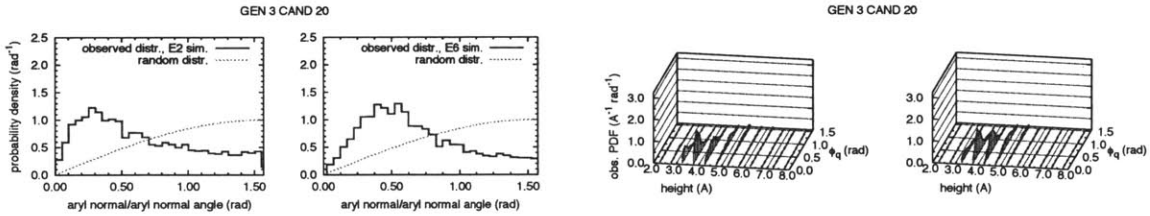
Figure 6-14. Alignment of E2 (*left*) and E6 (*right*) molecules with three different ligand designs, for states in which $\phi_c < \phi_{\text{cutoff}} = 60°$. The ligand is pictured in the center, and purple arrows represent the normal vector to the phenyl ring of the adsorbed (E2 or E6) molecule at each recorded configuration. The ligand design in subfigure (c) is a non-positive control. For clarity, only 1,000 such arrows are shown in each rendering.



(a) Ligand design CH_2=CH–($C_{10}H_6$)–CH=CH_2 (*selective*; exp. II, gen. 45, cand. 1).



(b) Ligand design CH_2=CH–($C_{10}H_6$)–CO–NH–CH=CH_2 (*selective*; exp. II, gen. 23, cand. 31).



(c) Ligand design CH#C–CH(COOH)–CH=C=CH–(o)Ph–CO–CH=C=CH (*non-selective*; exp. II, gen. 3, cand. 20).

Figure 6-15. Histogram of relative orientation ϕ_q (*left*) and histogram of relative orientation ϕ_q as a function of separation height h. As before, these measurements are restricted to states in which the relative orientation angle $\phi_c < \phi_{\text{cutoff}} = 60°$. The ligand design in subfigure (c) is a non-positive control.

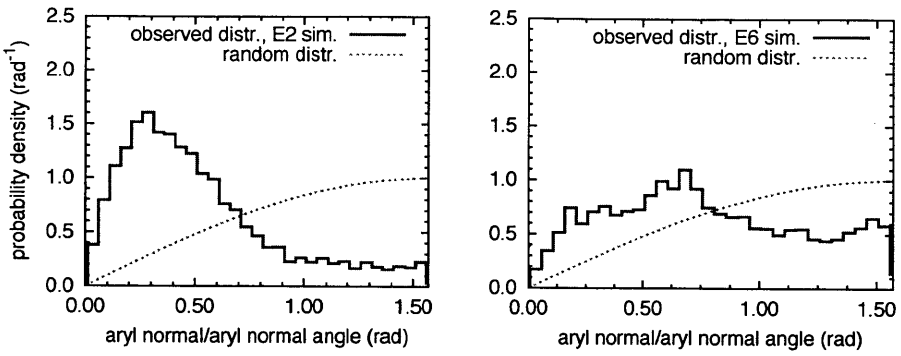
6.3.1 Top-scoring ligand designs

Experiments III and IV were designed to take advantage of several practices we believed would improve the molecular evolution procedure: use of final coordinates as initial coordinates in repeated evaluations; accumulative scoring for ligand designs; population consolidation, in which duplicate designs were replaced by random designs (discussed in detail in the next section); and slightly shortened evaluation (compared to Experiment III), to enable frequent reproduction steps.

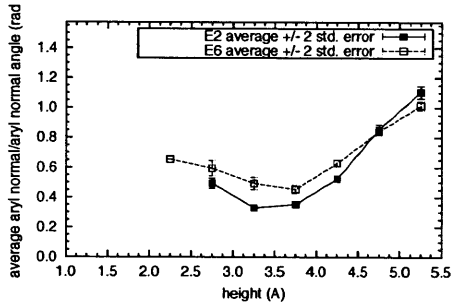
In these latter two experiments, the populations of designs (each viewed as a collective whole) did not surpass the populations in Experiments I and II, as can be seen by comparing the median and 80^{th} percentile scores in Figures Figures 6-7 through 6-10. But these latter two experiments did identify several ligand designs that made up the top-scoring ligands overall, including the top 12 scorers (Table 6-1).

These ligands, like those discussed above, contained internal naphthyl groups (Experiment III), or multiple phenyl rings (Experiment IV).[155] In all the top-scoring ligands from Experiments III and IV, the same planarity-based favorability to alignment and close adsorption of E2 was observed, as confirmed by $\langle \phi_q \rangle$ values at a near-close-contact distance of 3.25 Å(listed in Table 6-2) and alignment analyses of the kind shown above. These are shown in Figures 6-16, 6-17, and 6-18 for three ligands from Experiment III, and in Figures 6-19 through 6-22 for four ligands from Experiment IV.

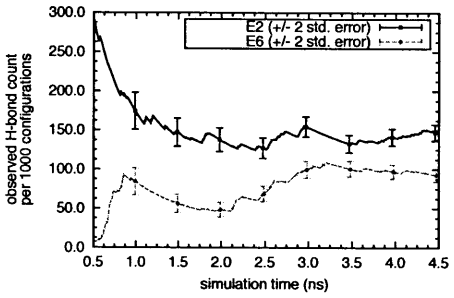
To understand why the ligands in these latter experiments obtained higher scores than the "solely planar" ligands in Experiments I and II, we analyzed the number of hydrogen bonds between the ligand and the E2 or E6 molecule. These ligands contain functional groups that accept, rather than donate, hydrogen bonds, like ether linkages, difluoro groups, and (to some extent) amide groups. As listed in Table 6-2, the ratio of the number of hydrogen bonds observed in E2 simulations to that observed in E6 simulations is between 1.5:1 and 2.5:1 for these high-scoring ligand designs. In addition, in E2 simulations, it was consistently observed that the ligand served as acceptor for over half of those hydrogen bonds.



(a) Histograms of relative orientation of E2 or E6's phenyl normal to normal of nearest aryl ring in ligand.



(b) Average relative orientation as a function of height from E2 or E6's phenyl center to plane of nearest aryl ring in ligand. Error bars are one standard error of the average.

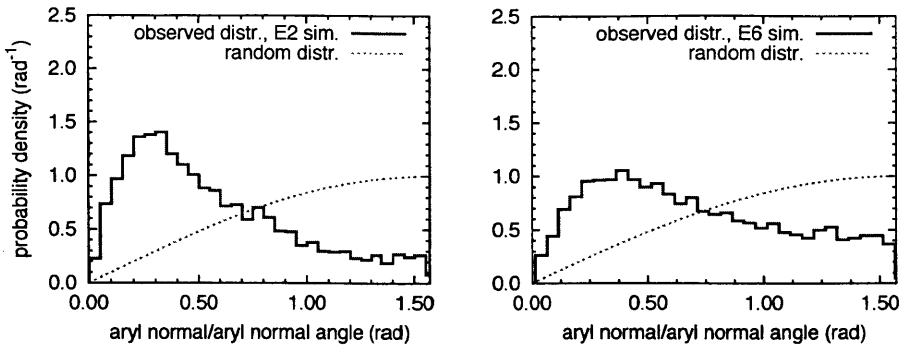


(c) Number of H-bonds observed per 1000 configurations among simulation frames recorded every 1 ps.

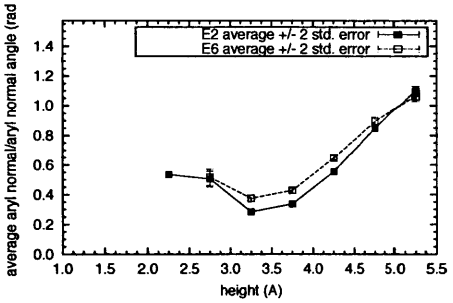
Figure 6-16. Alignment and hydrogen bonding analysis of ligand with sequence $CONH_2-C(CH_3)_2-C_{10}H_6-C_{10}H_6-C(CH_3)_2-H$ and selectivity score 3.02 ± 0.10 kcal/mol. The ligand design's final 4.5-ns evaluation (Experiment III, generation 69, candidate 20) was used for this analysis.

In fact, this hydrogen-bond acceptor motif was in agreement with our initial chemical reasoning: because the E2 molecule contains a hydrogen bond donor in its second hydroxyl group (which differentiates it from E6 and its tertiary amine), a ligand with accepting-only groups would exhibit favorable hydrogen bonds with E2 more frequently than E6, as noted above.

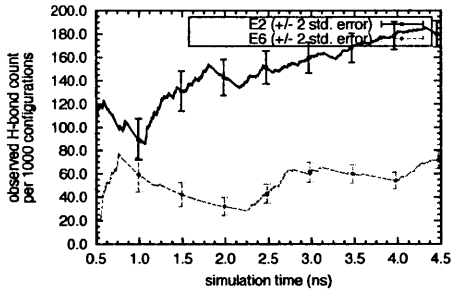
The negative controls listed in Table 6-2 have E2 alignment angles between 25 and 35°, in contrast to values from 14 to 20° in the selective designs. Additionally, differences between the average alignment angle values for E2 and E6 are smaller for the non-selective control cases than in the selective designs. In three of the four control cases, the ligand makes many more hydrogen bonds with E2 than with E6



(a) Histograms of relative orientation of E2 or E6's phenyl normal to normal of nearest aryl ring in ligand.



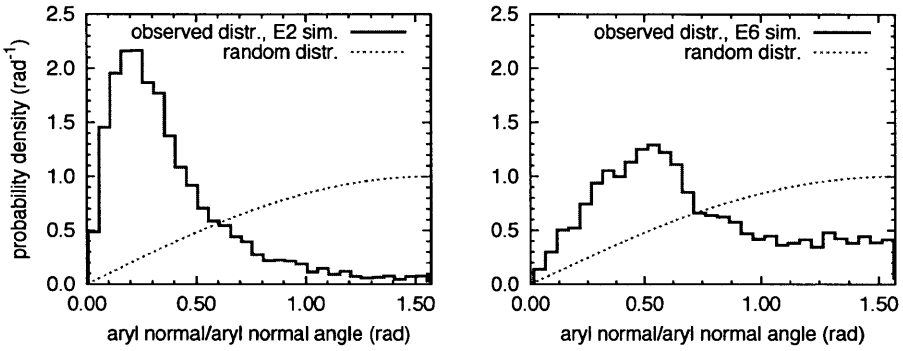
(b) Average relative orientation as a function of height from E2 or E6's phenyl center to plane of nearest aryl ring in ligand. Error bars are one standard error of the average.



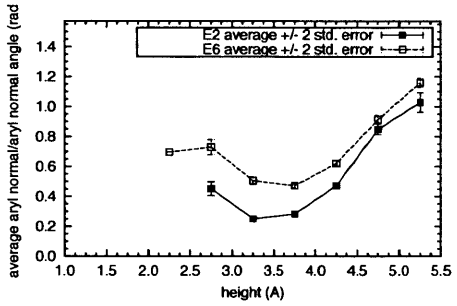
(c) Number of H-bonds observed per 1000 configurations among simulation frames recorded every 1 ps.

Figure 6-17. Alignment and hydrogen bonding analysis of ligand with sequence $CONH_2-C(CH_3)_2-C_{10}H_6-C_{10}H_6-C_{10}H_6-CH_3$ and selectivity score 2.57 ± 0.05 kcal/mol. The ligand design's final 4.5-ns evaluation (Experiment III, generation 84, candidate 12) was used for this analysis.

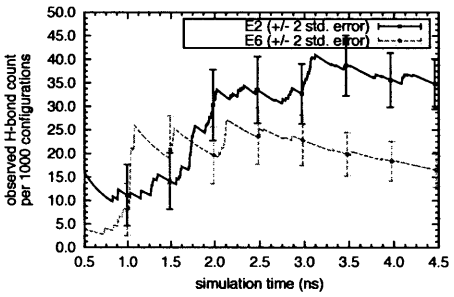
during the simulation, suggesting that a hydrogen bonding advantage alone does not necessarily confer energetic selectivity.



(a) Histograms of relative orientation of E2 or E6's phenyl normal to normal of nearest aryl ring in ligand.



(b) Average relative orientation as a function of height from E2 or E6's phenyl center to plane of nearest aryl ring in ligand. Error bars are one standard error of the average.



(c) Number of H-bonds observed per 1000 configurations among simulation frames recorded every 1 ps.

Figure 6-18. Alignment and hydrogen bonding analysis of ligand with sequence $CH_3-C_{10}H_6-$(trans)CH=CH$-$COOH and selectivity score 2.48 ± 0.05 kcal/mol. The ligand design's final 4.5-ns evaluation (Experiment III, generation 55, candidate 24) was used for this analysis.

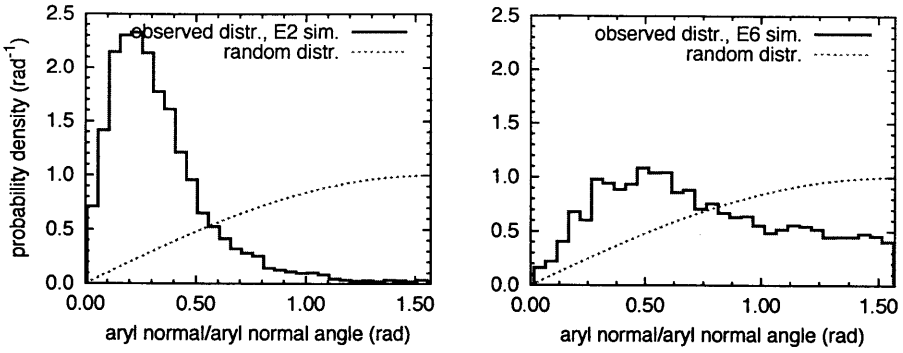
Table 6-2. Observed alignment and hydrogen bonding behavior of selected high-scoring ligands. Entries below the dotted line are non-selective controls containing aromatic groups.

sequence	score[a] (kcal/mol)	$\langle\phi_q\rangle_{h=3.25\ \text{Å}}$ (deg)[b] E2	E6	H-bonds observed[c] E2	E6
$CONH_2-C(CH_3)_2-C_{10}H_6-C_{10}H_6-C(CH_3)_2-H$	3.02 ± 0.10	19	28	652 (650)	440 (50)
$CONH_2-C(CH_3)_2-C_{10}H_6-C_{10}H_6-C_{10}H_6-CH_3$	2.57 ± 0.05	16	21	800 (211)	333 (63)
$CH_3-C_{10}H_6-$(trans)CH=CH−COOH	2.48 ± 0.05	14	31	149 (108)	75 (63)
COOH−(m)Ph−(m)Ph−Ph	3.67 ± 0.04	14	25	120 (100)	61 (61)
CH_3-(m)Ph−CH(COOH)−[(m)Ph]$_3$−O−$CH(CH_2CH_3)$−COOH	2.00 ± 0.10	14	29	192 (128)	159 (141)
CH_3-(m)Ph−CF_2−O−(m)Ph−Ph	1.42 ± 0.09	20	28	98 (98)	3 (3)
CH_3-(m)Ph−CH(COOH)−(m)Ph−(m)Ph−CH_3	1.42 ± 0.05	18	35	183 (139)	155 (92)
HO−CH(iBut)−(m)Ph−Ph	0.04 ± 0.08	25	35	438 (246)	181 (46)
CH≡C−CH(COOH)−CH=C=CH−(o)Ph−CO−CH=C=CH	-0.36 ± 0.12	28	41	315 (165)	154 (37)
F−$CH(CH_2CH_3)$−NH−NH−C(Ph)H−SH	-0.47 ± 0.14	35	41	1479 (512)	610 (96)
Cl−CH(iBut)−(p)Ph−S−(trans)CH=CH−CH(COOH)−OH	-0.84 ± 0.15	30	34	441 (328)	427 (280)

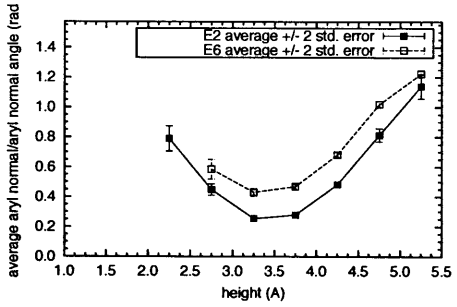
[a] Including score penalty.

[b] Statistical errors of these average values were typically about 1°.

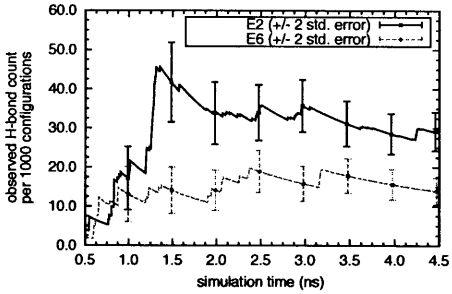
[c] Total number of hydrogen bonds observed in a 4.5-ns simulation, with configurations recorded every 1 ps. Value in parentheses is number of instances in which the ligand was the hydrogen bond *acceptor*, and the E2 or E6 molecule was the donor.



(a) Histograms of relative orientation of E2 or E6's phenyl normal to normal of nearest aryl ring in ligand.

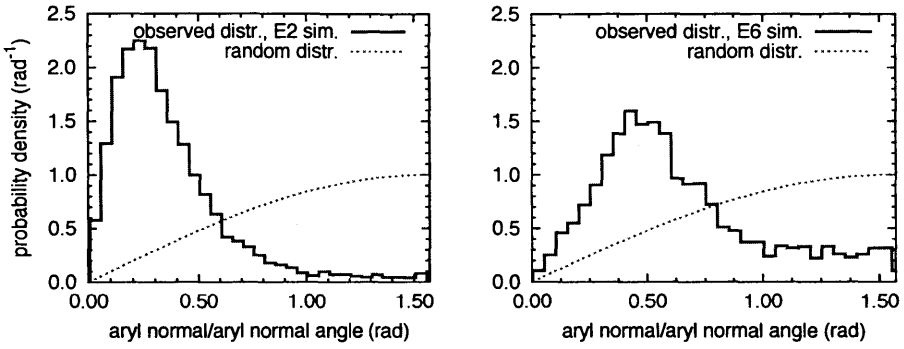


(b) Average relative orientation as a function of height from E2 or E6's phenyl center to plane of nearest aryl ring in ligand. Error bars are one standard error of the average.

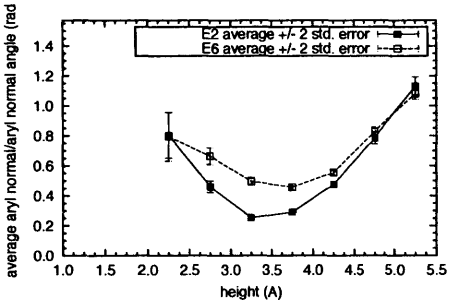


(c) Number of H-bonds observed per 1000 configurations among simulation frames recorded every 1 ps.

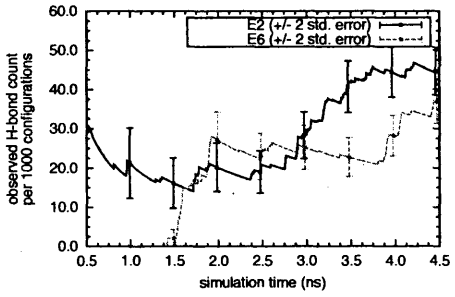
Figure 6-19. Alignment and hydrogen bonding analysis of ligand with sequence COOH−(m)Ph−(m)Ph−Ph and selectivity score 3.67 ± 0.10 kcal/mol. The ligand design's final 4.5-ns evaluation (Experiment IV, generation 68, candidate 20) was used for this analysis.



(a) Histograms of relative orientation of E2 or E6's phenyl normal to normal of nearest aryl ring in ligand.

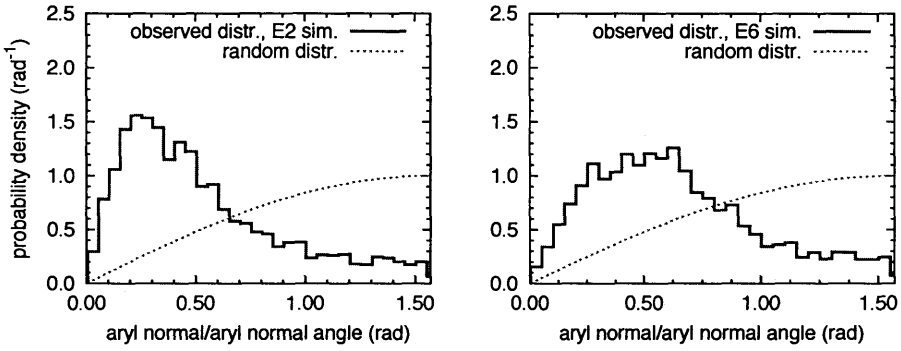


(b) Average relative orientation as a function of height from E2 or E6's phenyl center to plane of nearest aryl ring in ligand. Error bars are one standard error of the average.

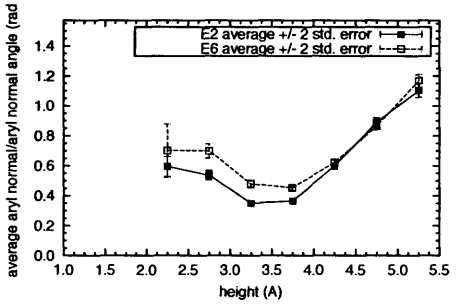


(c) Number of H-bonds observed per 1000 configurations among simulation frames recorded every 1 ps.

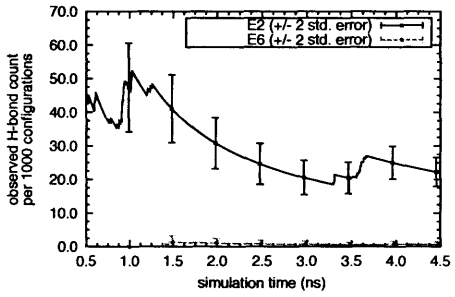
Figure 6-20. Alignment and hydrogen bonding analysis of ligand with sequence CH_3–(m)Ph–CH(COOH)–(m)Ph–(m)Ph–(m)Ph–O–CH(CH_2CH_3)–COOH and selectivity score 2.00 ± 0.10 kcal/mol. The ligand design's final 4.5-ns evaluation (Experiment IV, generation 44, candidate 37) was used for this analysis.



(a) Histograms of relative orientation of E2 or E6's phenyl normal to normal of nearest aryl ring in ligand.

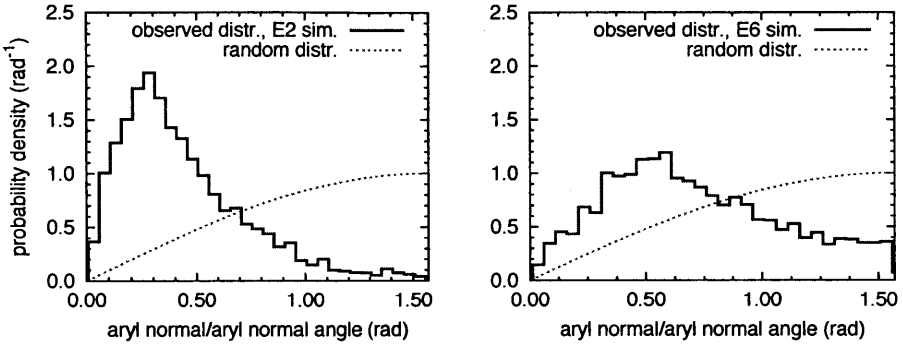


(b) Average relative orientation as a function of height from E2 or E6's phenyl center to plane of nearest aryl ring in ligand. Error bars are one standard error of the average.

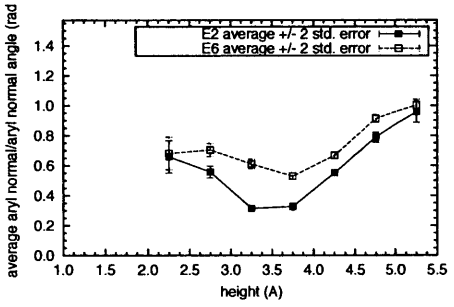


(c) Number of H-bonds observed per 1000 configurations among simulation frames recorded every 1 ps.

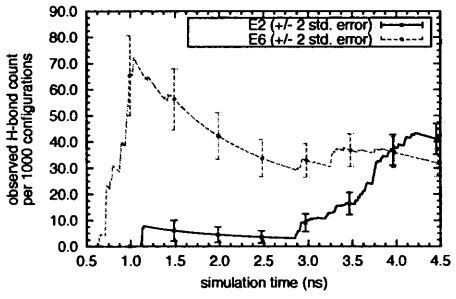
Figure 6-21. Alignment and hydrogen bonding analysis of ligand with sequence CH_3-(m)Ph$-CF_2-$O$-$(m)Ph$-$Ph and selectivity score 1.42 ± 0.09 kcal/mol. The ligand design's final 4.5-ns evaluation (Experiment IV, generation 62, candidate 16) was used for this analysis.



(a) Histograms of relative orientation of E2 or E6's phenyl normal to normal of nearest aryl ring in ligand.



(b) Average relative orientation as a function of height from E2 or E6's phenyl center to plane of nearest aryl ring in ligand. Error bars are one standard error of the average.



(c) Number of H-bonds observed per 1000 configurations among simulation frames recorded every 1 ps.

Figure 6-22. Alignment and hydrogen bonding analysis of ligand with sequence CH_3–(m)Ph–CH(COOH)–(m)Ph–(m)Ph–CH_3 and selectivity score 1.42±0.05 kcal/mol. The ligand design's final 4.5-ns evaluation (Experiment IV, generation 42, candidate 28) was used for this analysis.

6.4 Evolution dynamics and effect of fitness uncertainty

To understand the dynamics of molecular evolution, we employed the "selection intensity" paradigm of Muhlenbein and coworkers. The intensity of the selection approach in a genetic algorithm is "the expected average fitness of a population after selection is performed on a population whose fitness is distributed according to the unit normal distribution $N(0,1)$".[146,156] If the fitness of a population at generation t is normally distributed as $N(\mu_t, \sigma_t^2)$, then the expected value of the mean fitness at generation $t+1$ can be related to the selection intensity I:

$$\mu_{t+1} = \mu_t + I\sigma_t$$

Values of the selection intensity depend on the selection scheme, *i.e.* how members of a parent generation are selected to reproduce. The selection intensity values for the techniques used in this study are listed in Table 6-3 below. In deriving this relationship, it was assumed that all improvement in a populations' fitness comes from *selection*; that is, when a population's fitness is distributed normally as $N(\mu_t, \sigma_t)$, the parental subset will have, on average, a mean fitness $\mu_t^P = \mu_t + I\sigma_t$, and the subsequent generation, will have the same fitness level as the parental subset.[146]

Table 6-3. Selection intensity of selection schemes used in this work. Adapted from Ref. 146.

selection scheme	selection intensity I	experiments where employed
roulette wheel or proportionate	σ_t/μ_t	Exp. I and II
deterministic tournament, size s	$\mu_{s:s}$[a]	Exp. III
fuzzy tournament, size s	$\mu_{s:s}$	Exp. IV

[a] The order statistic $\mu_{n:k}$ is the expected value of the k^{th} ordered (largest) of n values sampled from a unit normal distribution. According to a standard reference,[157] the value of $\mu_{2:2}$ for a two-member tournament is 0.564.

6.4.1 Changes in driving force for population improvement

This framework makes it possible to understand our evolution results, and in particular the reason that the rate of improvement of fitness scores (per generation) diminishes as evolution is carried out. In this study, fitness scores plateaued in the late stages of evolution in all four experiments. In Experiments I and II, as evolution proceeded, the populations grew more homogeneous in terms of members' phenotypic properties (as shown in the top panels of Figures 6-7 and 6-8), *and* in terms of their fitness scores (fitness score variance not shown in those figures).

To understand how the state of the population impacts evolution dynamics, the standard deviation of the fitness score in each generation was calculated. Under the analysis above, this variation in fitness scores can be thought of as a "driving force" behind selection-based improvements in evolution. Looking at Experiment II as an example, as evolution proceeded, the mean fitness score increased, and as the top panel of Figure 6-23(a) suggests, the absolute value of the population's score standard deviation σ_t decreased by about 40% (from 0.5 to 0.3 kcal/mol). One reason for this decrease was the population's becoming more homogeneous in phenotypic terms, as suggested by the decreasing diversity shown in Figure 6-8. A second possible reason for the decreasing fitness score variance could be that the population is reaching a region of genotype space in which mutations and crossover operations do not produce improved offspring, and therefore do not take hold—the discrete-genome equivalent of the population falling into a narrow local optimum in the fitness landscape, in which children produced through genetic operations would be located "away" from the neighborhood of the optimum and have low fitness score.

Referring to the equation above, this decrease in diversity (as measured by fitness score variance) would tend to slow the rate of population improvement, if selection intensity I were held constant, since $\mu_{t+1} = \mu_t + I\sigma_t$. But in this case, roulette wheel selection (also called proportionate selection) was used, so the selection intensity $I = \sigma_t/\mu_t$ *also decreased*, as shown in the middle panel of Figure 6-23(a), because the fitness landscape, as explored by the evolution-guided population, grew narrower

as fitness increased. This has a further, interacting effect of slowing the rate of improvement in the evolution process. This may explain why the fitness score plateaued at a value of 0.85 kcal/mol (for the median) in Experiment II: by generation 34 and thereafter, the selection intensity I had decreased from its initial value of 2.5 to about 0.25, so that the product $I\sigma_t$ decreased to roughly $\frac{1}{10} \times \frac{1}{2} \approx 5\%$ of its original value.

The function fitted to values of σ_t/μ_t in the middle panel of Figure 6-23(a) was used to predict the expected fitness score improvement $\mu_{t+1} - \mu_t = \frac{\sigma_t}{\mu_t}\sigma_t = \left(\frac{\sigma_t}{\mu_t}\right)^2 \mu_t$. This is shown as the solid blue curve in the bottom panel. In a general, for a function of a continuous variable in the vicinity of a local optimum, as fitness increases, the space available for a population decreases. If, analogously, there are fewer discrete genotypes with higher fitness scores than with lower, as would be expected for a challenging discrete optimization problem, this presents a problem for GAs employing roulette wheel selection: unless σ/μ decreases proportionally to $\sqrt{\mu}$ or more slowly—which is a property of the underlying fitness landscape, as explored by the GA—the product $I\sigma$ will shrink as μ increases, leading to a stalling of the selection-based evolutionary improvement

Diminshing fitnesss variance presents a similar, albeit less pressing, problem for evolutionary procedures employing tournament selection, like Experiments III and IV. In these cases, however, the value of the selection pressure I is constant, so that $I\sigma_t$ decreases less severely (as μ_t increases) than in the proportionate case. One possible comparison of these selection schemes in Experiments I through IV is to count the number of generations required to increase the mean fitness score of 0.35 to 0.70 kcal/mol; the lower cutoff was chosen to be slightly higher than each experiments' initial mean score. This improvement required 20 generations in Exp. I; 15 in Exp. II; 10 in Exp. III; and 15 in Exp. IV, so that when the MD production time in each set of simulations is accounted for, Experiments III and IV initial rate of fitness score improvement was greater than that of Experiments I and II.

6.4.2 Noise in fitness function evaluation

Another complication when performing MD using automated molecular dynamics simulations is that thermodynamic measurements made from such simulations include statistical errors. In this work, the typical estimates of statistical error for the fitness function, due to finite sampling, are 0.15 to 0.20 kcal/mol for 4.5-ns production simulations. For comparison, the range of variation of fitness scores in initial, randomly-generated populations is about 0.6, as measured by interquartile range.

Miller and Goldberg[146,158] analyzed how evolution dynamics would differ when fitness functions include "noise" from physical processes, measurement imprecision, or limited sampling. They considered a fitness score f' which is the sum of a true fitness score f plus noise:

$$f' = f + \text{noise}$$

In modeling evolution dynamics, they assumed that the true fitness function was normally distributed as $N(\mu_t, \sigma_t^2)$ at generation t, and that the noise was unbiased and normally distributed as $N(0, \sigma_N^2)$. They showed that under these assumptions, the evolution dynamics would be impeded by the inclusion of noise. The expected value of μ_{t+1}, the mean of the fitness score in the next generation would be:

$$\mu_{t+1} = \mu_t + I\left(\frac{\sigma_t}{\sqrt{\sigma_t^2 + \sigma_N^2}}\right)\sigma_t$$

In this noisy case, the selection-based improvement in evolution (usually equal to $I\sigma_t$) in each generation is diminished by a factor $\left(\frac{\sigma_t}{\sqrt{\sigma_t^2+\sigma_N^2}}\right)$ which can be thought of as a signal-to-(signal plus noise) ratio.

In this case, as the population in Experiment III (discussed below) evolved from generation 1 to generation 35, the standard deviation of the fitness scores in the population decreased from about 0.6 to about 0.1 or 0.2 (again, see Figure 6-23), while the "noise" level resulting from finite sampling remained approximately constant. These values correspond to diminishment in the effective selection intensity ranging from 30% (initially) to 60% (at end of evolution), compared to hypothetical noise-free

evaluations.

6.4.3 Genetic algorithm performance

The four experiments performed in this study performed allow us to compare the performance of the genetic algorithm, using different selection techniques and evolution parameters.

For example, Experiments I and II differed by the amount of MD production time used (3.0 and 6.0 ns, respectively). As in any simulation, the longer production time leads to smaller statistical errors, and a greater likelihood of each ligand evaluation being near its true, long-run-average value. Because of this improved consistency, the GA was able to achieve a score-based convergence, defined as a median score of 0.75 kcal/mol or higher, at 19 generations, compared to 24 in Experiment I. But because of the longer production time, the 19 generations in Experiment II required computer time equivalent to 38 generations in Experiment I.

This suggests that when decreased statistical error can be achieved by performing longer MD (or more generally, using additional computational resources to increase accuracy), doing so will not necessarily accelerate convergence, when measured in total simulation time. A genetic algorithm, as a stochastic optimization technique, can accept some "noise" in function evaluation values, as discussed above; the most important aspect of evaluation is that it consistently (and correctly) rank each generations' members. With shorter evaluation times, the GA can more frequently select and propagate successful designs, and introduce genetic operations like crossover and mutation.

Experiments III and IV used different techniques for evolution (two-member tournament selection), and included accumulative scoring, in which an evaluation in the current generation was averaged with all previous production MD (see Table 5-3). This was done with the aim of increasing the consistency of evaluation, in a manner consistent with physical reasoning.

In addition, motivated by the fact that homogeneous populations tended to decrease the rate of score improvement in a genetic algorithm, we implemented *consol-*

idation, in which duplicate copies of a ligand design in the population were deleted, and replaced by randomly-generated ligand designs. The rationale for this step was to keep the value of σ_t high, in order to prevent this driving force for improvement from diminishing.[159] As noted above, allowing multiple copies of a design to exist in a population was designed to enable convergence; in a similar way, consolidation was a way to step back from impending convergence when further exploration of the chemical space was desired.

The effects of consolidation in Experiment III can be seen in Figure 6-23(b). Before consolidation, the value of and $\frac{\sigma_t}{\mu_t}$ as a function of μ_t is similar to those observed in Experiment II. (The quantity $\frac{\sigma_t}{\mu_t}$ is used for comparison because it seems to exhibit less variation than σ_t) .The first consolidation was effected after generation 35, because the population diversity had dropped sharply. After consolidation, the $\frac{\sigma_t}{\mu_t}$-versus-μ_t curve was, in effect, shifted to the right, *increasing* the value of $\frac{\sigma_t}{\mu_t}$ at any given μ_t. This resulted in an immediate increase in population diversity and an immediate downward shift in the population's fitness distribution, due to the introduction of random designs; however, subsequent selection and genetic operations led to a fast increase in fitness scores (generations 36 to 41 in Figure 6-9), although the median scores then returned to approximately the same level (approx. 0.88 kcal/mol) as before consolidation.

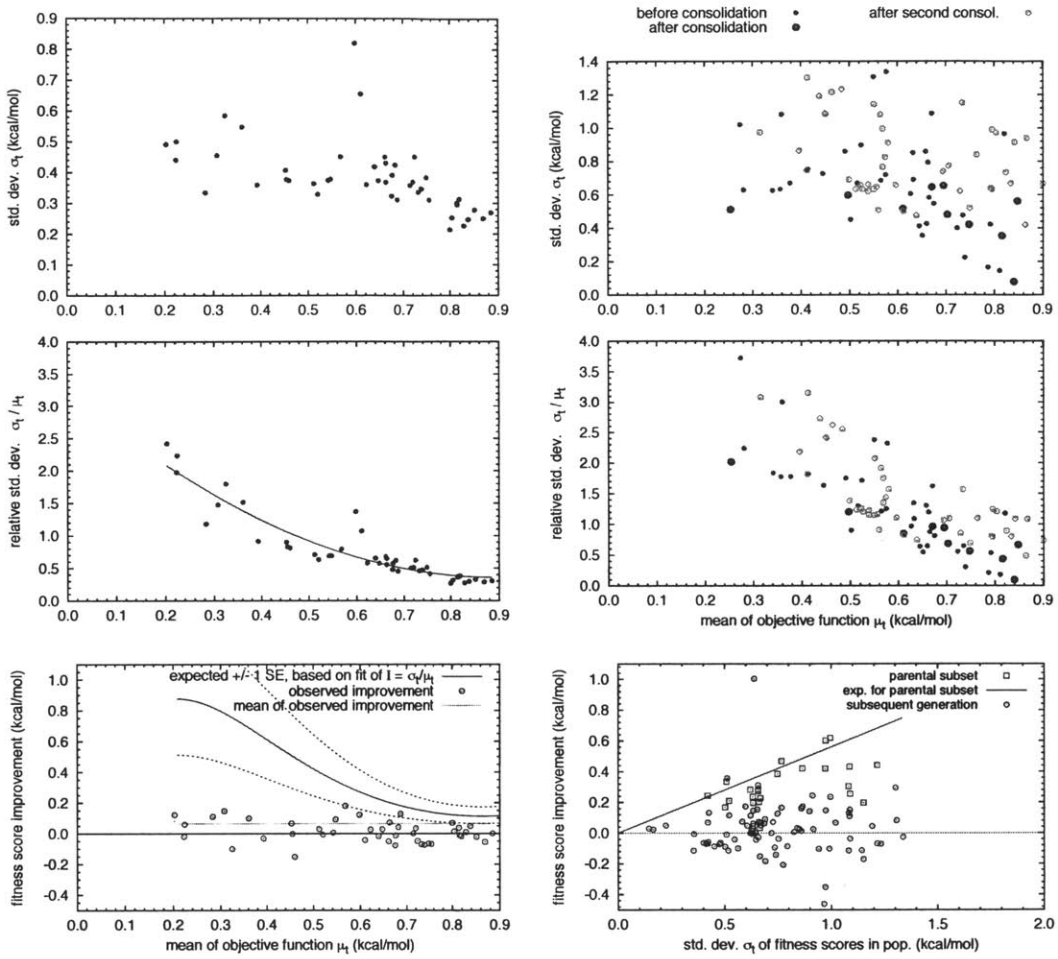
The bottom panel of Figure 6-23(b) illustrates another challenge in the molecular evolution process. It shows the fitness of the "parental subset" in generations with different diversity of fitness score, measured by standard deviation. In the analysis of evolution dynamics discussed above, Miller and co-workers assumed that all fitness improvement would come from *selection*, and that reproduction steps would be fitness-neutral; that is, the children of every parental subset would have, on average, the same fitness as their parents: $\mu_{t+1} = \mu_t^P$.

The bottom panel of Figure 6-23(b) shows that, indeed, the parental subset did constitute an improvement upon the population's fitness (measured by average scores), and that the selection-based improvement was greater when the fitness diversity was larger. The solid line represents the theoretical relationship based on

normally-distributed fitness scores, and the blue points fall around, and slightly below, that line.

The challenge presented is that the children of these parental subsets *do not* display the same fitness as their parents, as shown by the red points in the bottom panel of Figure 6-23(b). This is likely the result of the discrete representation of chemical species as a string of enumerated functional groups: any change, such as a single-point mutation or an insertion, could introduce significant changes in physicochemical properties, which could result in diminished fitness. In the "Conclusions and Outlook" section below, we propose changes that could reduce the impact of this issue.

A second possible explanation is that the selective ligands, which are the solutions obtained by the molecular evolution scheme, are topologically "fragile." That is, they may achieve selectivity through specific combinations of functional groups; for example, the ligand may bind E2 strongly by presenting a large hydrophobic surface to interact with E2's phenyl core, and provide one or two hydrogen bond acceptors to interact with E2's two hydroxyl groups, when correctly positioned by small functional groups between them. Once these combinations are disturbed by genetic operations like crossover or deletion, the child ligands can be much less selective.



(a) Experiment II. The expected fitness score improvement (*bottom panel*) for roulette wheel selection is $I\sigma_t = \frac{\sigma_t}{\mu_t}\sigma_t$.

(b) Experiment III. The expected fitness score improvement (*bottom panel*) for two-member tournament selection is $I\sigma_t = (0.56)\sigma_t$. Note that the lower panel uses a different abscissa from the corresponding panel for Experiment II.

Figure 6-23. Evolution dynamics in Experiments II (45 generations) and III (88 generations): standard deviation and relative standard deviation of fitness scores in evolving population, as evolution proceeded and mean fitness score increased.

Chapter 7

Conclusions and outlook for future work

The conclusions from each of the applications of interfacial modification design are presented in the two sections below, along with my outlook for possible future work.

7.1 Mechanistic understanding of evaporation

In this study, the minimum free energy path for evaporation (under the eight restrained interfacial order parameters) was determined. During evaporation, the evaporating molecule sheds its second donated and second accepted hydrogen bonds, and rotates (relative to the interfacial normal) into a position unlike most water molecules in the interfacial region. It then loses its remaining donated hydrogen bond, and then loses its final accepted hydrogen bond at a time that its two hydrogen atoms are pointing outwards. For details, see Figure 4-4 on page 65, and description on page 63).

During this evaporation process, the orientation of nearby molecules relative to the evaporating molecules becomes *less aligned*, as measured by dipole-dipole angle η. In particular, the mean dipole-dipole angles shifts values of partial alignment, at $\sim 60°$, to an partially anti-aligned state (average value $\sim 60°$).

Using Voronoi milestoning, the evaporation process was found to take place on

a plateau-like free energy landscape, with $\Delta F = 7.4$ kcal/mol for the SPC/E water model. The mean first-passage time for evaporation was found to be 1375 ns for an individual molecule. This corresponds to an evaporation coefficient of $\gamma_E = 0.24$.

The directions in which order parameters most varied were analyzed, by examining contributing trajectories collectively in the portion of the string where FE and MFPT values changed, and by examining those trajectories directionality on an individual basis. These analyses suggested that the relative z-position, the orientation of nearby water molecules, and the number of hydrogen bonds accepted and donated (q_0^z, q_2^{ravg}, q_7^{acc}, and q_6^{don}, respectively) order parameters that changed most in this important region of the string.

When the MFPT values were regressed against the OP values at each milestone, the orientation of nearby molecules (q_2^{ravg}) and the number of accepted hydrogen bonds (q_7^{acc}) appeared most frequently in the combinations of OPs as explanatory variables with best BIC values.

Together, these results suggest that the loss of accepted hydrogen bonds, and the reorganization of the first solvation shell, play a critical role in the evaporation process. This conclusion would be consistent with the those of Cappa *et al.*,[16] cited above, which were based on strong isotopic dependence of the evaporation coefficient.

Based on the understanding of the evaporation mechanism suggested by the above conclusions, specific features of an additive to impede evaporation are put forward below.

Finally, we note that future work could include refinements of the order parameters used in this study. For example, one could measure the z-coordinate of an evaporating water molecule with respect to a calculated instantaneous or time-averaged interfacial surface,[160] rather than using a fixed position. Recently-published general order parameters for molecular crystals[161] could be used to measure the location and orientation of a water molecule's individual neighbors,[162] rather than measuring the mean and variance of nearby molecules' orientation, as was done here. And the unusual oxygen–oxygen or hydrogen–hydrogen distances which played a role in evaporation in a previous study[37] could be examined further as order parameters using a tech-

nique other than the SMCV, because the string method and milestoning would not be appropriate for a case in which kinetic energy propels a system along a reaction path.

7.2 Design of surface-bound molecules for selective separation

We developed a molecular evolution approach that optimizes molecular structures using a genetic algorithm. Our FORM2GEOM software can construct three-dimensional structures for molecules which are described as sequences of functional groups; these structures can then be used for automated molecular dynamics evaluation of thermodynamic properties. We then applied selection and genetic operations to generate new molecular designs from the most suitable designs in an initial population.

We applied this technique to a particular separation problem, namely the removal of an unconverted reactant from an API solution in a pharmaceutical manufacturing process. This separation was particularly challenging because the two molecules were structurally similar, differing by a single functional group.

The selectivity energy estimates (*i.e.* $\Delta\Delta E_{ads}$) obtained from our simplified simulations for the top-scoring molecular designs were in the range 0.60 to 1.60 kcal/mol, corresponding to separation factors of approximately 3 to 15, if the entropic component $\Delta\Delta S_{ads}$ is negligible. These top-scoring designs were selective because they contained planar regions consisting of a naphthalene or phenyl core, with attached groups containing sp^2 atoms. These planar regions allowed the E2 molecule to lay its phenyl core against the surface-bound ligand, while the E6 molecule was prevented from doing so by its bulky dimethyl amine group. This mechanism of preferential adsorption of E2, and the ligand motifs to achieve it, had not been anticipated by chemical intuition. Hydrogen bond-accepting groups in the ligand enhanced its selectivity, as had been anticipated.

More generally, the ability to quickly identify potentially selective surface-bound

ligands could enable economically viable development of adsorption-based separation processes. Such processes could obviate the need for alternative separation techniques that are more energy- and time-intensive like crystallization or solvent exchange, and could be configured for continuous manufacturing.

In the course of applying this GA-based molecular design procedure, we learned about its performance. We observed that shorter MD-based evaluations can lead to faster convergence of the GA to a homogeneous population, despite their more uncertain fitness values, when measured in computer time. We also saw that the fitness improvements effected by the genetic algorithm depended on the population diversity, as is commonly known in the genetic programming field, but that conventional analyses of fitness improvement do not necessarily apply to molecular evolution, because the offspring of two succesful parents often exhibits fitness lower than either parent's. This is likely a result of the necessarily discrete description of a molecule's constituent functional groups or atoms, and adjusting the genetic encoding of molecular topology to mitigate this problem is a priority for future work, as discussed below.

Based on the work in this study, we believe the molecular evolution approach could be improved and applied to new problems in molecular design. That is, improvements could be made to the evolution process, and in identifying new applications in which efficient *in silico* screening can be used to solve challenging design problems.

To improve the molecular evolution methodology, the molecular genome could be expanded in descriptive capability: it could describe functional groups using a base-2 string and the Gray encoding,[163] which would then enumerate functional groups as an ordered list (by chemical characteristics), and allow the algorithm to change functional groups in an incremental way. An expanded base-2 genome would also provide a natural way to add other, non-topological information, such as molecular conformations;[164] lattice spacings[165] or crystal habit (for solid systems); or composition or concentrations (for solution-based systems). To explore chemical space more broadly, rather than focusing on roughly linear molecules, it will also be helpful to develop a tree-based molecular genome with branching, as in some studies in Table 1-1.[66,69,73] And of course, the set of functional groups that serve as building blocks

of molecules can be expanded to include hetero-aromatic groups, amino acids, and other groups. In addition, the chemical space could be preemptively pruned through the use of "molecular abstraction," in which rough evaluations are used to eliminate entire sets of designs.[54]

Within our work, we saw tradeoffs between the tendency of a GA to broadly explore the chemical space, and its tendency to converge to a particular solution or related solutions. In our case, we used consolidation to address this practical challenge: that is, we added randomly-generated members to "temper" the population when it was near convergence to genetic homogeneity. In future work, more systematic approaches like selection techniques with adjustable intensity, or island models with migration,[166] could be used to achieve the right balance between converging the optimization problem, and efficiently exploring the chemical space.

Of course, the molecular evolution procedure can produce useful results to the extent that its underlying simulation-based evaluations correspond to experimentally-observed, real behavior. In future work, we plan to include such validation, and to compare the outcome of molecular evolution with molecular designs obtained through more traditional approaches.

In future work, exploring other selection/sampling techniques may also be worthwhile—in particular, applying stochastic universal sampling[167] as a selection technique, which is know to circumvent the problem of premature "takeover" of a population by successful members through bias-free sampling over the entire range of member fitnesses.

As noted above, first-principles-based computer simulations (like *ab initio* and empirical electronic structure calculations, or MD/MC with molecular force fields) are particularly well-suited for challenging problems in molecular design, since they exhibit no "bias" toward a particular kind of solution, or particular mechanisms or physical processes that are believed to be important in the design problem. If the improvements listed above were made to the genomic representation of chemical species, so that non-linear molecules could be included in the optimization search space, then any property that could be evaluated in an automated way could be screened for and optimized. Potential applications for this molecular design approach could then in-

clude: (*i*) solution additives for viscosity modification, solubility enhancement,[168,169] or other property modification;[170–172] (*ii*) organic molecules/materials with desireable electronic properties like small band gap or nonlinear optical response;[78,79] (*iii*) chelation or binding with small molecules; (*iv*) protein docking and drug design; (*v*) novel materials for drug delivery;[173] and (*vi*) catalysis[174–176] .

Chapter 8

References

1. Humphrey, W.; Dalke, A. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

2. Annable, T.; Cordwell, R.; Ewing, P. Demands on drops: ink-jet technology in industrial applications poses challenges to both ink formulations and pigments. *Eur. Coatings J.* **2006**, *44*, 44–50.

3. Holl, Y.; Keddie, J. L.; McDonald, P.; Winnik, W. In *Film formation in coatings: Mechanisms, Properties and Morphology*; Provder, T., Urban, M. W., Eds.; ACS Symposium Series; American Chemical Society, 2001; Vol. 790; Chapter 1, pp 2–29.

4. Chayen, N. A novel technique to control the rate of vapour diffusion, giving larger protein crystals. *J. Appl. Crystallogr.* **1997**, *30*, 198–202.

5. Talreja, S. et al. Screening and optimization of protein crystallization conditions through gradual evaporation using a novel crystallization platform. *J. Appl. Crystallogr.* **2005**, *38*, 988–995.

6. Chayen, N. E.; Saridakis, E. Protein crystallization: from purified protein to diffraction-quality crystal. *Nat. Meth.* **2008**, *5*, 147–153.

7. Knudsen, M. Die maximale Verdampfungsgeschwindigkeit des Quecksilbers. *Ann. Phys.* **1915**, *352*, 697–708.

8. Mortensen, E. M.; Eyring, H. Transmission Coefficients for Evaporation and Condensation. *J. Phys. Chem.* **1960**, *64*, 846–849.

9. Jones, F. *Evaporation of Water: With Emphasis on Applications and Measurements*; CRC Press, 1992.

10. Rideal, E. The influence of thin surface films on the evaporation of water. *J. Phys. Chem.* **1925**, *29*, 1585–1588.

11. Alty, T. The maximum rate of evaporation of water. *Philos. Mag.* **1933**, *15*, 82–103.

12. Alty, T.; Mackay, C. The accommodation coefficient and the evaporation coefficient of water. *Proc. R. Soc. Lon.* **1935**, *A149*, 104–116.

13. Marek, R.; Straub, J. Analysis of the evaporation coefficient and the condensation coefficient of water. *Int. J. Heat Mass Trans.* **2001**, *44*, 39–53.

14. Davidovits, P. et al. Update 1 of: Mass Accommodation and Chemical Reactions at Gas-Liquid Interfaces. *Chem. Rev.* **2011**, *111*, PR76–PR109.

15. Li, Y. et al. Mass and thermal accommodation coefficients of H2O(g) on liquid water as a function of temperature. *J. Phys. Chem. A* **2001**, *105*, 10627–10634.

16. Cappa, C. et al. Isotope fractionation of water during evaporation without condensation. *J. Phys. Chem. B* **2005**, *109*, 24391–24400.

17. Smith, J. D. et al. Raman thermometry measurements of free evaporation from liquid water droplets. *J. Am. Chem. Soc.* **2006**, *128*, 12892–12898.

18. Winkler, P. M. et al. Condensation of water vapor: Experimental determination of mass and thermal accommodation coefficients. *J. Geophys. Res.–Atmos.* **2006**, *111*, D19202.

19. Davidovits, P. et al. Mass accommodation and chemical reactions at gas-liquid interfaces. *Chem. Rev.* **2006**, *106*, 1323–1354.

20. Faubel, M.; Kisters, T. Non-equilibrium molecular evaporation of carboxylic acid dimers. *Nature* **1989**, *339*, 527–529.

21. Wilson, M.; Pohorille, A.; Pratt, L. Molecular dynamics of the water liquid-vapor interface. *J. Phys. Chem.* **1987**, *91*, 4873–4878.

22. Wilson, M.; Pohorille, A.; Pratt, L. Surface potential of the water liquid–vapor interface. *J. Chem. Phys.* **1988**, *88*, 3281.

23. Matsumoto, M.; Kataoka, Y. Study on liquid–vapor interface of water. I. Simulational results of thermodynamic properties and orientational structure. *J. Chem. Phys.* **1988**, *88*, 3233.

24. Alejandre, J.; Tildesley, D.; Chapela, G. Molecular dynamics simulation of the orthobaric densities and surface tension of water. *J. Chem. Phys.* **1995**, *102*, 4574.

25. Taylor, R.; Dang, L.; Garrett, B. Molecular Dynamics Simulations of the Liquid/Vapor Interface of SPC/E Water. *J. Phys. Chem* **1996**, *100*, 711–720.

26. Shi, B.; Sinha, S.; Dhir, V. K. Molecular dynamics simulation of the density and surface tension of water by particle-particle particle-mesh method. *J. Chem. Phys.* **2006**, *124*, 204715.

27. Vassilev, P. et al. Ab initio molecular dynamics simulation of liquid water and water-vapor interface. *J. Chem. Phys.* **2001**, *115*, 9815–9820.

28. Kuo, I.; Mundy, C. An ab initio molecular dynamics study of the aqueous liquid-vapor interface. *Science* **2004**, *303*, 658–660.

29. Kuo, I. et al. Structure and dynamics of the aqueous liquid-vapor interface: A comprehensive particle-based simulation study. *J. Phys. Chem. B* **2006**, *110*, 3738–3746.

30. Wick, C.; Kuo, I.; Mundy, C.; Dang, L. The Effect of Polarizability for Understanding the Molecular Structure of Aqueous Interfaces. *J. Chem. Theo. Comput.* **2007**, *3*, 2002–2010.

31. Taylor, R.; Ray, D.; Garrett, B. Understanding the mechanism for the mass accommodation of ethanol by a water droplet. *J. Phys. Chem. B* **1997**, *101*, 5473–5476.

32. Taylor, R.; Garrett, B. Accommodation of Alcohols by the Liquid/Vapor Interface of Water: Molecular Dynamics Study. *J. Phys. Chem. B* **1999**, *103*, 844–851.

33. Vacha, R. et al. Adsorption of atmospherically relevant gases at the air/water interface: Free energy profiles of aqueous solvation of N2, O2, O3, OH, H2O, HO2, and H2O2. *J. Phys. Chem. A* **2004**, *108*, 11573–11579.

34. Vieceli, J.; Roeselova, M.; Tobias, D. Accommodation coefficients for water vapor at the air/water interface. *Chem. Phys. Lett.* **2004**, *393*, 249–255.

35. Dang, L.; Garrett, B. Molecular mechanism of water and ammonia uptake by the liquid/vapor interface of water. *Chem. Phys. Lett.* **2004**, *385*, 309–313.

36. Garrett, B.; Schenter, G.; Morita, A. Molecular simulations of the transport of molecules across the liquid/vapor interface of water. *Chem. Rev.* **2006**, *106*, 1355–1374.

37. Mason, P. E. Molecular Dynamics Study on the Microscopic Details of the Evaporation of Water. *J. Phys. Chem. A* **2011**, *115*, 6054–6058.

38. Tsuruta, T.; Nagayama, G. Molecular dynamics studies on the condensation coefficient of water. *J. Phys. Chem. B* **2004**, *108*, 1736–1743.

39. Morita, A. et al. Mass accommodation coefficient of water: molecular dynamics simulation and revised analysis of droplet train/flow reactor experiment. *J. Phys. Chem. B* **2004**, *108*, 9111–9120.

40. Vieceli, J. et al. Molecular dynamics simulations of atmospheric oxidants at the air-water interface: Solvation and accommodation of OH and O3. *J. Phys. Chem. B* **2005**, *109*, 15876–15892.

41. Ishiyama, T.; Yano, T.; Fujikawa, S. Molecular dynamics study of kinetic boundary condition at an interface between a polyatomic vapor and its condensed phase. *Phys. Fluids* **2004**, *16*, 4713–4726.

42. Yang, T.; Pan, C. Molecular dynamics simulation of a thin water layer evaporation and evaporation coefficient. *Int. J. Heat Mass Trans.* **2005**, *48*, 3516–3526.

43. Matsumoto, M. Molecular dynamics of fluid phase change. *Fluid Phase Equilibr.* **1998**, *144*, 307–314.

44. Caleman, C.; van der Spoel, D. Evaporation from water clusters containing singly charged ions. *Phys Chem Chem Phys* **2007**, *9*, 5105–5111.

45. Caleman, C.; van der Spoel, D. Temperature and structural changes of water clusters in vacuum due to evaporation. *J. Chem. Phys.* **2006**, *125*, 154508.

46. Li, L. et al. BioDrugScreen: a computational drug design resource for ranking molecules docked to the human proteome. *Nucleic Acids Res.* **2010**, *38*, D765–D773.

47. Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.

48. Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.

49. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **2001**, *46*, 3–26.

50. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug

discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, *1*, 55–68.

51. Matter, H. et al. Computational approaches towards the rational design of drug-like compound libraries. *Comb. Chem. High Throughput Screen.* **2001**, *4*, 453–75.

52. Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. In *Methods Mol. Biol.*; Zhou, J. Z., Ed.; Springer, 2011; Vol. 685; pp 253–276, and references therein.

53. Kutchukian, P. S. et al. Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery. *PLoS ONE* **2012**, *7*, e48476.

54. Joback, K. G.; Stephanopoulos, G. *Intelligent Systems in Process Engineering*; Advances in Chemical Engineering; Academic Press, 1995; Vol. 21; pp 257–311.

55. Patkar, P. R.; Venkatasubramanian, V. *Computer Aided Molecular Design: Theory and Practice*; Computer-Aided Chemical Engineering; Elsevier, 2003; Vol. 12; pp 95–128.

56. Eslick, J. C.; Shulda, S. M.; Spencer, P.; Camarda, K. V. *Molecular Systems Engineering*; Process Systems Engineering; Wiley-VCH, 2010; Vol. 6; pp 173–194.

57. Ref. 54–56, and references therein.

58. Siddhaye, S.; Camarda, K.; Southard, M.; Topp, E. Pharmaceutical product design using combinatorial optimization. *Comput. Chem. Eng.* **2004**, *28*, 425–434.

59. Siddhaye, S.; Camarda, K.; Topp, E.; Southard, M. Design of novel pharmaceutical products via combinatorial optimization. *Comput. Chem. Eng.* **2000**, *24*, 701–704.

60. Gillet, V. J. De Novo Molecular Design. **2000**, *8*, 49–66.

61. Terfloth, L.; Gasteiger, J. Neural networks and genetic algorithms in drug design. *Drug Discov. Today* **2001**, *6*, S102–S108.

62. Weininger, D. Method and Apparatus for Designing molecules with Desired Properties by Evolving Successive Populations. US Patent 5,434,796. 1995.

63. Rogers, D.; Tanimoto, T. A computer program for classifying plants. *Science* **1960**, *132*, 1115–1118.

64. Glen, R.; Payne, A. A genetic algorithm for the automated generation of molecules within constraints. *J. Comp.-Aided Mol. Des.* **1995**, *9*, 181–202.

65. Westhead, D.; Clark, D.; Frenkel, D.; Li, J. PRO_LIGAND: An approach to de novo molecular design. 3. A genetic algorithm for structure refinement. *Mol. Des.* **1995**, *9*, 139–148.

66. Nachbar, R. Molecular evolution: automated manipulation of hierarchical chemical topology and its application to average molecular structures. *Genetic Prog. Evolv. Machines* **2000**, 57–94.

67. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302.

68. Schneider, G.; Lee, M.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comp.-Aided Mol. Des.* **2000**, *14*, 487–494.

69. Pegg, S.; Haresco, J.; Kuntz, I. A genetic algorithm for structure-based de novo design. *J. Comp.-Aided Mol. Des.* **2001**, *15*, 911–933.

70. Ewing, T.; Kuntz, I. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175–1189.

71. Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. LEA3D: a computer-aided ligand design for structure-based drug design. *J. Med. Chem.* **2005**, *48*, 2457–2468.

72. Rarey, M.; Wefing, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comp.-Aided Mol. Des.* **1996**, *10*, 41–54.

73. Lameijer, E.; Kok, J.; Bäck, T.; IJzerman, A. The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *J. Chem. Inf. Modell.* **2006**, *46*, 545–552.

74. Dey, F.; Caflisch, A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Modell.* **2008**, *48*, 679–690.

75. Santiso, E.; Gubbins, K. Multi-scale molecular modeling of chemical reactivity. *Mol. Simulat.* **2004**, *30*, 699–748.

76. Kubinyi, H. In *Encyclopedia of Computational Chemistry*; von Rague Schleyer, P., Ed.; Van Nostrand Reinhold, Wiley, 1998; pp 2309–2320.

77. Jurs, P. C. In *Encyclopedia of Computational Chemistry*; von Rague Schleyer, P., Ed.; Van Nostrand Reinhold, Wiley, 1998; pp 2320–2330.

78. Hachmann, J. et al. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

79. Olivares-Amaya, R. et al. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **2011**, *4*, 4849–4861.

80. Wang, J. et al. Development and testing of a general amber force field. *J. Comp. Chem.* **2004**, *25*, 1157–1174.

81. Wang, J.; Wang, W.; Kollman, P.; Case, D. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.

82. Vanommeslaeghe, K. et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–90.

83. Arbuse, A. et al. Fine-tuning ligand-receptor design for selective molecular recognition of dicarboxylic acids. *Inorg. Chem.* **2007**, *46*, 10632–10638.

84. Monti, S. et al. Towards the design of highly selective recognition sites into molecular imprinting polymers: A computational approach. *Biosens. Bioelectron.* **2006**, *22*, 153–163.

85. Shen, X.-L.; Takimoto-Kamimura, M.; Wei, J.; Gao, Q.-Z. Computer-aided de novo ligand design and docking/molecular dynamics study of Vitamin D receptor agonists. *J. Mol. Model.* **2012**, *18*, 203–212.

86. Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **2005**, *48*, 4040–4048.

87. Ulman, A. Formation and structure of self-assembled monolayers. *Chem. Rev.* **1996**, *96*, 1533–1554.

88. Gembicki, S.; Rekoske, J.; Oroskar, A.; Johnson, J. Adsorption, liquid separation. *Kirk-Othmer Encyclopedia of Chemical Technology* **2002**, *1*, 678–691.

89. Gomes, P. S.; Minceva, M.; Rodrigues, A. E. Simulated moving bed technology: old and new. *Adsorption* **2006**, *12*, 375–392.

90. Seidel-Morgenstern, A.; Keßler, L.; Kaspereit, M. New developments in simulated moving bed chromatography. *Chem. Eng. Technol.* **2008**, *31*, 826–837.

91. Guest, D. Evaluation of simulated moving bed chromatography pharmaceutical process development. *J. Chrom. A* **1997**, *760*, 159–162.

92. Deveant, R. et al. Enantiomer Separation of a Novel Ca-Sensitizing Drug by simulated moving bed (SMB)–chromatography. *J. Prakt. Chem.* **1997**, *339*, 315–321.

93. Francotte, E.; Richert, P. Applications of simulated moving-bed chromatography to the separation of the enantiomers of chiral drugs. *J. Chrom. A* **1997**, *769*, 101–107.

94. Francotte, E.; Richert, P.; Mazzotti, M.; Morbidelli, M. Simulated moving bed chromatographic resolution of a chiral antitussive. *J. Chrom. A* **1998**, *796*, 239–248.

95. Wu, D.; Ma, Z.; Wang, N. Optimization of throughput and desorbent consumption in simulated moving-bed chromatography for paclitaxel purification. *J. Chrom. A* **1999**, *855*, 71–89.

96. Grill, C.; Miller, L.; Yan, T. Resolution of a racemic pharmaceutical intermediate—A comparison of preparative HPLC, steady state recycling, and simulated moving bed. *J. Chrom. A* **2004**, *1026*, 101–108.

97. Wei, F.; Shen, B.; Chen, M. From analytical chromatography to simulated moving bed chromatography: Resolution of omeprazole enantiomers. *Ind. Eng. Chem. Res.* **2006**, *45*, 1420–1425.

98. Huthmann, E.; Juza, A. Less common applications of simulated moving bed chromatography in the pharmaceutical industry. *J. Chrom. A* **2005**, *1092*, 24–35.

99. Juza, M.; Mazzotti, M.; Morbidelli, M. Simulated moving-bed chromatography and its application to chirotechnology. *Trends Biotech.* **2000**, *18*, 108–118.

100. Francotte, E. Enantioselective chromatography as a powerful alternative for the preparation of drug enantiomers. *J. Chrom. A* **2001**, *906*, 379–397.

101. Andersson, S.; Allenmark, S. Preparative chiral chromatographic resolution of enantiomers in drug discovery. *J. Biochem. Bioph. Meth.* **2002**, *54*, 11–23.

102. Zhang, Y.; Wu, D.; Wang-Iverson, D.; Tymiak, A. Enantioselective chromatography in drug discovery. *Drug Disc. Today* **2005**, *10*, 571–577.

103. Rajendran, A.; Paredes, G.; Mazzotti, M. Simulated moving bed chromatography for the separation of enantiomers. *J. Chrom. A* **2009**, *1216*, 709–738.

104. Juza, M. Development of an high-performance liquid chromatographic simulated moving bed separation from an industrial perspective. *J. Chrom. A* **1999**, *865*, 35–49.

105. Wu, D. J.; Ma, Z.; Wang, N. H. Optimization of throughput and desorbent consumption in simulated moving-bed chromatography for paclitaxel purification. *J. Chrom. A* **1999**, *855*, 71–89.

106. Jupke, A.; Epping, A.; Schmidt-Traub, H. Optimal design of batch and simulation moving bed chromatographic separation processes. *J. Chrom. A* **2002**, *944*, 93–117.

107. Schaber, S. D. et al. Economic Analysis of Integrated Continuous and Batch Pharmaceutical Manufacturing: A Case Study. *Ind. Eng. Chem. Res.* **2011**, *50*, 10083–10092.

108. Miller, L. et al. Chromatographic resolution of the enantiomers of a pharmaceutical intermediate from the milligram to the kilogram scale. *J. Chrom. A* **1999**, *849*, 309–317.

109. Zorita, S. et al. Selective determination of acidic pharmaceuticals in wastewater using molecularly imprinted solid-phase extraction. *Anal. Chim. Acta* **2008**, *626*, 147–154.

110. Widstrand, C.; Billing, J.; Boyd, B.; Yilmaz, E. Selective polymers speed sample preparation. *Manuf. Chemist (London, UK)* **Sept. 8, 2008**, p. 24.

111. Mohamed, R. et al. Use of Molecularly Imprinted Solid-Phase Extraction Sorbent for the Determination of Four 5-Nitroimidazoles and Three of Their Metabolites from Egg-Based Samples before Tandem LC-ESIMS/MS Analysis. *J. Agr. Food Chem.* **2008**, *56*, 3500–3508.

112. Kincaid, J.; Eyring, H. Free volumes and free angle ratios of molecules in liquids. *J. Chem. Phys.* **1938**, *6*, 620–629.

113. Berendsen, H.; Grigera, J.; Straatsma, T. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

114. Chen, F.; Smith, P. Simulated surface tensions of common water models. *J. Chem. Phys.* **2007**, *126*, 221101.

115. Kirkwood, J. G.; Buff, F. P. The statistical mechanical theory of surface tension. *J. Chem. Phys.* **1949**, *17*, 338–343.

116. Phillips, J. et al. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

117. Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089.

118. Essmann, U. et al. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

119. Wynveen, A.; Bresme, F. Interactions of polarizable media in water: A molecular dynamics approach. *J. Chem. Phys.* **2006**, *124*, 104502.

120. Alejandre, J.; Lynden-Bell, R. M. Phase diagrams and surface properties of modified water models. *Mol. Phys.* **2007**, *105*, 3029–3033.

121. Sakamaki, R. et al. Thermodynamic properties of methane/water interface predicted by molecular dynamics simulations. *J. Chem. Phys.* **2011**, *134*, 144702.

122. Chau, P.; Hardwick, A. A new order parameter for tetrahedral configurations. *Mol. Phys.* **1998**, *93*, 511–518.

123. Phillips, J. et al. Scalable molecular dynamics with NAMD. *J. Comp. Chem* **2005**, *26*, 1781–1802.

124. Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125*, 024106.

125. E, W.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B* **2002**, *66*, 052301.

126. E, W.; Ren, W. Q.; Vanden-Eijnden, E. Finite temperature string method for the study of rare events. *J. Phys. Chem. B* **2005**, *109*, 6688–6693.

127. Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *J. Chem. Phys.* **2005**, *123*, 134109.

128. E, W.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Ann. Rev. Phys. Chem.* **2010**, *61*, 391–420.

129. Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. On the assumptions underlying milestoning. *J. Chem. Phys.* **2008**, *129*, 174102.

130. Vanden-Eijnden, E.; Venturoli, M. Markovian milestoning with Voronoi tessellations. *J. Chem. Phys.* **2009**, *130*, 194101.

131. Vanden-Eijnden, E.; Venturoli, M. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **2009**, *130*, 194103.

132. Maragliano, L.; Vanden-Eijnden, E.; Roux, B. Free energy and kinetics of conformational transitions from Voronoi tessellated milestoning with restraining potentials. *J. Chem. Theo. Comput.* **2009**, *5*, 2589–2594.

133. Eiter, T.; Mannila, H. *Computing Discrete Frechet Distance*; Technical report CD-TR 94/64, TU Vienna: Vienna, Austria, 1994.

134. Maragliano, L.; Vanden-Eijnden, E.; Roux, B. Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials. *J. Chem. Theo. Comput.* **2009**, *5*, 2589–2594.

135. BenNaim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81*, 2016.

136. Mardia, K. V.; Jupp, P. E. *Directional Statistics*; Wiley, 2000.

137. Welsch, R. E. In *Encyclopedia of Statistical Sciences*, 2nd ed.; Balakrishnan, N., Read, C. B., Vidakovic, B., Eds.; Wiley, 2006.

138. Prescott, P. In *Encyclopedia of Statistical Sciences*, 2nd ed.; Balakrishnan, N., Read, C. B., Vidakovic, B., Eds.; Wiley, 2006.

139. Single-point calculations from low-energy MD frames, performed using the correlation-consistent cc-pVTZ[177] basis set and RIMP2 correlation with a matching auxiliary basis set.[178,179] The dipole moments from two configurations were Boltzmann-weighted to calculate to the average values.

140. Linstrom, P., Mallard, W., Eds. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology, Gaithersburg, MD, 2010–2012; (retrieved October 2010).

141. Jakalian, A.; Bush, B.; Jack, D.; Bayly, C. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

142. Jakalian, A.; Jack, D.; Bayly, C. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

143. Marcus, Y. *The Properties of Solvents*; Wiley, 1998.

144. Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford, 1987.

145. Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley, 1989.

146. Miller, B. L.; Goldberg, D. E. Genetic algorithms, selection schemes, and the varying effects of noise. *Evol. Comput.* **1996**, *4*, 113–131.

147. Augustson, J. G.; Minker, J. An analysis of some graph theoretical cluster techniques. *J. Assoc. Comput. Mach.* **1970**, *17*, 571–588.

148. Oprea, T.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157–166.

149. Oprea, T.; Zamora, I.; Ungell, A. Pharmacokinetically based mapping device for chemical space navigation. *J Comb Chem* **2002**, *4*, 258–266.

150. Holder, A. J. In *Encyclopedia of Computational Chemistry*; von Rague Schleyer, P., Ed.; Van Nostrand Reinhold, Wiley, 1998.

151. Bodor, N.; Buchwald, P. Molecular Size Based Approach To Estimate Partition Properties for Organic Solutes. *J. Phys. Chem. B* **1997**, *101*, 3404–3412, doi: 10.1021/jp9638503.

152. Edward, J. Calculation of octanol-water partition coefficients of organic solutes from their molecular volumes. *Can. J. Chem.* **1998**, *76*, 1294–1303.

153. The reason for using the 80^{th} percentile score, rather than the maximum, is that the former shows much less variation from generation to generation than the maximum. This is analogous to the fact that, within a sample of N independent random variables from identical distributions, the maximum would be expected to vary more than the median or the 80^{th} percentile score as illustrated in Figure B-2 in the Supplementary Information.

154. Santiso, E. E.; Trout, B. L. A general set of order parameters for molecular crystals. *J. Chem. Phys.* **2011**, *134*, 064109.

155. In such cases, our GAFF atomtype assignments allowed directly-linked aromatic groups (as in diphenyl) to assume a co-planar configuration.

156. Mühlenbein, H.; Schlierkamp-Voosen, D. Predictive models for the breeder genetic algorithm i. continuous parameter optimization. *Evol. Comput.* **1993**, *1*, 25–49.

157. Harter, H. L.; Balakrishnan, N. *CRC Handbook of Tables for the Use of Order Statistics in Estimation*; CRC Press, 1996.

158. Miller, B. L. *Noise, sampling, and efficient genetic algorithms*; Technical report 97001, Illinois Genetic Engineering Laboratory, University of Illinois: Urbana-Champaign, Ill, 1997.

159. Although if, after the consolidation, the population has a fitness distribution with a "fat" left tail (from newly-generated designs) or right tail (from the high-scoring retained ligand designs), the assumption of normality underlying the score improvement does not apply, and the improvement from selection may be less than $I\sigma_t$.

160. Willard, A. P.; Chandler, D. Instantaneous liquid interfaces. *J. Phys. Chem. B* **2010**, *114*, 1954–1958.

161. Santiso, E. E.; Trout, B. L. A general set of order parameters for molecular crystals. *J. Chem. Phys.* **2011**, *134*, 064109.

162. Mason, P. E.; Brady, J. W. "Tetrahedrality" and the relationship between collective structure and radial distribution functions in liquid water. *J. Phys. Chem. B* **2007**, *111*, 5669–5679.

163. Savage, C. A Survey of Combinatorial Gray Codes. *SIAM Rev.* **1997**, *39*, 605–629.

164. Wood, G. P.; Santiso, E. E.; Trout, B. L. A Simple Genetic Algorithm Using Quaternion Encoding for Molecular Orientations. *J. Chem. Theo. Comput.* **2012**, submitted for publication.

165. Bianchi, E. et al. Predicting patchy particle crystals: Variable box shape simulations and evolutionary algorithms. *J. Chem. Phys.* **2012**, *136*, 214102–214102–9.

166. Whitley, D. A Genetic Algorithm Tutorial. *Stat. Comput.* **1994**, *4*, 65–85, and references therein.

167. Baker, J. E. Reducing Bias and Inefficiency in the Selection Algorithm. Proceedings of the Second International Conference on Genetic Algorithms. 1987; pp 14–21.

168. Sagisaka, M. et al. Water/supercritical CO2 microemulsions with mixed surfactant systems. *Langmuir* **2008**, *24*, 10116–10122.

169. Sagisaka, M. et al. Super-Efficient Surfactant for Stabilizing Water-in-Carbon Dioxide Microemulsionst. *Langmuir* **2011**, *27*, 5772–5780.

170. Cellmer, T.; Bratko, D.; Prausnitz, J. M.; Blanch, H. W. Protein aggregation in silico. *Trends Biotechnol.* **2007**, *25*, 254–61.

171. Hamada, H.; Arakawa, T.; Shiraki, K. Effect of additives on protein aggregation. *Curr. Pharm. Biotechnol.* **2009**, *10*, 400–7.

172. Mohamed, A. et al. Low Fluorine Content CO2-philic Surfactants. *Langmuir* **2011**, *27*, 10562–10569.

173. Moulin, E.; Cormos, G.; Giuseppone, N. Dynamic combinatorial chemistry as a tool for the design of functional materials and devices. *Chem. Soc. Rev.* **2012**, *41*, 1031–49.

174. Andersson, M. et al. Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts. *J. Catal.* **2006**, *239*, 501–506.

175. Baerns, M.; Holena, M. *Combinatorial Development of Solid Catalytic Materials*; Catalytic Science Series; Imperial College Press, 2009; Vol. 7.

176. Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nature Chem.* **2009**, *1*, 37–46.

177. Peterson, K.; Dunning, T. Accurate correlation consistent basis sets for molecular core-valence correlation effects: The second row atoms Al-Ar, and the first row atoms B-Ne revisited. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

178. Weigend, F.; Haser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem Phys Lett* **1998**, *294*, 143–152.

179. Hättig, C. Optimization of auxiliary basis sets for RI-MP2 and RI-CC2 calculations: Core–valence and quintuple- basis sets for H to Ar and QZVPP basis sets for Li to Kr. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59–66.

180. Macke, T.; Svrcek-Seiler, W.; Brown, R. AmberTools Users' Manual. 2011.

Appendices

Appendix A

Water behavior as characterized by order parameters

A.1 Behavior in bulk water

To understand the behavior of the ten order parameters defined in Section 3.4, a restraint-free simulation of bulk water was performed in the NPT ensemble, with $N = 1,025$ molecules. After minimization and equilibration for 1 ns of MD, the order parameters' values were measured every 5 fs during a 1-ns production run. During equilibration and production, temperature was controlled using Langevin dynamics (298 K, damping coefficient 4 ps^{-1}) and pressure was controlled with the Langevin piston approach (at 1.0 atm, with osciallation period 200 fs and decay time 100 fs) in NAMD.[116] During this simulation, the average density was 1.027 g/cm^3.

The distributions of order parameter values are shown in Figures A-1 and A-2. The order parameters listed in Figure A-1 are well-defined in bulk water; in contrast, those in Figure A-2 are not, because they are defined with respect to the interfacial normal (coincident with $\hat{z}$ in interfacial simulations), and no interface existed in this bulk-liquid simulation. Thus, in this simulation, the order parameters describing position and orientation with respect to the interface were measured with respect to an arbitrary direction in the simulations "lab frame."

Table A-1 lists the instantaneous correlation of order parameters, with values

greater in magnitude than 0.25 highlighted. Order parameters 1, 6, and 7 are all correlated with each other because they measure the presence of atoms within a few angstroms of the molecule's water atom; order parameter 1 records the contirbution for any atom in this range, while order parameters 6 and 7 record contibutions when one atom is an oxygen and the other is a hydrogen.

Figures A-3 and A-4 show the autocorrelation of the order parameters, which have decay times between between 100 fs and several picoseconds.

Table A-1. Correlation of order parameters in simulation of bulk liquid SPC/E water. OP values were recorded every 5 fs.

OP	0	1	2	3	4	5	6	7	8	9
0	1.000	0.016	0.047	-0.017	0.147	0.002	-0.004	0.008	0.006	-0.025
1		1.000	0.057	0.143	0.038	-0.000	0.682	0.663	-0.243	0.007
2			1.000	0.109	-0.038	0.000	0.015	0.017	-0.012	-0.027
3				1.000	-0.016	-0.002	0.049	0.083	-0.038	0.007
4					1.000	0.003	0.009	0.025	0.006	-0.031
5						1.000	-0.014	-0.011	0.868	0.059
6							1.000	0.386	-0.206	0.016
7								1.000	-0.243	-0.004
8									1.000	0.069
9										1.000

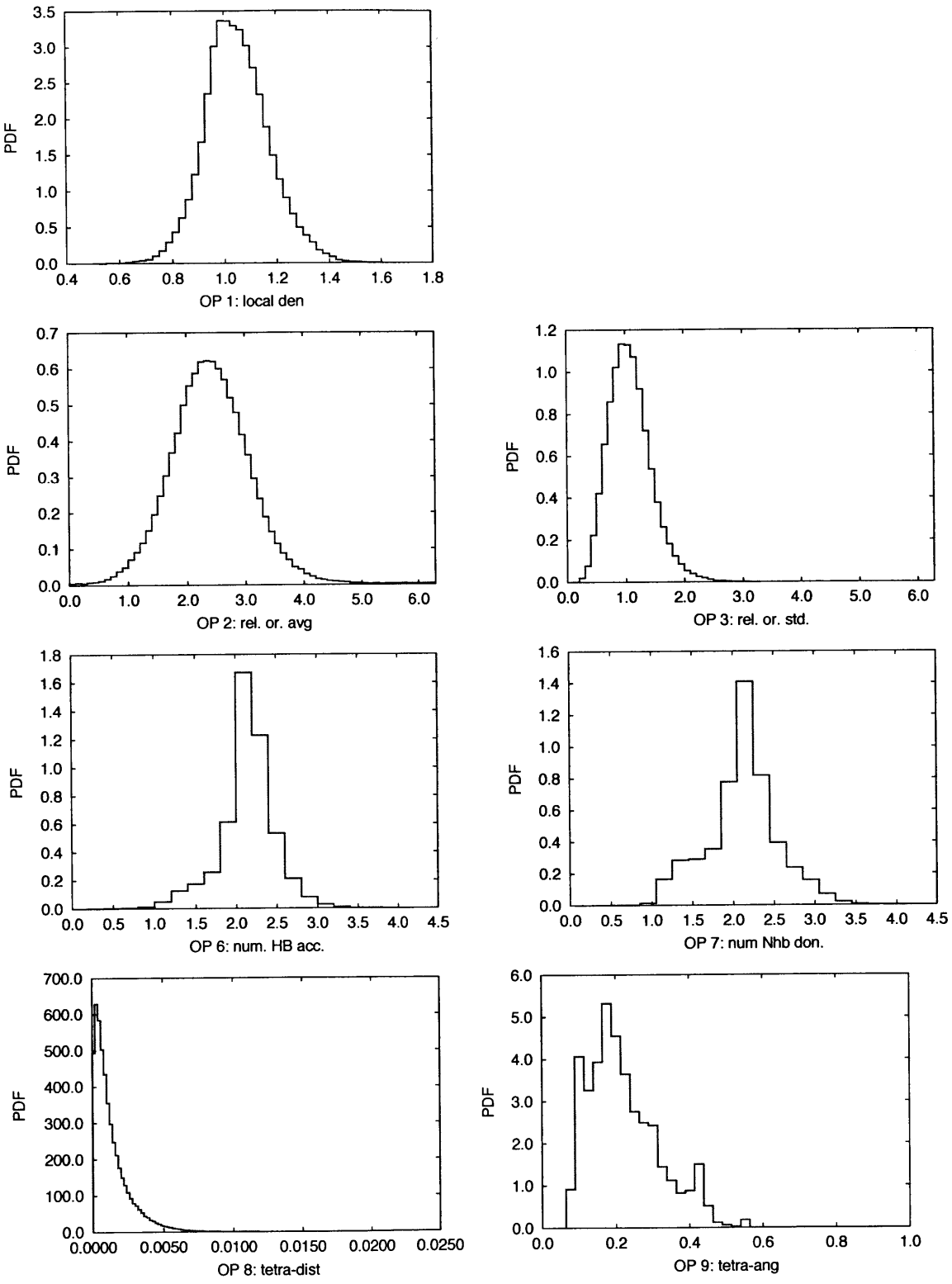


Figure A-1. Order parameter distributions in bulk SPC/E water, measured in NPT ensemble at 1.0 bar and 298 K. The order parameters above have meaningful ranges of values even in a completely bulk case.

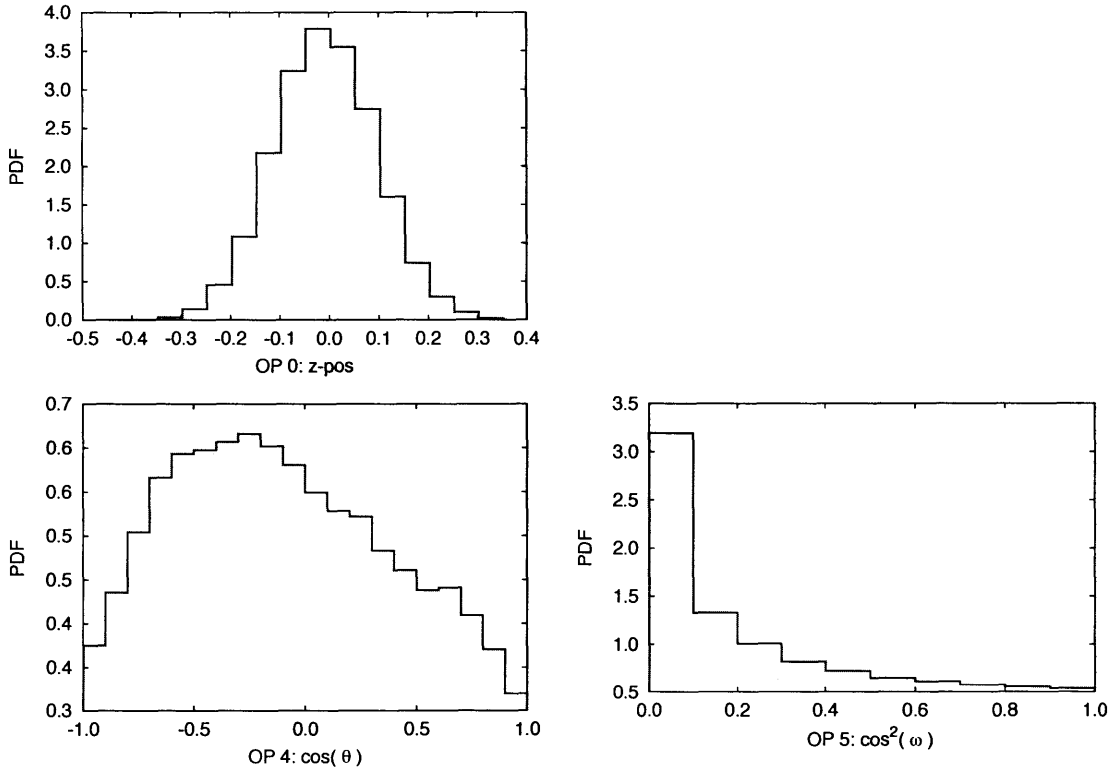


Figure A-2. Order parameter distributions in bulk SPC/E water, measured in NPT ensemble at 1.0 bar and 298 K. In this bulk simulation, the values of the order parameters were measured with respect to a lab frame, since they are defined in relation to an interfacial normal that did not exist in the bulk-liquid case.

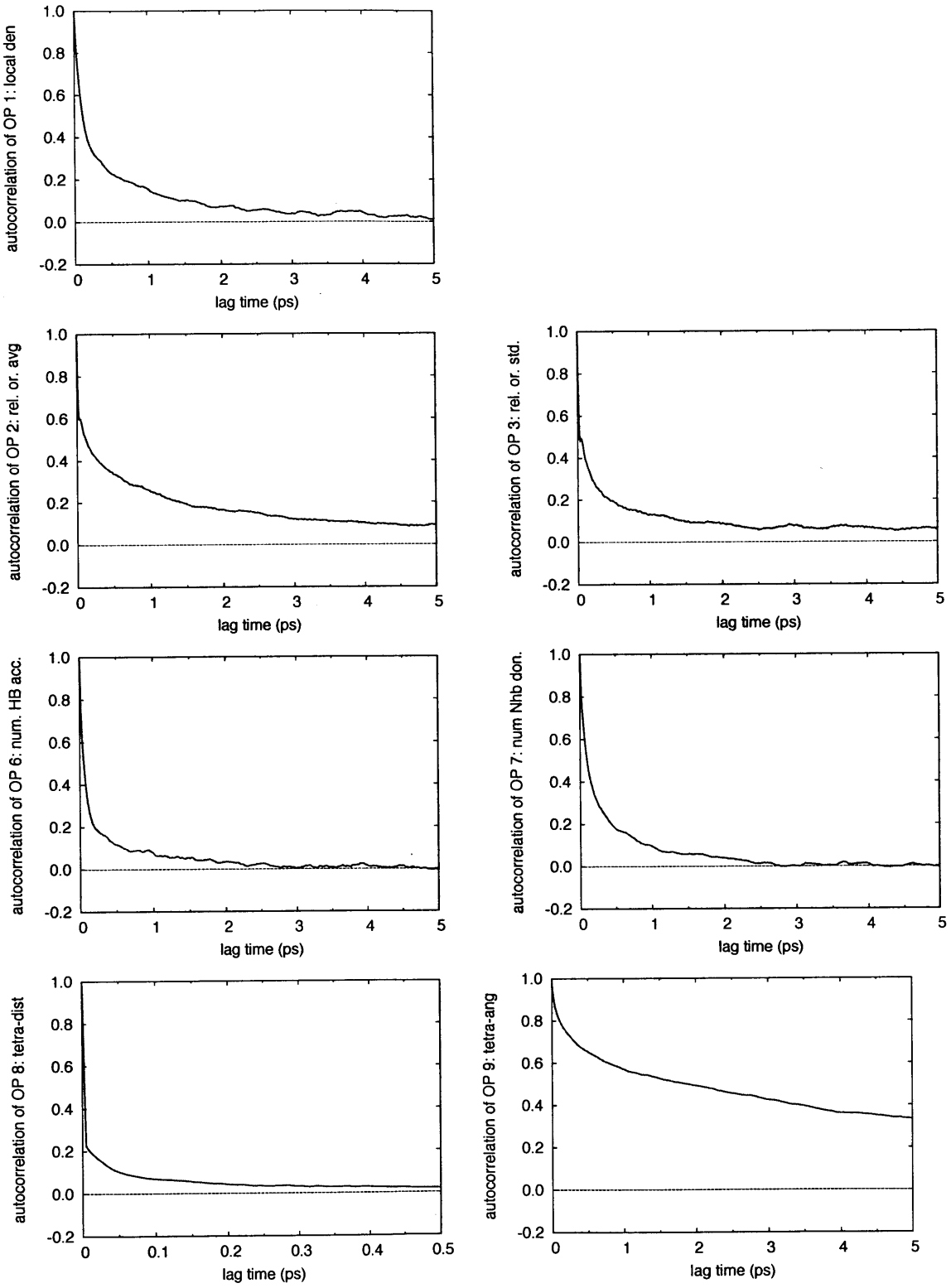


Figure A-3. Autocorrelation of order parameters in bulk SPC/E water at 1.0 bar and 298 K.

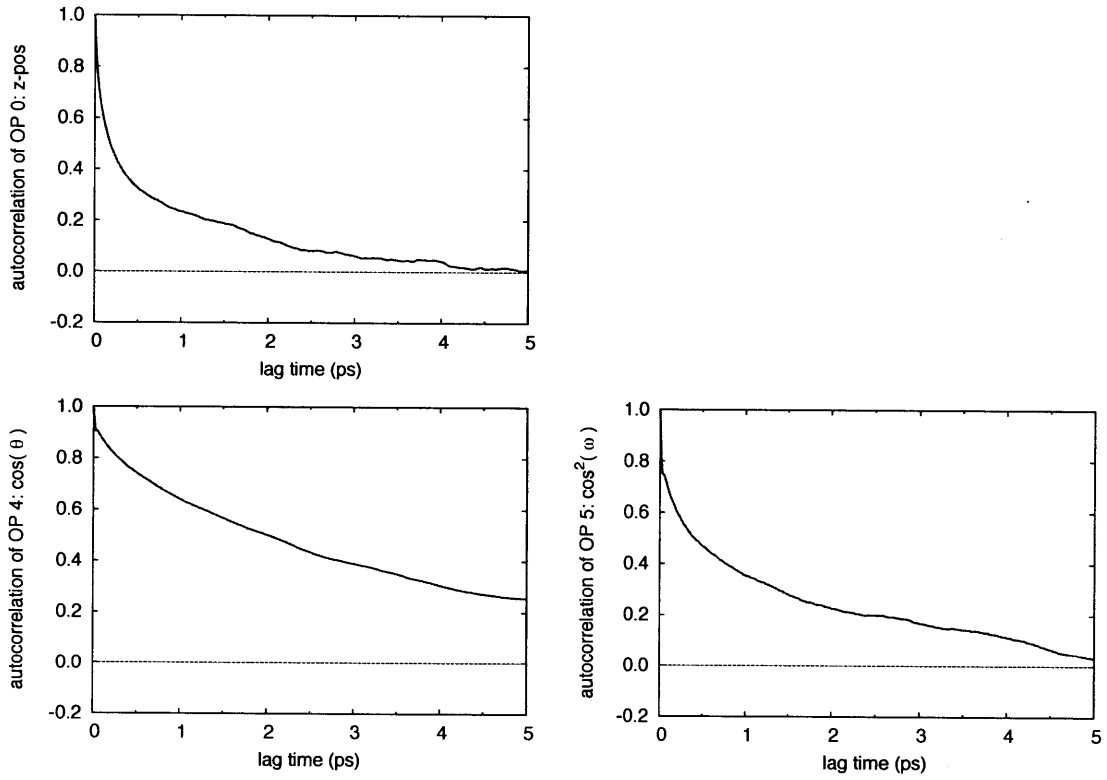


Figure A-4. Autocorrelation of order parameters in bulk SPC/E water at 1.0 bar and 298 K. See caption to Figure A-2.

Appendix B

Details of molecular evolution approach

The overall approach for ligand design, as discussed in Section 5.1 above, is to subject a population of linear, surface-modifying ligands to a genetic algorithm.

Finally, to understand whether design problem is amenable to a genetic algorithm (GA), and to better understand the impact of GA parameters on the evolution process, we performed a series of evolution experiments using a surrogate objective function on populations (and population members) similar to those in our full problem. This approach and the results obtained are described in Section B.4 (page 221).

B.1 Functional group information

Table A-1. Terminal functional groups used in design of linear ligands. See text for explanation of functional group properties listed as column headings.

CODON	NAME	STRUCTURE	v_i (Å^3)	n_{BB}	$N_{hb\ don}$	$N_{hb\ acc}$
0	hydrogen	$-H$	5.5	0	0	0
1	methyl	$-CH_3$	22.1	0	0	0
2	hydroxyl	$-OH$	13.4	2	1	2
3	aldehyde	$-CHO$	24.9	2	0	0
4	carboxyl	$-COOH$	32.8	2	1	2
5	primary amino	$-NH_2$	17.6	2	2	1
6	phenyl	$-Ph$	81.6	0	0	0
7	vinyl	$-CH{=}CH_2$	32.7	0	0	0
8	acetylenyl	$-C{\equiv}CH$	32.3	0	0	0
9	allenyl	$-CH{=}C{=}CH_2$	33.3	0	0	0
10	isopropyl	$-CH(CH_3)_2$	55.3	0	0	0
11	terbutyl	$-C(CH_3)_3$	71.9	0	0	0
12	amide	$-CONH_2$	37.0	4	2	1
13	thiol	$-SH$	23.4	0	1	2
14	fluoride	$-F$	9.5	0	1	0
15	chloride	$-Cl$	19.0	0	0	0
16	bromide	$-Br$	26.9	0	0	0

Table A-2. Intermediate functional groups used in design of linear ligands. See text for explanation of functional group properties listed as column headings.

CODON	GROUP NAME	STRUCTURE	v_i (Å^3)	n_{BB}	$N_{hb\ don}$	$N_{hb\ acc}$
1000	methylene	$-CH_2-$	16.6	0	0	0
1001	ether	$-O-$	6.2	2	0	0
1002	carbonyl	$-(CO)-$	19.4	2	0	0
1003	ester	$-COO-$	25.6	2	0	2
1004	secondary amino	$-NH-$	12.4	2	1	1
1005	*o*-didehydrobenzene	$-$(o)$Ph-$	70.6	0	0	0
1006	*m*-didehydrobenzene	$-$(m)$Ph-$	70.6	0	0	0
1007	*p*-didehydrobenzene	$-$(p)$Ph-$	70.6	0	0	0
1008	*cis*-ethylene-1,2-diyl	$-$(cis)$CH{=}CH-$	27.2	0	0	0
1009	*trans*-ethylene-1,2-diyl	$-$(trans)$CH{=}CH-$	27.2	0	0	0
1010	acetylene-1,2-diyl	$-C{\equiv}C-$	26.8	0	0	0
1011	allene-1,3-diyl	$-CH{=}C{=}CH-$	27.8	0	0	0
1012	methanol-1,1-diyl	$-CHOH-$	24.5	0	1	2
1013	thioether	$-S-$	17.9	0	0	0
1014	isopropyl-methylene	$-$CH(iPr)$-$	66.4	0	0	0
1015	methyl-methylene	$-CH(CH_3)-$	33.2	0	0	0
1016	ethyl-methylene	$-CH(CH_2CH_3)-$	49.8	0	0	0
1017	dimethyl-methylene	$-C(CH_3)_2-$	49.8	0	0	0
1018	phenyl-methylene	$-CHPh-$	87.2	0	0	0
1019	carboxyl-methylene	$-CHCOOH-$	43.9	2	1	2
1020	amine-methylene	$-CHNH_2-$	28.7	2	2	1
1021	1,5-didehydronapthalene	$-C_{10}H_6-$	114.0	0	0	0
1022	2,6-didehydronapthalene	$-C_{10}H_6-$	114.0	0	0	0

B.2 Evolution with a surrogate objective function

B.2.1 Multi-ligand layer simulations

To set up ligand layer simulations, it was necessary to establish a close-packed ligand layer; we did so using a two-dimensional square lattice. To ascertain the appropriate two-dimensional density of the close-packed layer, the ligand was divided into 2.0 Å slices, parallel to the xy-plane, and the extent of the ligand in each slice in the x- and y-directions was measured. This slicing and measurement process is depicted in Figure B-1. If these extent values are denoted Δx_i and Δy_i for each slice i, then the preliminary per-ligand area (equal to the inverse of the two-dimensional number density) is equal to the RMS-average of the areas of the slices $A_i = (\Delta x_i)(\Delta y_i)$. The RMS average was used after we saw linear average produced a ligand layer which was too tightly packed.

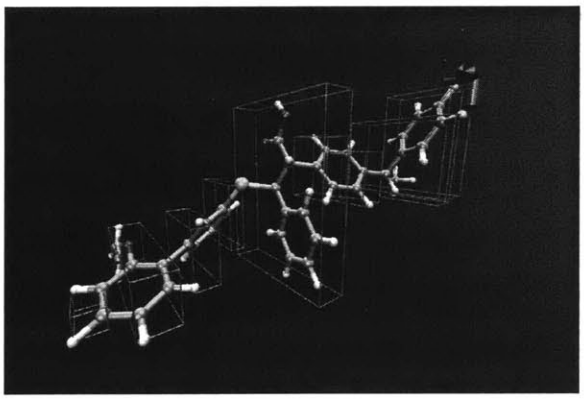


Figure B-1. Division of ligand into "slices" perpendicular to the z-axis (blue arrow) and measurement of the extent of the ligand in the x- and y-directions.

Finally, the true ligand spacing distance is set to the square root of the preliminary per-ligand area plus 1.7 Å. This additional buffer accounts for atoms' van der Waals radii. Both the ligand alignment and area-determination steps were performed using VMD.[1]

Next, the TLEAP program[180] was used to generate a two-dimensional array of ligands in a square matrix, using the ligand spacing distance calculated in the manner described above. A 5×5 arrangement of ligands is used, unless the supercell produced

would be less than 25 Å in width, in which case the number of ligands used is chosen to produce a square supercell greater than 25 Å. In the MD simulations, three-dimensional periodic boundary conditions are then used, with a periodic length in the z-direction of $L_z = 4L_x = 4L_y$. The dielectric constant was set to 6.4, to match the experimental value of pure ethyl acetate.

B.3 Supplemental Information: Evolution experiment results

The sections below include detailed characterization of the four molecular evolution experiments in this study. The populations' score distributions are characterized by their median and 80^{th} percentile value. The reason for using the latter quantity, rather than their maximum, is that the 80^{th} percentile would be expected to have less variance than the max value; an example of this is shown in Figure B-2. This was found to be true with the actual results from these four experiments, and so the 80^{th} percentile values were used to illustrate with clarity the trends in the evolving populations' score distributions.

Other properties are also given for the four evolving populations in Experiments I through IV, such as the number of unique ligands; the prevalence of functional groups and chemical motifs; and trends in molecular properties estimated with QSPRs.

B.3.1 Experiment I: Seventy-six generations of roulette-wheel selection after window scaling, using 3-ns production MD for evaluation

Table C-3. Prevalence of structural motifs in three different populations in Experiment I. The "top 90" population are the top-scoring ligand candidates, ranked by average score, for which *at least* 6 ns of production MD was performed (as in Table C-4). These 90 top-scoring members had selectivity scores ranging from 1.60 to 0.58 kcal/mol $\approx kT$. Motif entries are listed in order of prevalence in the top-90 scorers.

	in gen. 1		in gen. 75		in top 90 scorers	
MOTIF	count (out of 45)	fraction (%)	count (out of 45)	fraction (%)	count (out of 90)	fraction (%)
napthalene group	10	22	44	98	74	82
alkene/alkyne group	28	62	45	100	68	76
aromatic group[a] adjacent to alkene/alkyne group	13	29	42	93	59	66
H-bond acceptor group[b]	40	89	1	2	44	49
initial vinyl group	2	4	45	100	43	48
aromatic group adjacent to alkane group	10	22	44	98	41	46
fluoro group	8	18	0	0	38	42
H-bond donor group[c]	38	84	1	2	36	40
aromatic group adjacent to H-bond acceptor	19	42	1	2	31	34
aromatic group adjacent to fluoro group	3	7	0	0	29	32
amine group	22	49	1	2	22	24
aromatic group adjacent to H-bond donor	15	33	1	2	21	23
phenyl group	19	42	1	2	20	22
carbonyl group[d]	21	47	0	0	20	22
H-bond donor within 2 groups of acceptor group	21	47	0	0	13	14
aromatic group adjacent to carbonyl group	9	20	0	0	13	14
aromatic group adjacent to amino group	7	16	1	2	13	14
H-bond donor adjacent to acceptor group	18	40	0	0	11	12
soft, bulky group[e]	24	53	1	2	10	11
carboxyl group	15	33	0	0	10	11
amino group within 2 units of internal phenyl ring	8	18	0	0	9	10
hydroxyl group	12	27	0	0	8	9
thiol group	8	18	0	0	7	8
aromatic group adjacent to bulky, soft group	9	20	0	0	7	8
terminal phenyl ring	5	11	0	0	4	4
terminal halide (F or Cl)	0	0	0	0	4	4
aromatic group adjacent to hydroxyl group	3	7	0	0	4	4
hydroxyl group within 2 units of internal phenyl ring	4	9	0	0	3	3
aromatic group adjacent to carboxyl group	3	7	0	0	3	3
isopropyl group	10	22	0	0	2	2
ether or thioether group	10	22	1	2	2	2
E6-like: amino–any group–int. phenyl ring–hydroxyl	1	2	0	0	1	1
aromatic group adjacent to thiol group	1	2	0	0	1	1
aromatic group adjacent to isopropyl group	1	2	0	0	1	1
ligand is a saturated alkane	0	0	0	0	0	0
E2-like: hydroxyl–any group–int. phenyl ring–hydroxyl	0	0	0	0	0	0
aromatic group adjacent to ether/thioether group	2	4	1	2	0	0

[a] Aromatic groups are phenyl groups or napthalene groups.

[b] Acceptors are hydroxyl, aldehyde, carbonyl, carboxyl, ester, primary amino, secondary amino, amide, and thiol groups.

[c] Donors are hydroxyl, carboxyl, primary amino, secondary amino, amide, and thiol groups.

[d] Excludes carbonyl group within carboxyl groups.

[e] Includes isopropyl groups, *t*-butyl groups, ethyl-methylene [–CH(CH_2CH_3)–], dimethyl-methylene [–C(CH_3)$_2$–], and phenyl methylene [–CH(Ph)–].

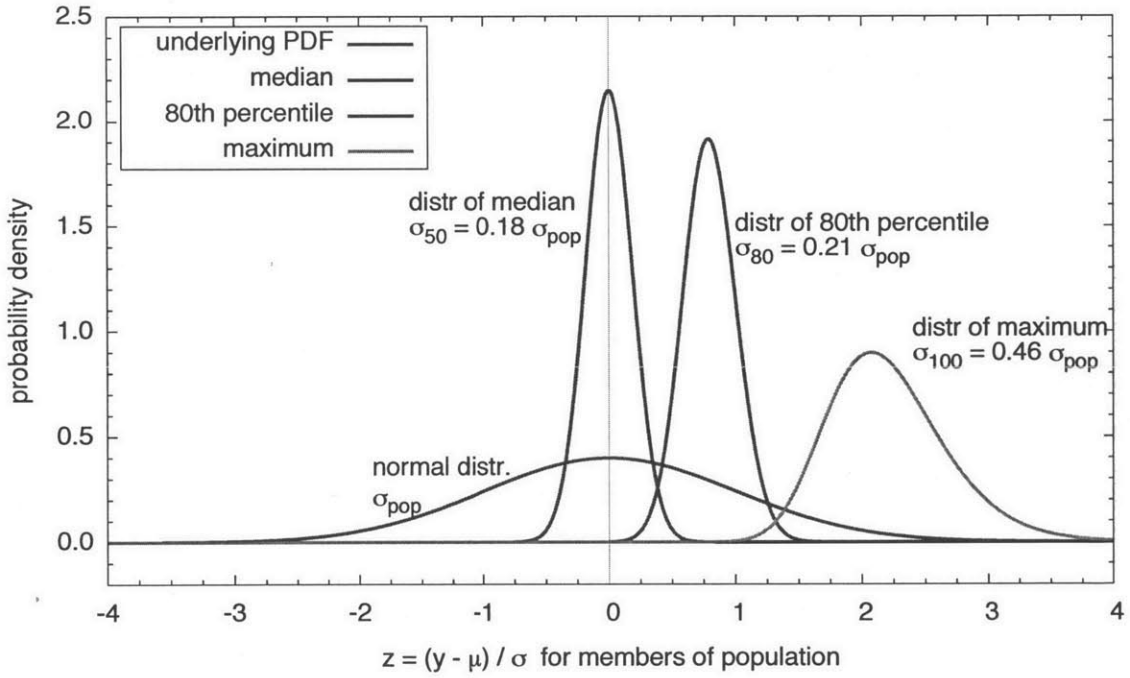


Figure B-2. Distribution of order statistics for a sample of $N = 45$ independent random variables, each drawn from a standard normal distribution $N(0, 1)$.

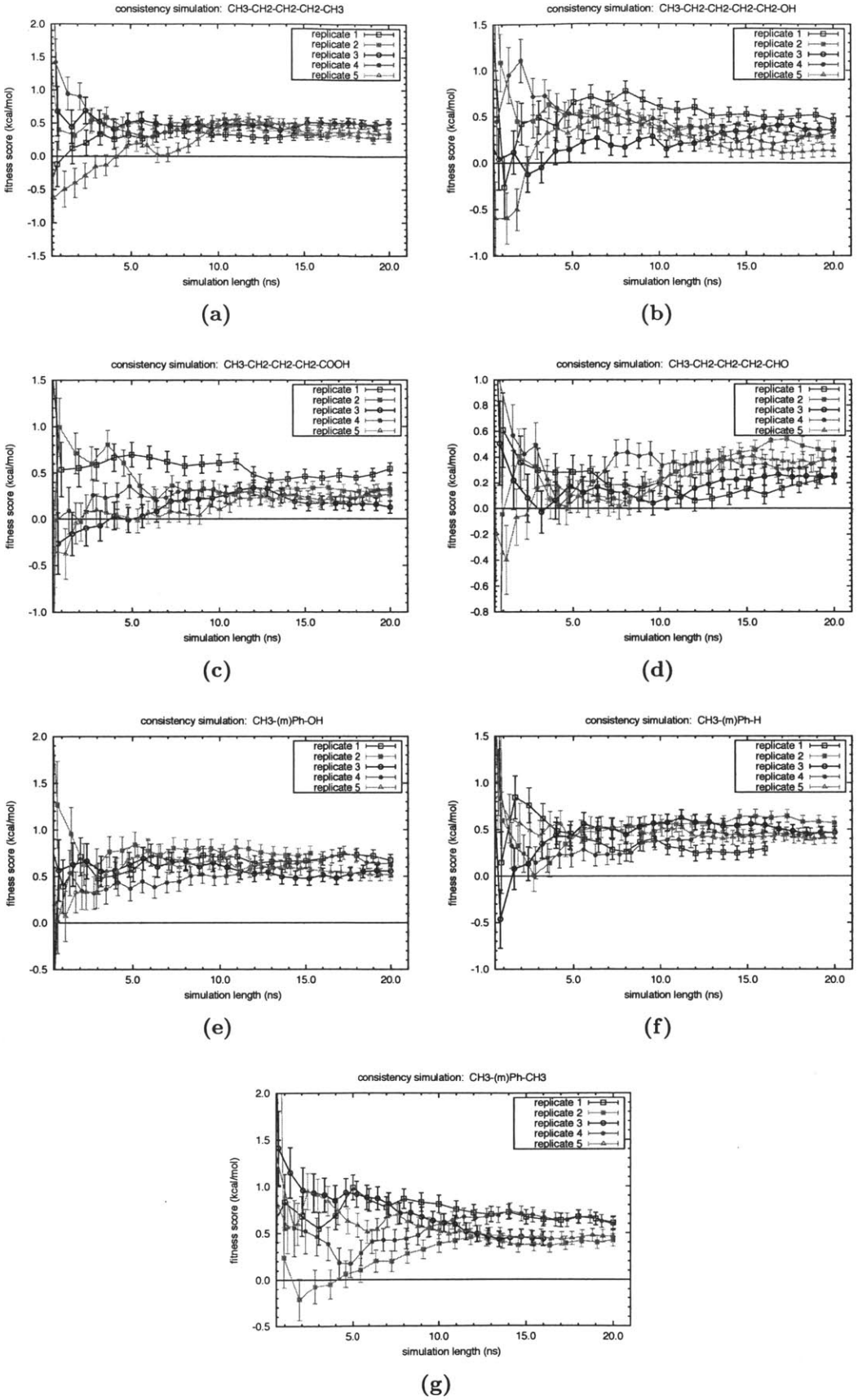


Figure B-3. Convergence of fitness score of ligands with seven distinct sequences. Error bars are ±1 standard error. The structure of each ligand is listed in the title of each plot.

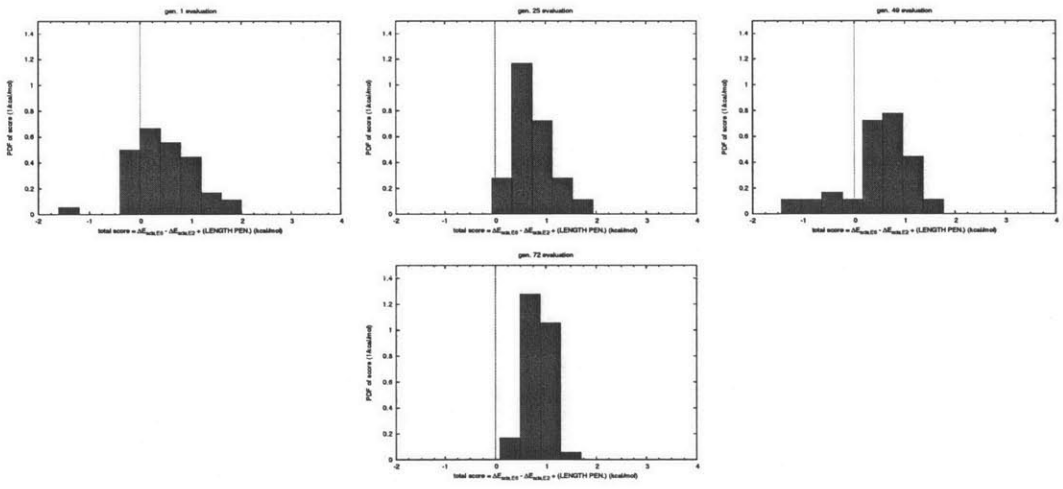
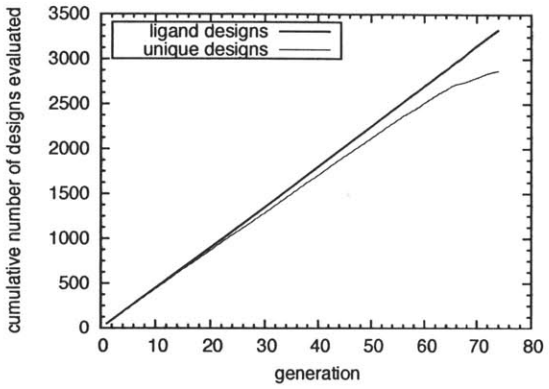
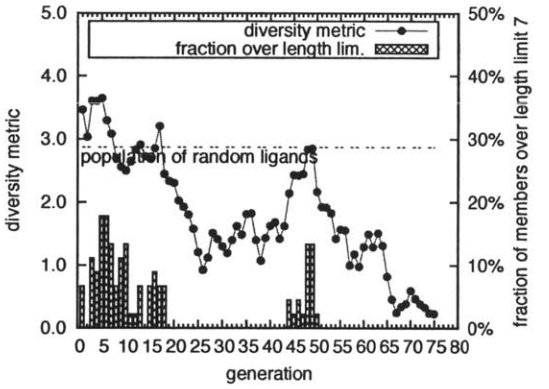


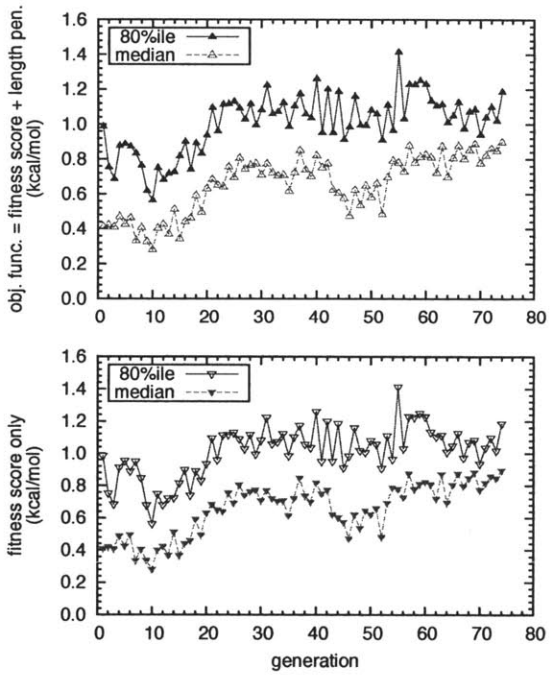
Figure B-4. Distribution of fitness scores (including length penalties) of members of generations 1, 25, 49, and 72 in experiment I.



(a) Cumulative number of ligand evaluations.

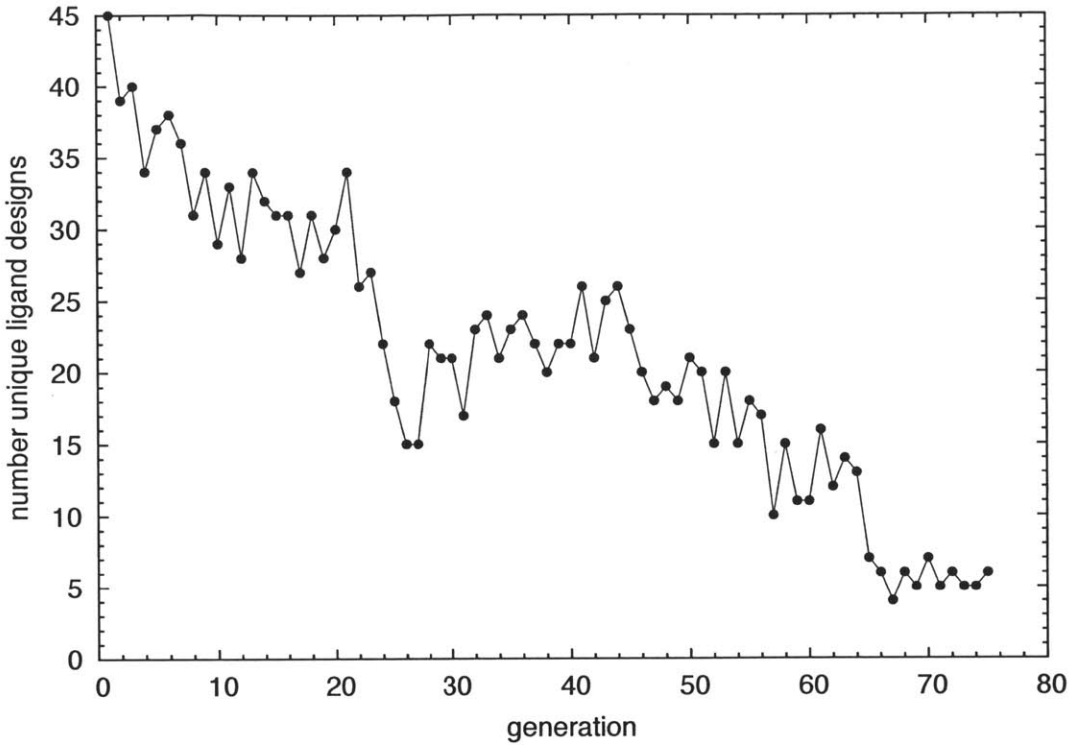


(b) Phenotypic diversity of evolving population of ligands. Bars show the fraction of ligands which exceed the length limit of 7 functional groups.

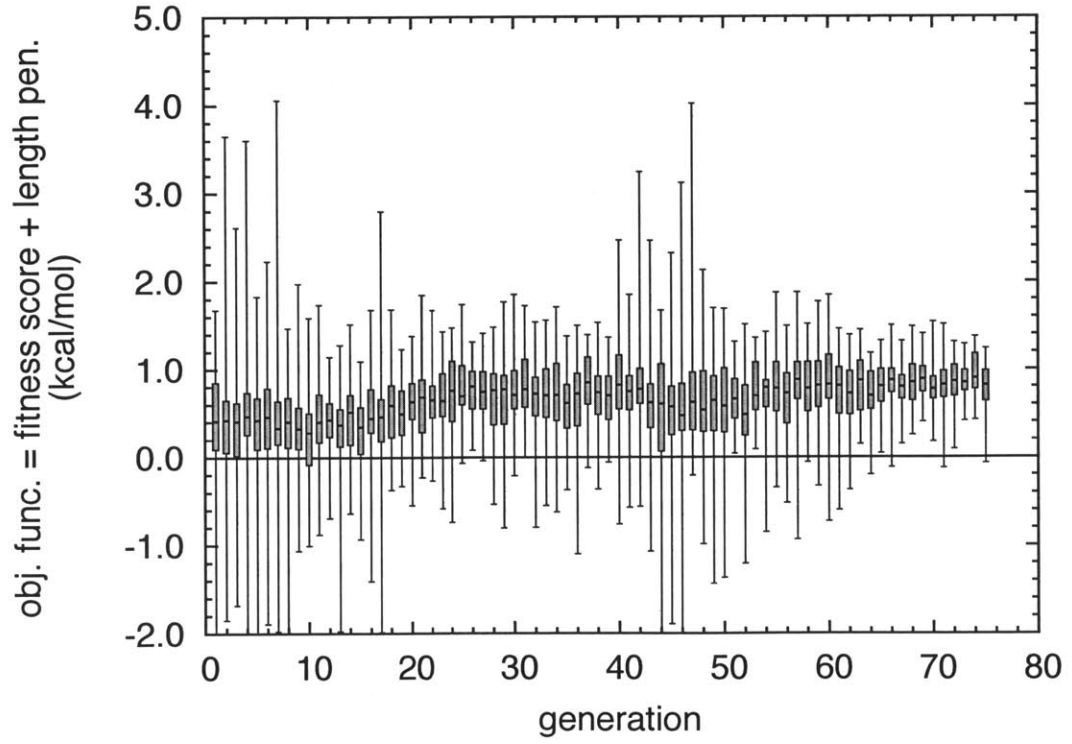


(c) Value of objective function at 80th and 50th percentiles in each generation. Depicted are the fitness score plus length penalty (*top*), and the finess score alone (*bottom*).

Figure B-5. Characterization of evolution over generations 1 to 74 in experiment I, which featured $N = 45$ ligands, roulette wheel selection after window-based scaling, and 3.0-ns production MD in each evaluation.



(a) Number of unique ligand designs (out of a total of 45) within each generation.



(b) Box-and-whisker plots of objective function (selection score + length penalty) measurements for members of each generation.

Figure B-6. Evolution dynamics in Experiment I.

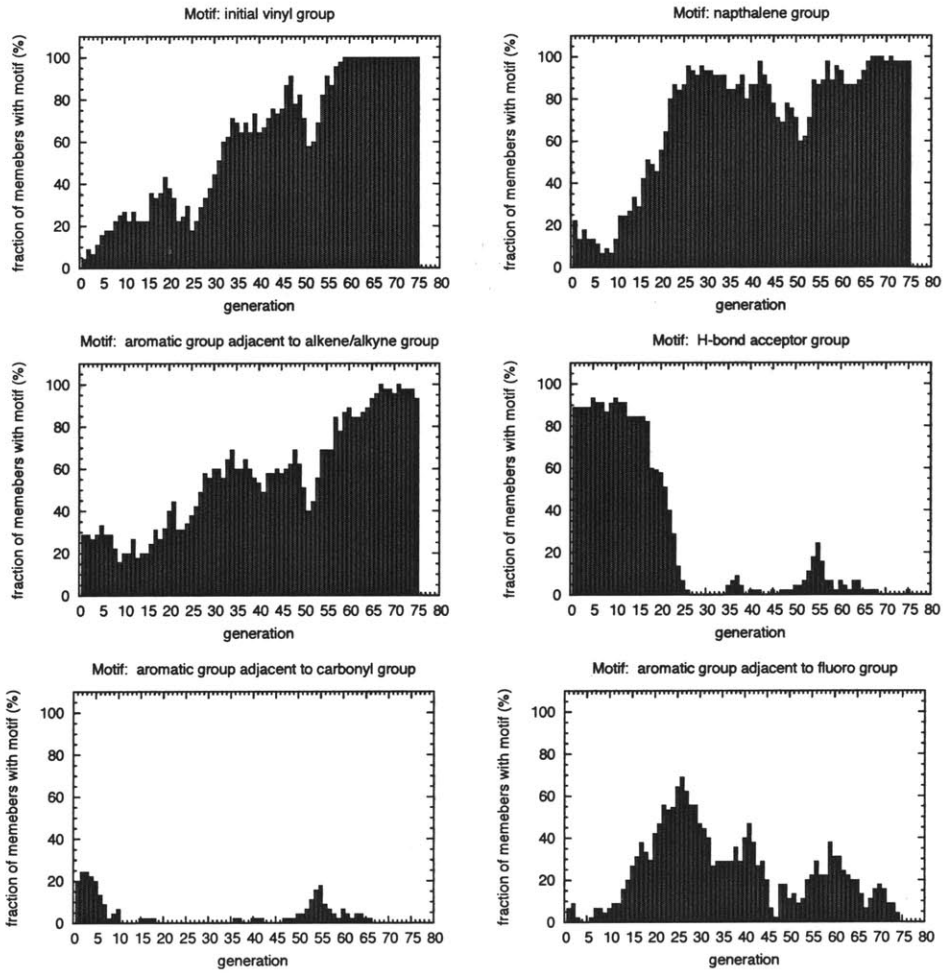


Figure B-7. Prevalence of motifs in generations 1 to 75 of Experiment I. Motif descriptions are listed in the title of each graph.

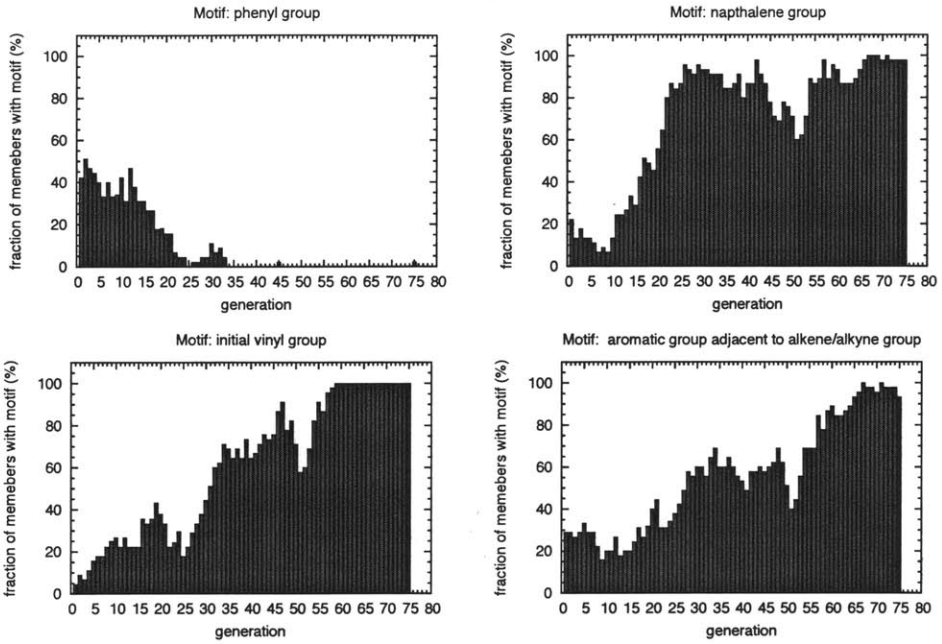


Figure B-8. Prevalence of motifs involving unsaturated/aromatic groups in generations 1 to 75 of Exp. I. Motif descriptions are listed in the title of each graph.

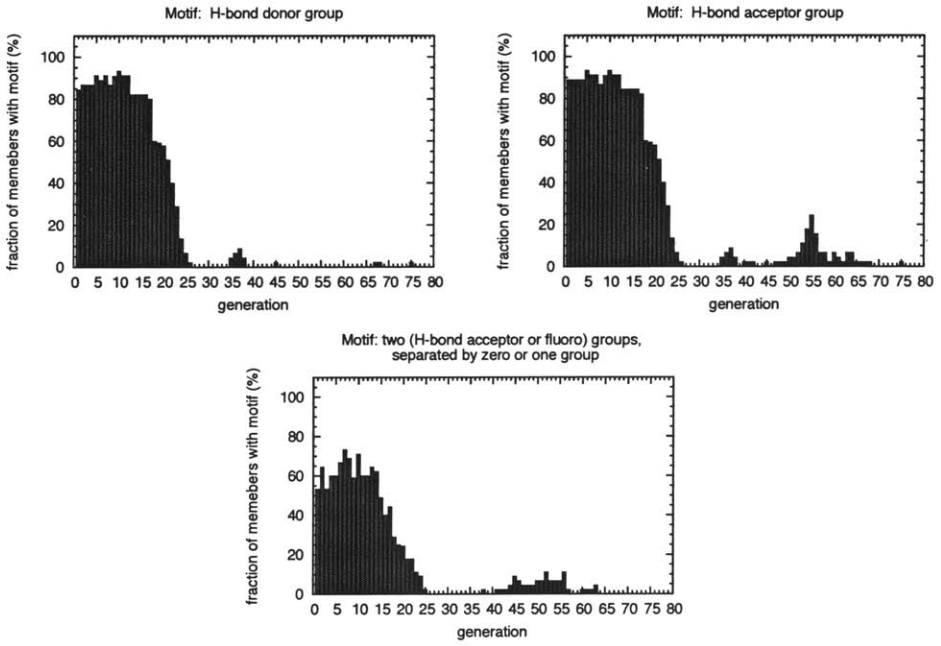


Figure B-9. Prevalence of motifs involving hydrogen-bond donors and acceptors in generations 1 to 75 of Experiment I.

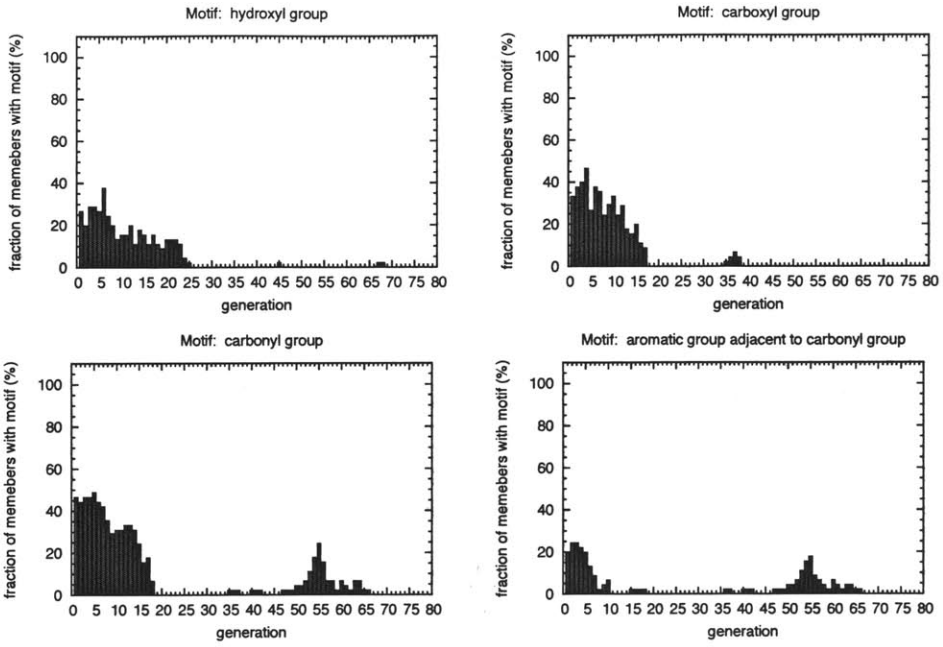


Figure B-10. Prevalence of motifs involving oxygen- or sulfur-containing groups in generations 1 to 75 of Experiment I. Note that the "carbonyl groups" do not include the carbonyl carbon within carboyxl groups.

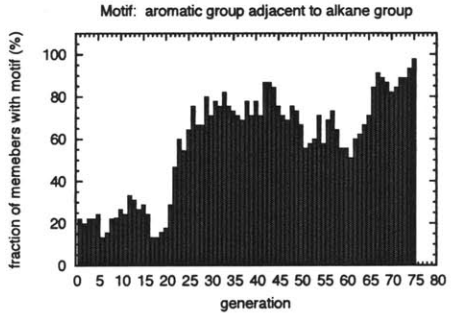


Figure B-11. Prevalence of other motifs in generations 1 to 75 in Experiment I. Note that ether groups can consist of either oxygen-based ether groups or thioether groups.

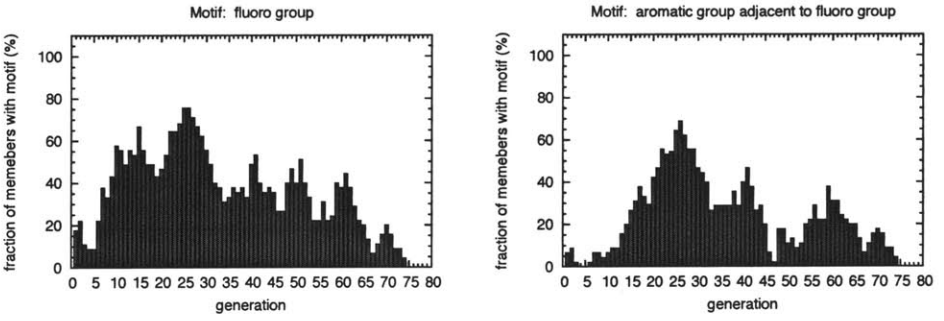
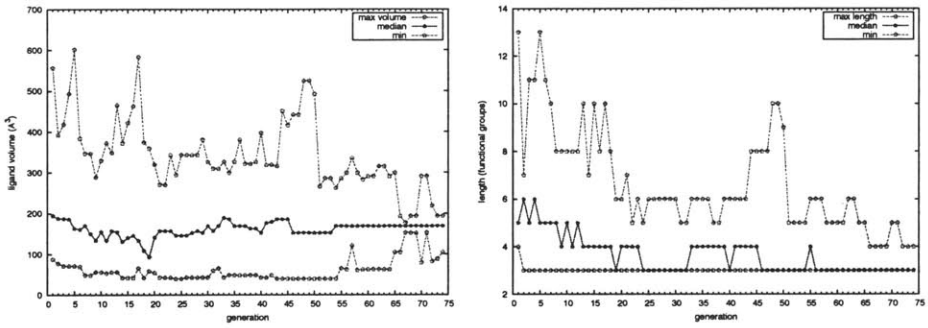


Figure B-12. Prevalence of motifs involving halide and amino groups in generations 1 to 75 in Experiment I.

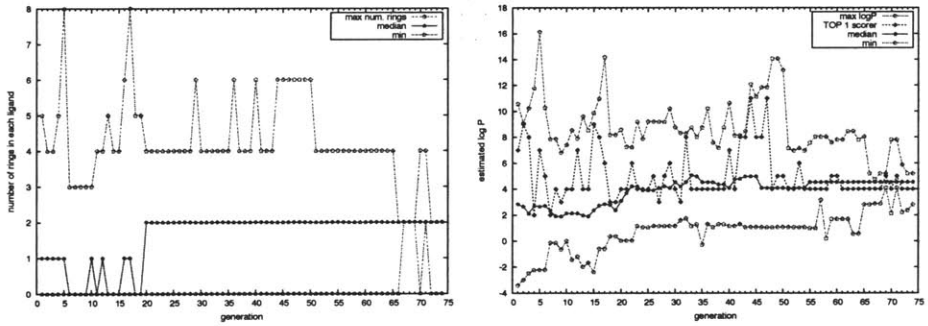
Table C-4. Top-scoring forty-five ligand designs from generations 1 through 74 in experiment I, *for which at least 6 ns of production MD* (corresponding to two ligand evaluations) was performed. Listed scores are averages of all evaluations for each design, and inlcude the length penalty each ligand. A sharp sign ('#') represents a triple bond. All values are in kcal/mol.

AVG SCOR	PEN	$\Delta E_{ads,\mathrm{E2}}$	$\Delta E_{ads,\mathrm{E6}}$	EVALUATIONS	SEQUENCE
1.60 ± 0.19	0.0	-7.2	-5.6	2	F–$C_{10}H_6$–CH=C=CH–(trans)CH=CH–H
1.50 ± 0.25	0.0	-10.3	-8.8	3	Cl–C(Ph)H–$C_{10}H_6$–CH=C=CH–CH(COOH)–CH(NH_2)–C(CH_3)$_3$
1.44 ± 0.21	-0.1	-7.4	-5.9	2	CH_2=C=CH–(trans)CH=CH–CH(NH_2)–(p)Ph–C≡C–(trans)CH=CH–CF_2–CH_3
1.39 ± 0.17	0.0	-6.8	-5.4	5	COOH–CH(COOH)–CH(COOH)–Cl
1.32 ± 0.18	0.0	-6.7	-5.7	2	SH–$C_{10}H_6$–CH_3
1.17 ± 0.24	0.0	-7.2	-6.7	2	CH_2=CH–CH(NH_2)–(p)Ph–(p)Ph–$C_{10}H_6$–CH(CH_3)–CH_3
1.17 ± 0.23	0.0	-9.4	-8.6	3	Cl–C(Ph)H–$C_{10}H_6$–C(Ph)H–CH_3
1.16 ± 0.20	0.0	-7.3	-6.2	3	F–$C_{10}H_6$–CH=C=CH–CH=C=CH–CH_3
1.12 ± 0.21	0.0	-8.1	-7.0	3	CH_2=CH–$C_{10}H_6$–C≡C–$C_{10}H_6$–H
1.10 ± 0.19	0.0	-7.5	-6.7	18	CH_2=CH–$C_{10}H_6$–CH=C=CH–H
1.10 ± 0.18	0.0	-6.9	-5.8	2	NH_2–$C_{10}H_6$–C≡C–CHO
1.08 ± 0.19	0.0	-7.4	-6.4	2	CH_2=CH–$C_{10}H_6$–$CONH_2$
1.04 ± 0.21	0.0	-8.3	-7.2	2	CH_2=CH–$C_{10}H_6$–CHOH–(p)Ph–NH_2
1.04 ± 0.20	0.0	-7.7	-6.7	9	CH_2=CH–$C_{10}H_6$–CO–CH(CH_3)–CH_3
1.02 ± 0.19	0.0	-7.7	-6.7	5	Cl–(o)Ph–COO–(o)Ph–COOH
0.99 ± 0.20	0.0	-7.6	-7.0	5	CH_2=CH–$C_{10}H_6$–CH=C=CH–CH_3
0.98 ± 0.21	0.0	-7.1	-6.2	2	CH_2=CH–CH(CH_3)–$C_{10}H_6$–CH(CH_3)–CH_3
0.98 ± 0.19	0.0	-7.5	-5.8	2	F–$C_{10}H_6$–CH=C=CH–CH(NH_2)–H
0.97 ± 0.21	0.0	-8.0	-7.1	3	CH_2=CH–$C_{10}H_6$–CO–$C_{10}H_6$–H
0.97 ± 0.19	0.0	-6.4	-5.4	5	F–CF_2–CF_2–$C_{10}H_6$–H
0.96 ± 0.19	0.0	-7.2	-6.3	6	CH_2=CH–$C_{10}H_6$–CH=C=CH–NH_2
0.93 ± 0.13	0.0	-6.8	-5.9	2	CH_2=CH–$C_{10}H_6$–OH
0.92 ± 0.22	0.0	-7.2	-6.2	3	CH(CH_3)$_2$–C(CH_3)$_2$–O–C≡C–$C_{10}H_6$–F
0.91 ± 0.17	0.0	-7.0	-6.1	11	CH_2=CH–$C_{10}H_6$–CO–H
0.90 ± 0.23	0.0	-7.0	-6.1	2	C(CH_3)$_3$–CH_2–CH(CH_3)–COO–(o)Ph–CH(CH_2CH_3)–CHO
0.89 ± 0.20	0.0	-7.6	-6.7	6	CH_2=CH–$C_{10}H_6$–CO–CH(CH_3)–H
0.89 ± 0.19	0.0	-7.0	-6.1	25	F–$C_{10}H_6$–CH=C=CH–NH_2
0.89 ± 0.16	0.0	-6.9	-6.0	142	CH_2=CH–$C_{10}H_6$–CF_2–H
0.88 ± 0.20	0.0	-7.8	-7.3	25	CH_2=CH–$C_{10}H_6$–CH(CH_3)–CH_3
0.88 ± 0.19	0.0	-7.5	-6.9	14	CH_2=CH–CH(CH_3)–$C_{10}H_6$–CH_3
0.88 ± 0.15	0.0	-7.1	-6.3	21	CH_2=CH–$C_{10}H_6$–CF_2–CH_3
0.88 ± 0.16	0.0	-7.4	-6.5	2	CH_2=CH–$C_{10}H_6$–CF_2–$C_{10}H_6$–CF_2–H
0.87 ± 0.15	0.0	-7.0	-6.2	602[a]	CH_2=CH–$C_{10}H_6$–CH_3
0.86 ± 0.20	0.0	-7.5	-6.7	2	CH_2=CH–$C_{10}H_6$–C(CH_3)$_3$
0.85 ± 0.20	0.0	-7.2	-6.4	23	CH_2=CH–C≡C–$C_{10}H_6$–CH(CH_3)–CH_3
0.83 ± 0.24	0.0	-9.3	-8.4	2	$CONH_2$–(o)Ph–$C_{10}H_6$–CH(COOH)–CH(CH_2CH_3)–NH–COOH
0.83 ± 0.22	0.0	-8.7	-7.9	13	F–$C_{10}H_6$–CH(CH_3)–C≡C–$C_{10}H_6$–H
0.83 ± 0.20	0.0	-6.0	-5.2	7	CH_2=CH–CF_2–CH(CH_3)–$C_{10}H_6$–H
0.83 ± 0.19	0.0	-6.9	-6.1	57	CH_2=CH–$C_{10}H_6$–CH(CH_3)–H
0.83 ± 0.18	0.0	-6.4	-5.7	2	F–$C_{10}H_6$–NH_2
0.82 ± 0.21	0.0	-8.2	-7.4	2	F–$C_{10}H_6$–CO–$C_{10}H_6$–CO–H
0.82 ± 0.21	0.0	-6.6	-5.8	2	SH–CH_2–C(Ph)H–(p)Ph–C≡C–CH=C=CH_2
0.82 ± 0.18	0.0	-6.6	-5.8	71	F–$C_{10}H_6$–CH=C=CH–H
0.82 ± 0.17	0.0	-7.5	-6.6	14	CH_2=CH–$C_{10}H_6$–CO–CH_3
0.80 ± 0.18	0.0	-6.4	-5.6	2	CH_2=CH–(p)Ph–CO–COO–SH

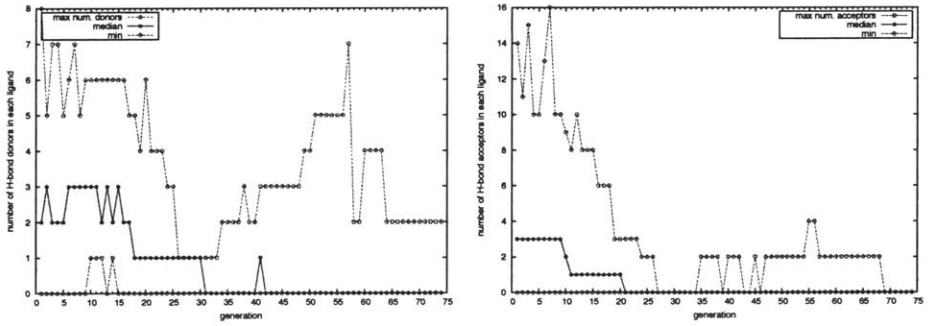
[a]Most prevalent design in final generation.



(a) Molecular volume and length in functional groups of each ligand.



(b) Number of rings and log of partition coefficient $\log P$ of each ligand.



(c) Number of hydrogen bond donors and acceptors in each ligand.

Figure B-13. Property distribution evolution in generations 1 through 74 in experiment I. In each figure, the filled points with solid lines represents the median value of that property, while the open points connected with dotted lines are the maximum and minimum value in that generation.

B.3.2 Experiment II: Forty-five generations of roulette-wheel selection after window scaling, using 6-ns production MD for evaluation

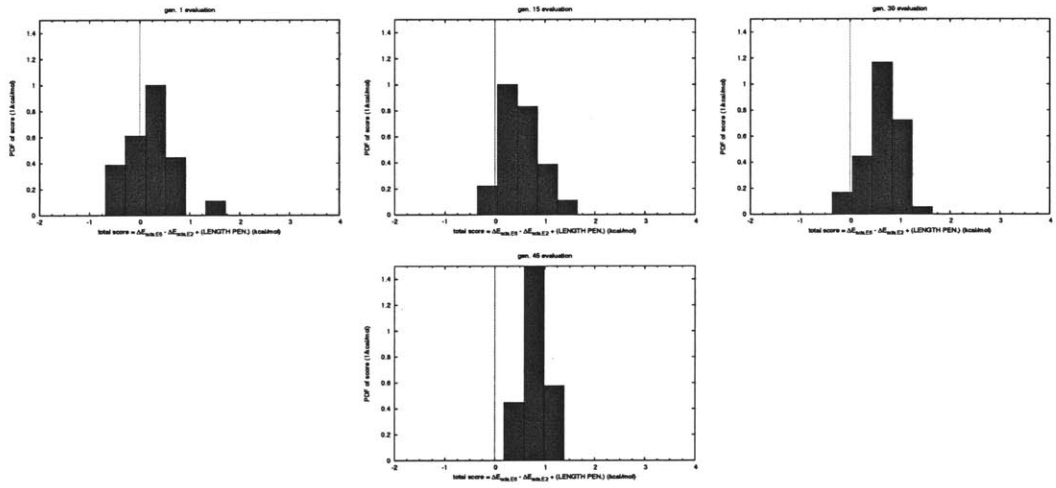
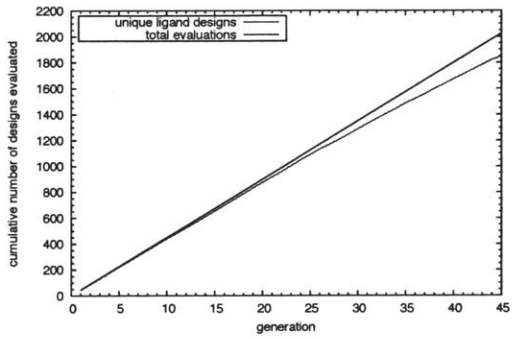
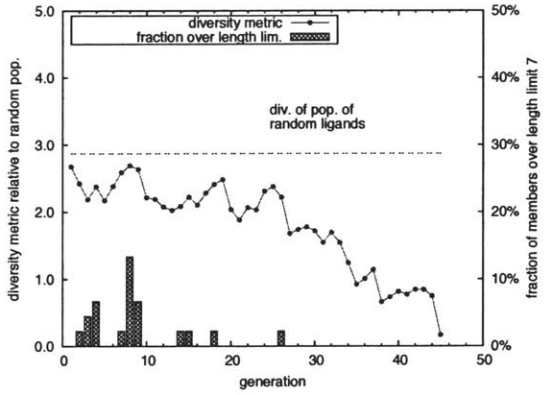


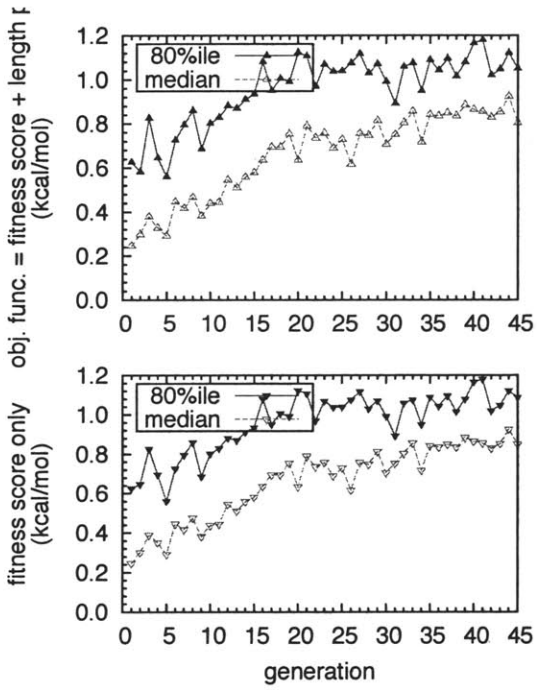
Figure B-14. Distribution of fitness scores (including length penalties) of members of generations 1, 15, 30, and 45 in experiment II.



(a) Cumulative number of ligand evaluations.



(b) Phenotypic diversity of evolving population of ligands. Bars show the fraction of ligands which exceed the length limit of 7 functional groups.

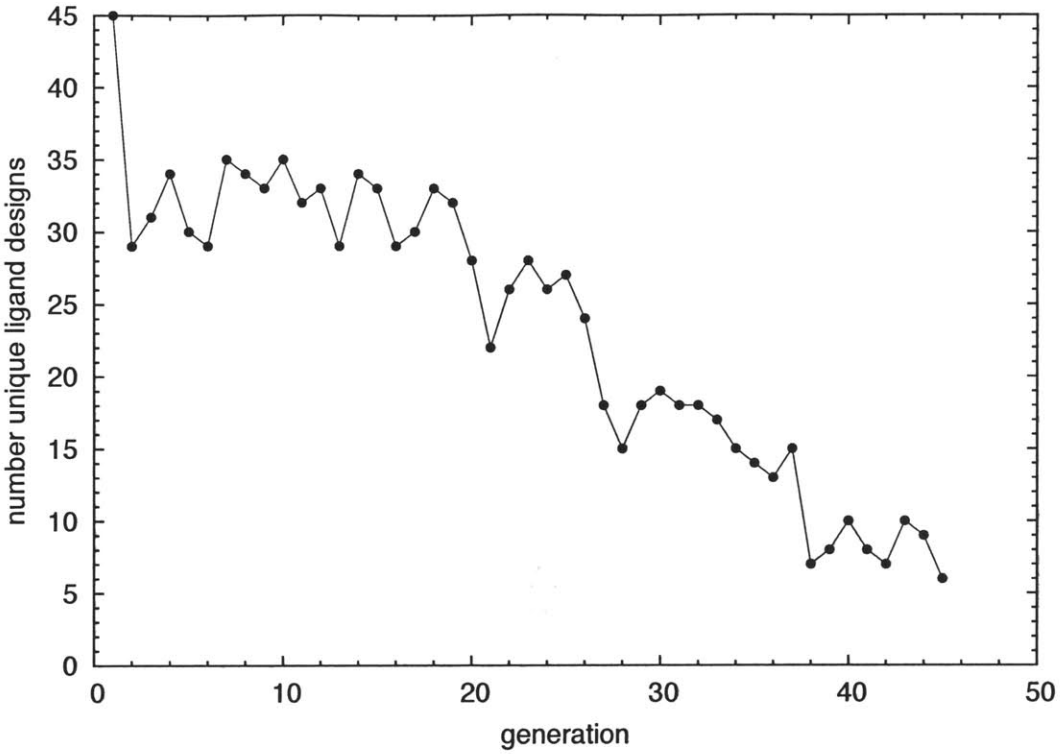


(c) Value of objective function at 80th and 50th percentiles in each generation. Depicted are the fitness score plus length penalty (*top*), and the finess score alone (*bottom*).

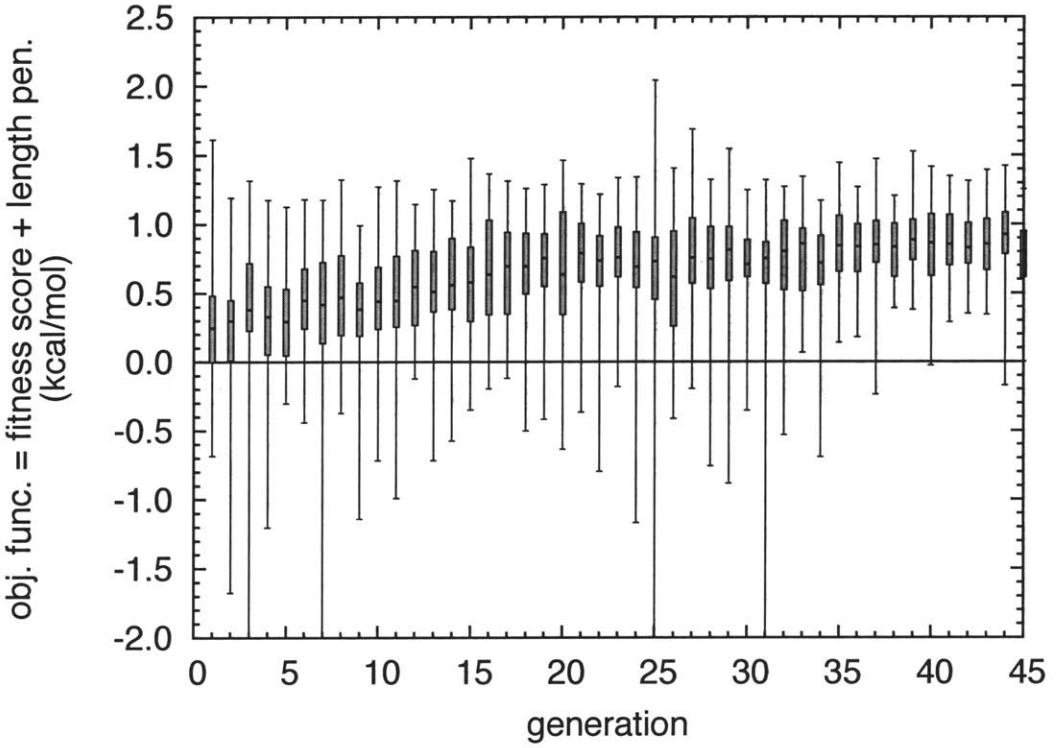
Figure B-15. Characterization of evolution over generations 1 to 45 in Experiment II, which featured $N = 45$ ligands, roulette wheel selection after window-based scaling, and 6.0-ns production MD in each evaluation.

Table C-5. Prevalence of structural motifs in three different populations in Experiment II. The "top 45" population are the top-scoring ligand candidates, ranked by average score, for which *at least* 12 ns of production MD was performed (as in Table C-6). These 45 top-scoring members had selectivity scores ranging from 1.0 to 0.72 kcal/mol $\approx 1.2kT$. Motif entries are listed in order of prevalence in the top-45 scorers.

	in gen. 1		in gen. 45		in top 45 scorers	
MOTIF	count (out of 45)	fraction (%)	count (out of 45)	fraction (%)	count (out of 45)	fraction (%)
aromatic group: phenyl or napthalene group	25	56	45	100	45	100
napthalene group	6	13	44	98	41	91
alkene/alkyne group	33	73	45	100	38	84
H-bond donor or acceptor group	40	89	2	4	37	82
H-bond acceptor group	40	89	2	4	37	82
H-bond donor group	34	76	2	4	34	76
aromatic group adjacent to H-bonding group	15	33	2	4	33	73
aromatic group adjacent to H-bond acceptor	15	33	2	4	33	73
aromatic group adjacent to alkene/alkyne group	11	24	45	100	30	67
aromatic group adjacent to H-bond donor	14	31	2	4	27	60
hydroxyl group	18	40	2	4	22	49
amine group	15	33	0	0	22	49
H-bond donor within 2 groups of acceptor group	19	42	0	0	19	42
aromatic group adjacent to amino group	5	11	0	0	18	40
H-bond donor adjacent to acceptor group	12	27	0	0	15	33
initial vinyl group	3	7	43	96	14	31
carbonyl group	20	44	0	0	14	31
aromatic group adjacent to hydroxyl group	8	18	2	4	13	29
aromatic group adjacent to carbonyl group	6	13	0	0	11	24
phenyl group	20	44	1	2	9	20
soft, bulky group	26	58	0	0	7	16
terminal phenyl ring	7	16	0	0	6	13
terminal halide (F or Cl)	0	0	0	0	6	13
isopropyl group	11	24	0	0	6	13
fluoro group	9	20	0	0	6	13
aromatic group adjacent to alkane group	9	20	0	0	5	11
aromatic group adjacent to fluoro group	2	4	0	0	4	9
aromatic group adjacent to bulky, soft group	6	13	0	0	4	9
aromatic group adjacent to isopropyl group	3	7	0	0	3	7
amino group within 2 units of internal phenyl ring	4	9	0	0	2	4
hydroxyl group within 2 units of internal phenyl ring	8	18	0	0	1	2
carboxyl group	10	22	0	0	1	2
aromatic group adjacent to carboxyl group	2	4	0	0	1	2
thiol group	4	9	0	0	0	0
ligand is a saturated alkane	0	0	0	0	0	0
ether or thioether group	11	24	0	0	0	0
E6-like: amino–any group–int. phenyl ring–hydroxyl	1	2	0	0	0	0
E2-like: hydroxyl–any group–int. phenyl ring–hydroxyl	0	0	0	0	0	0
aromatic group adjacent to thiol group	2	4	0	0	0	0
aromatic group adjacent to ether/thioether group	4	9	0	0	0	0



(a) Number of unique ligand designs (out of a total of 45) within each generation.



(b) Box-and-whisker plots of objective function (selection score + length penalty) measurements for members of each generation.

Figure B-16. Evolution dynamics in Experiment II.

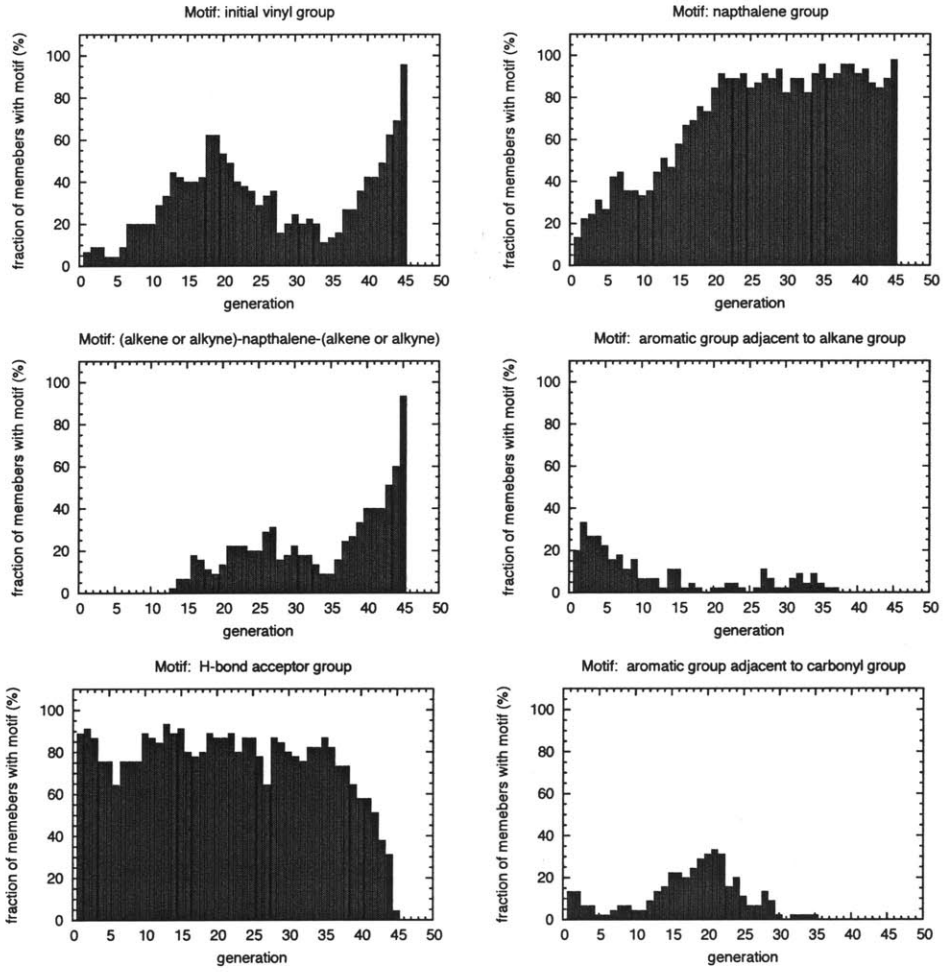


Figure B-17. Prevalence of motifs in generations 1 to 45 of Experiment II. Motif descriptions are listed in the title of each graph.

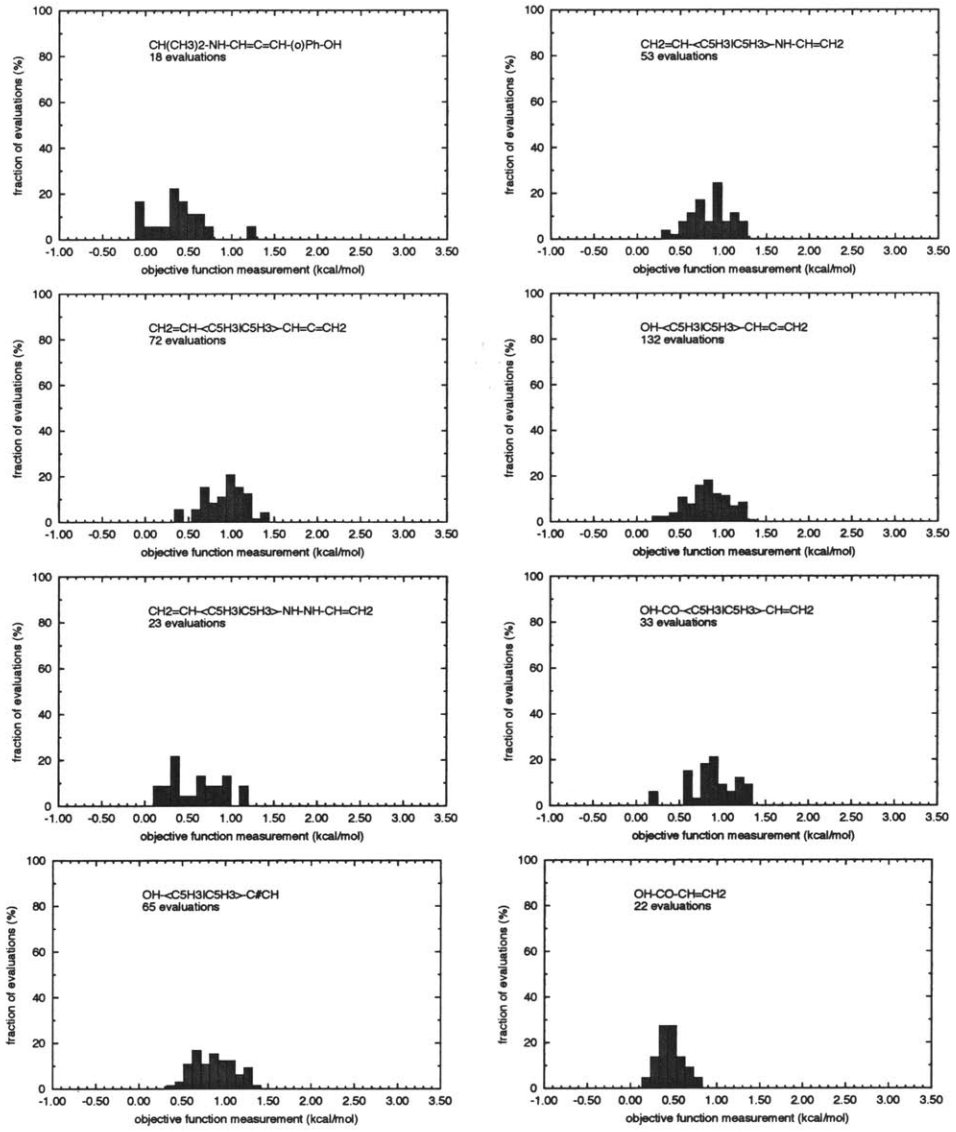


Figure B-18. Histograms of measured objective function values (fitness score plus length penalty) for frequently-occurring ligand designs in Experiment II, which used 6 ns production MD for evaluation.

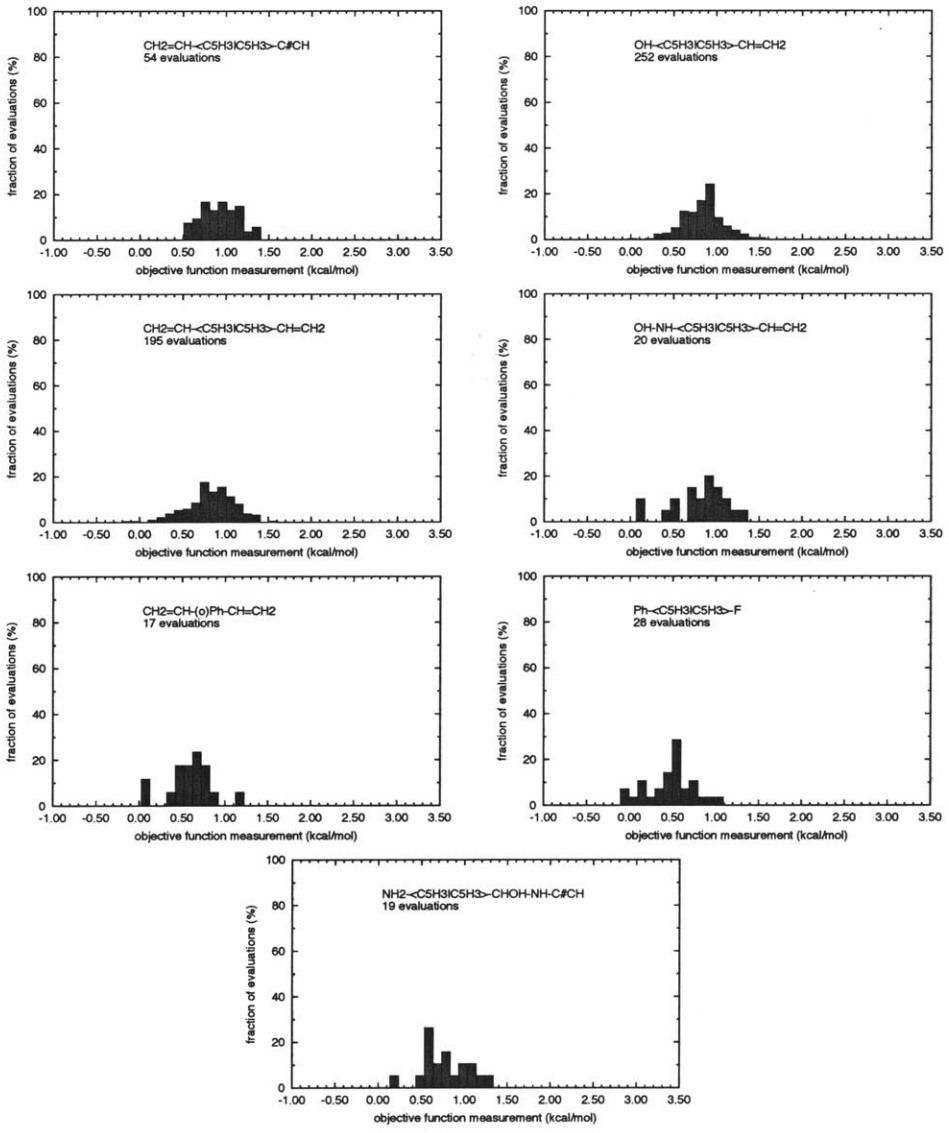


Figure B-19. Histograms of measured objective function values (fitness score plus length penalty) for frequently-occurring ligand designs in Experiment II, which used 6 ns production MD for evaluation.

Table C-6. Top-scoring forty-five ligand designs from generations 1 through 45 in Experiment II, *for which at least 12 ns of production MD* (corresponding to two ligand evaluations) was performed. Listed scores are averages of all evaluations for each design, and inlcude the length penalty each ligand. A sharp sign ('#') represents a triple bond. All values are in kcal/mol.

AVG SCOR	PEN	$\Delta E_{ads,\mathrm{E2}}$	$\Delta E_{ads,\mathrm{E6}}$	EVALUATIONS	SEQUENCE
1.04 ± 0.14	0.0	-7.8	-6.8	2	CH2=CH-C10H6-NH-CO-NH-CH=CH2
0.99 ± 0.14	0.0	-7.3	-6.3	8	OH-CH=C=CH-C10H6-CH=C=CH2
0.97 ± 0.14	0.0	-7.5	-6.5	3	CH2=CH-C10H6-CO-CH=CH2
0.95 ± 0.15	0.0	-7.7	-6.7	2	CH2=CH-(p)Ph-(m)Ph-CH(CH3)-NH-CH=CH2
0.94 ± 0.14	0.0	-7.8	-6.9	23	CH2=CH-C10H6-CO-NH-CH=CH2
0.93 ± 0.16	0.0	-9.1	-8.2	2	CH2=CH-C10H6-CHOH-C10H6-CH=CH2
0.93 ± 0.15	0.0	-7.7	-6.8	5	CH2=CH-C10H6-CH=C=CH-CO-NH-CH=CH2
0.93 ± 0.13	0.0	-7.3	-6.4	4	OH-CO-C10H6-C#CH
0.92 ± 0.14	0.0	-7.3	-6.4	2	CH#C-CH(COOH)-(p)Ph-(o)Ph-F
0.92 ± 0.13	0.0	-7.1	-6.2	72	CH2=CH-C10H6-CH=C=CH2
0.92 ± 0.13	0.0	-6.8	-5.9	54	CH2=CH-C10H6-C#CH
0.91 ± 0.14	0.0	-7.3	-6.4	2	CH2=CH-CO-C10H6-CH=CH2
0.88 ± 0.14	0.0	-7.3	-6.4	5	OH-CH(NH2)-C10H6-CH=CH2
0.87 ± 0.12	0.0	-6.2	-5.4	3	OH-C10H6-F
0.86 ± 0.15	0.0	-7.5	-6.6	2	CH(CH3)2-C10H6-CH=C=CH-CH=CH2
0.86 ± 0.13	0.0	-7.5	-6.6	35	OH-CO-C10H6-CH=CH2
0.86 ± 0.13	0.0	-6.6	-5.8	7	CH2=CH-C10H6-OH
0.85 ± 0.14	0.0	-7.1	-6.2	53	CH2=CH-C10H6-NH-CH=CH2
0.84 ± 0.15	0.0	-8.1	-7.2	4	CH(CH3)2-CO-C10H6-NH-CH=CH2
0.84 ± 0.13	0.0	-7.2	-6.4	7	OH-NH-C10H6-CH=C=CH2
0.84 ± 0.13	0.0	-6.9	-6.1	188[a]	CH2=CH-C10H6-CH=CH2
0.84 ± 0.12	0.0	-6.1	-5.3	8	NH2-C10H6-F
0.83 ± 0.14	0.0	-7.3	-6.5	24	OH-C10H6-NH-CH=CH2
0.83 ± 0.14	0.0	-7.0	-6.2	14	CH(CH3)2-C10H6-C#CH
0.83 ± 0.13	0.0	-7.6	-6.8	5	OH-CO-C10H6-NH-CH=CH2
0.83 ± 0.13	0.0	-6.9	-6.0	20	OH-NH-C10H6-CH=CH2
0.82 ± 0.15	0.0	-8.6	-7.7	3	OH-(p)Ph-CO-C10H6-NH-CONH2
0.82 ± 0.15	0.0	-7.7	-6.9	3	Ph-C10H6-C#C-C#C-NH-CH3
0.82 ± 0.14	0.0	-7.4	-6.6	3	Ph-C10H6-C#C-C#C-F
0.82 ± 0.13	0.0	-6.9	-6.1	298	OH-C10H6-CH=CH2
0.82 ± 0.13	0.0	-6.9	-6.1	132	OH-C10H6-CH=C=CH2
0.81 ± 0.14	0.0	-7.9	-7.1	10	OH-CO-C10H6-NH-NH-CH=CH2
0.81 ± 0.14	0.0	-7.1	-6.3	2	CH(CH3)2-CO-(p)Ph-(o)Ph-F
0.80 ± 0.13	0.0	-6.8	-6.0	2	NH2-C10H6-CHOH-F
0.80 ± 0.13	0.0	-6.7	-5.9	72	OH-C10H6-C#CH
0.78 ± 0.14	0.0	-7.7	-6.9	15	OH-C10H6-NH-NH-CH=CH2
0.78 ± 0.14	0.0	-7.5	-6.8	5	CH2=CH-C10H6-Ph
0.78 ± 0.12	0.0	-4.9	-4.1	2	CH#C-(p)Ph-NH-CH3
0.77 ± 0.14	0.0	-7.3	-6.6	19	NH2-C10H6-CHOH-NH-C#CH
0.76 ± 0.14	0.0	-7.5	-6.7	14	OH-C10H6-Ph
0.74 ± 0.15	0.0	-8.5	-7.7	5	CH2=CH-C10H6-NH-NH-CO-NH-CH=CH2
0.74 ± 0.14	0.0	-7.1	-6.4	7	OH-NH-NH-C10H6-CH=CH2
0.74 ± 0.14	0.0	-7.1	-6.3	6	OH-C10H6-C(CH3)3
0.73 ± 0.15	0.0	-7.9	-7.1	5	CH(CH3)2-CO-C10H6-NH-CH=C=CH2
0.72 ± 0.14	0.0	-7.3	-6.6	5	CH(CH3)2-C10H6-NH-CH=CH2

[a]Most prevalent design in final generation.

B.3.3 Experiment III: Eighty-nine generations of tournament selection with no scaling, using 4.5-ns production MD and accumulative scoring for evaluation

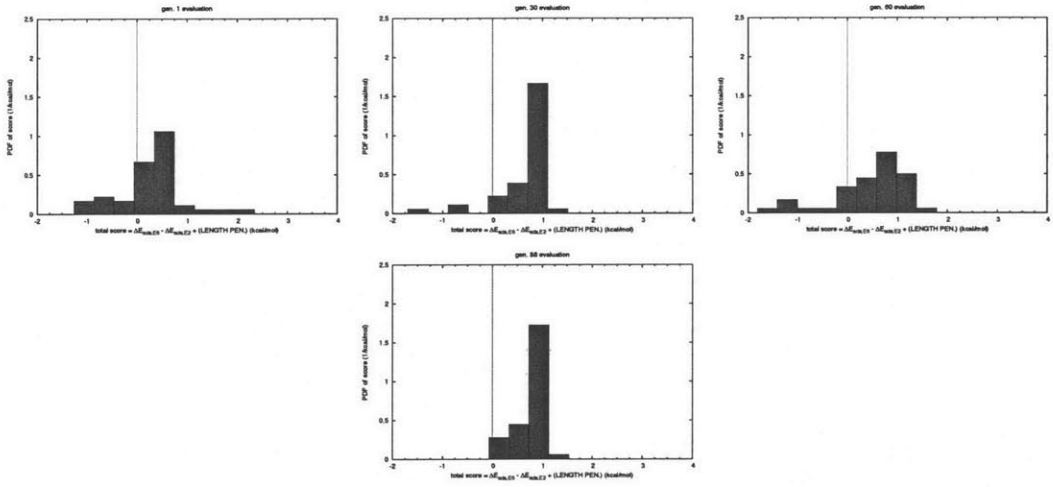
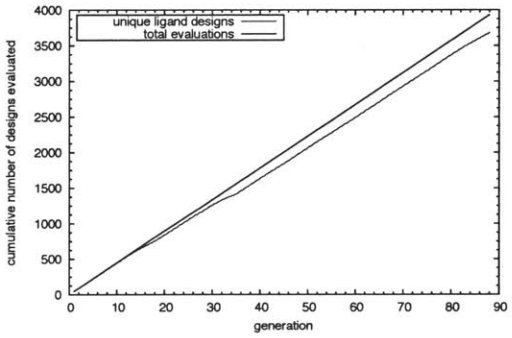


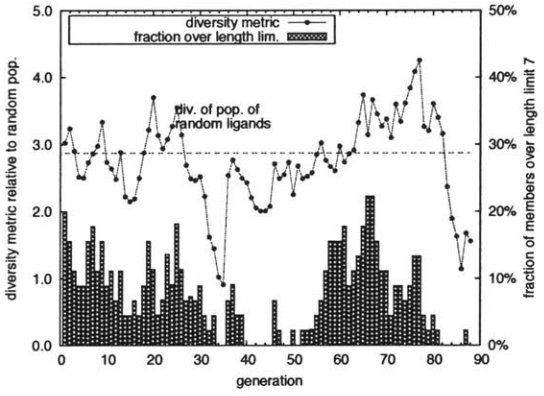
Figure B-20. Distribution of fitness scores (including length penalties) of members of generations 1, 30, 68, and 88 in Experiment III.

Table C-7. Prevalence of structural motifs in three different populations of Experiment III. The "top 15" population are the top-scoring ligand candidates, ranked by average score, for which *at least* 13.5 ns of production MD was performed (as in Table C-8). These 15 top-scoring members had selectivity scores ranging from 1.8 to 0.52 kcal/mol $\approx kT$. Motif entries are listed in order of prevalence in the top-15 scorers.

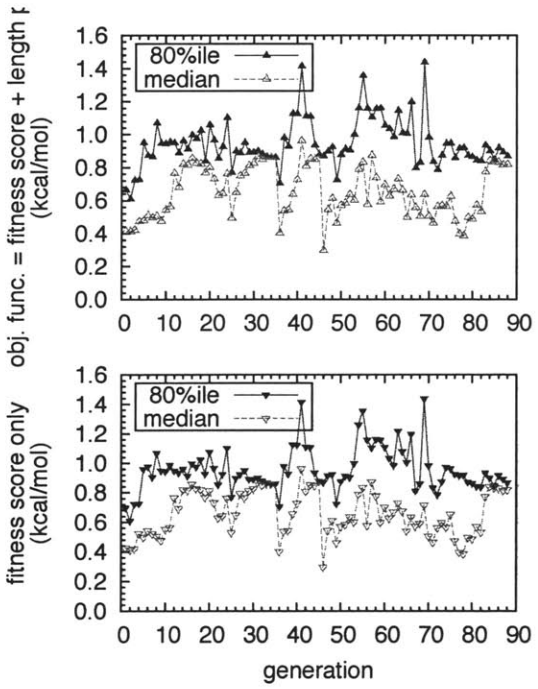
	in gen. 1		in gen. 88		in top 11 scorers	
MOTIF	count (out of 45)	fraction (%)	count (out of 45)	fraction (%)	count (out of 15)	fraction (%)
aromatic group: phenyl or napthalene group	28	62	39	87	9	56
H-bond donor or acceptor group	40	89	45	100	8	50
H-bond acceptor group	40	89	45	100	8	50
ether or thioether group	10	22	13	29	8	50
fluoro group	12	27	0	0	7	44
soft, bulky group	28	62	4	9	6	38
carbonyl group	23	51	45	100	5	31
aromatic group adjacent to ether/thioether group	6	13	7	16	5	31
aromatic group adjacent to alkane group	14	31	23	51	5	31
alkene/alkyne group	35	78	0	0	5	31
terminal phenyl ring	6	13	0	0	4	25
phenyl group	21	47	4	9	4	25
aromatic group adjacent to bulky, soft group	12	27	1	2	4	25
H-bond donor group	32	71	41	91	3	19
napthalene group	6	13	35	78	2	12
hydroxyl group	17	38	0	0	2	12
aromatic group adjacent to fluoro group	3	7	0	0	2	12
aromatic group adjacent to alkene/alkyne group	16	36	0	0	2	12
carboxyl group	8	18	0	0	1	6
aromatic group adjacent to H-bonding group	17	38	36	80	1	6
aromatic group adjacent to H-bond acceptor	17	38	36	80	1	6
aromatic group adjacent to carbonyl group	9	20	36	80	1	6
thiol group	4	9	0	0	0	0
terminal halide (F or Cl)	0	0	0	0	0	0
ligand is a saturated alkane	0	0	0	0	0	0
isopropyl group	12	27	0	0	0	0
initial vinyl group	1	2	0	0	0	0
hydroxyl group within 2 units of internal phenyl ring	5	11	0	0	0	0
H-bond donor within 2 groups of acceptor group	18	40	9	20	0	0
H-bond donor adjacent to acceptor group	12	27	1	2	0	0
E6-like: amino–any group–int. phenyl ring–hydroxyl	1	2	0	0	0	0
E2-like: hydroxyl–any group–int. phenyl ring–hydroxyl	0	0	0	0	0	0
aromatic group adjacent to thiol group	1	2	0	0	0	0
aromatic group adjacent to isopropyl group	3	7	0	0	0	0
aromatic group adjacent to hydroxyl group	7	16	0	0	0	0
aromatic group adjacent to H-bond donor	12	27	33	73	0	0
aromatic group adjacent to carboxyl group	0	0	0	0	0	0
aromatic group adjacent to amino group	6	13	0	0	0	0
amino group within 2 units of internal phenyl ring	5	11	0	0	0	0
amine group	15	33	0	0	0	0



(a) Cumulative number of ligand evaluations.

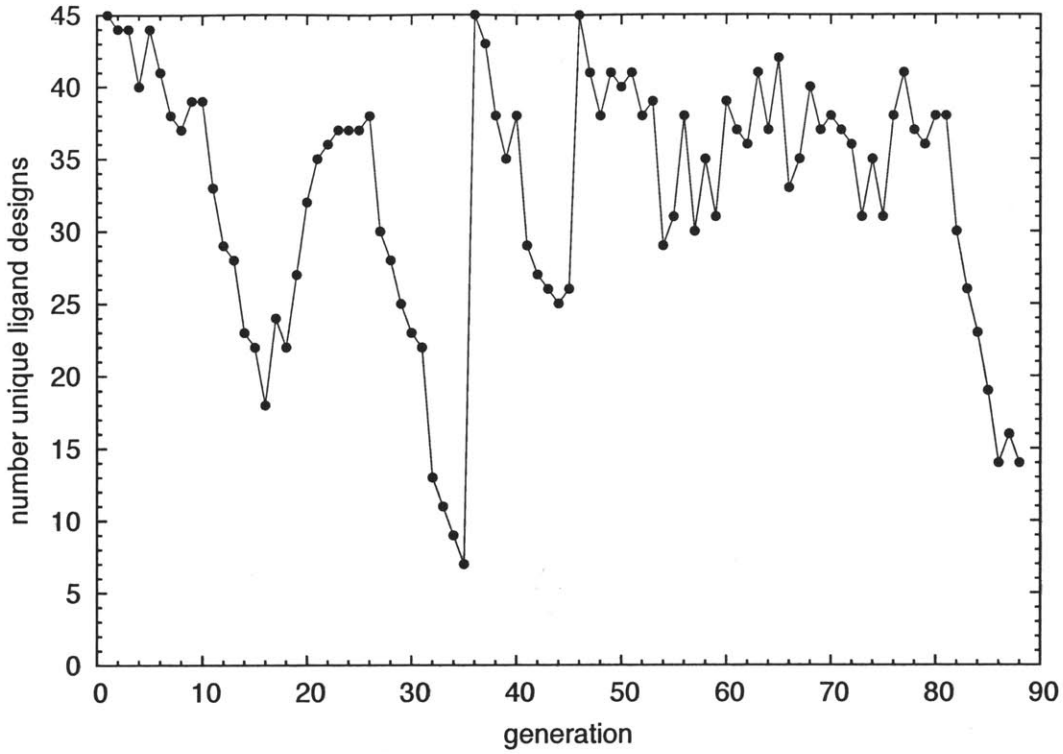


(b) Phenotypic diversity of evolving population of ligands. Bars show the fraction of ligands which exceed the length limit of 7 functional groups.

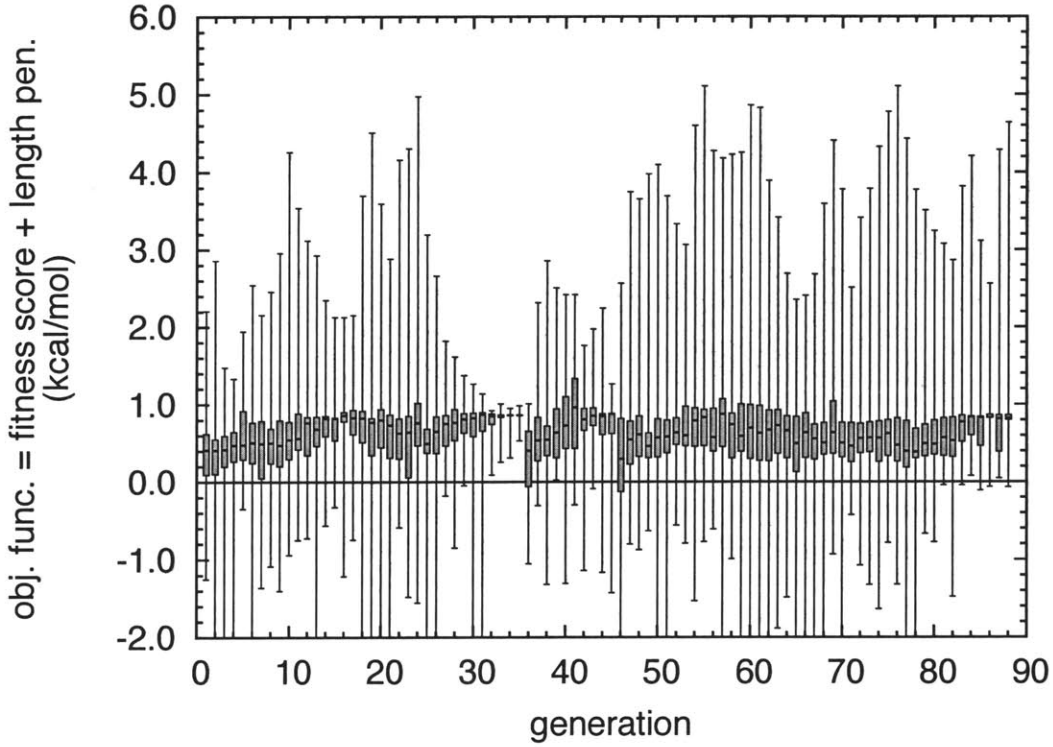


(c) Value of objective function at 80th and 50th percentiles in each generation. Depicted are the fitness score plus length penalty (*top*), and the finess score alone (*bottom*).

Figure B-21. Characterization of evolution over generations 1 to 88 in Experiment III. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).



(a) Number of unique ligand designs (out of a total of 88) within each generation.



(b) Box-and-whisker plots of objective function (selection score + length penalty) measurements for members of each generation.

Figure B-22. Evolution dynamics in Experiment III. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

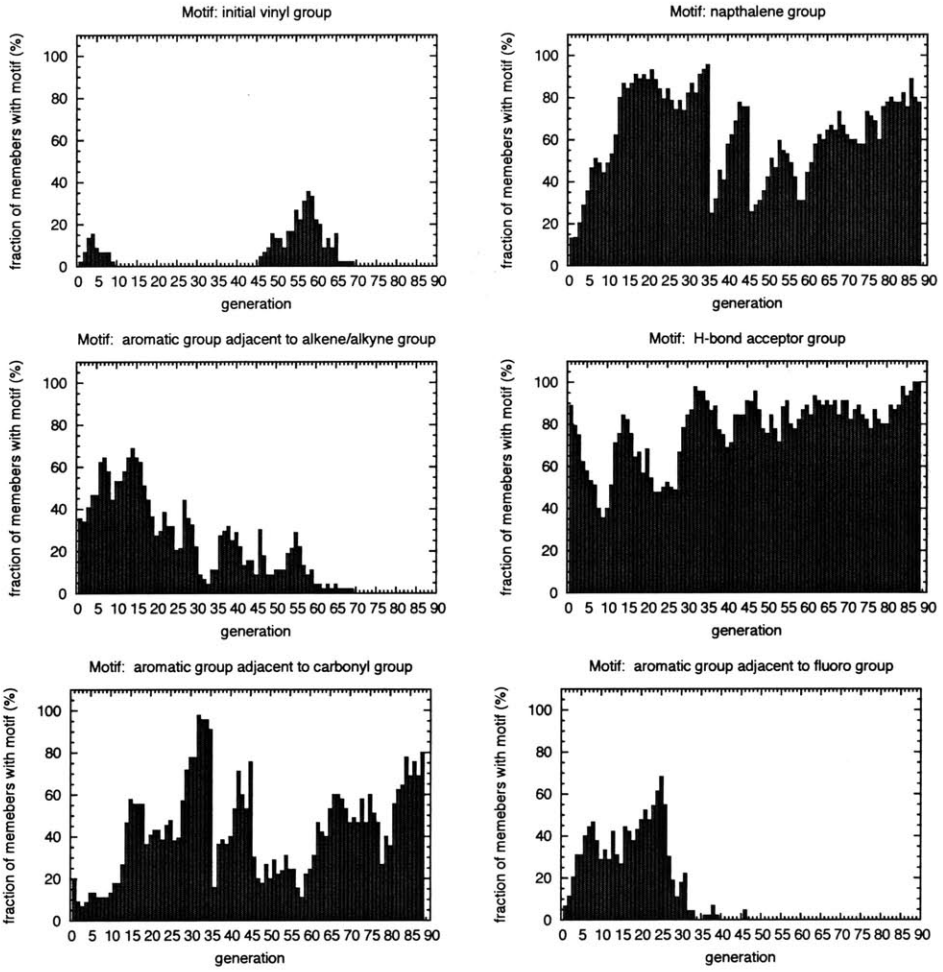


Figure B-23. Prevalence of motifs in generations 1 to 74 of Experiment III. Motif descriptions are listed in the title of each graph. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

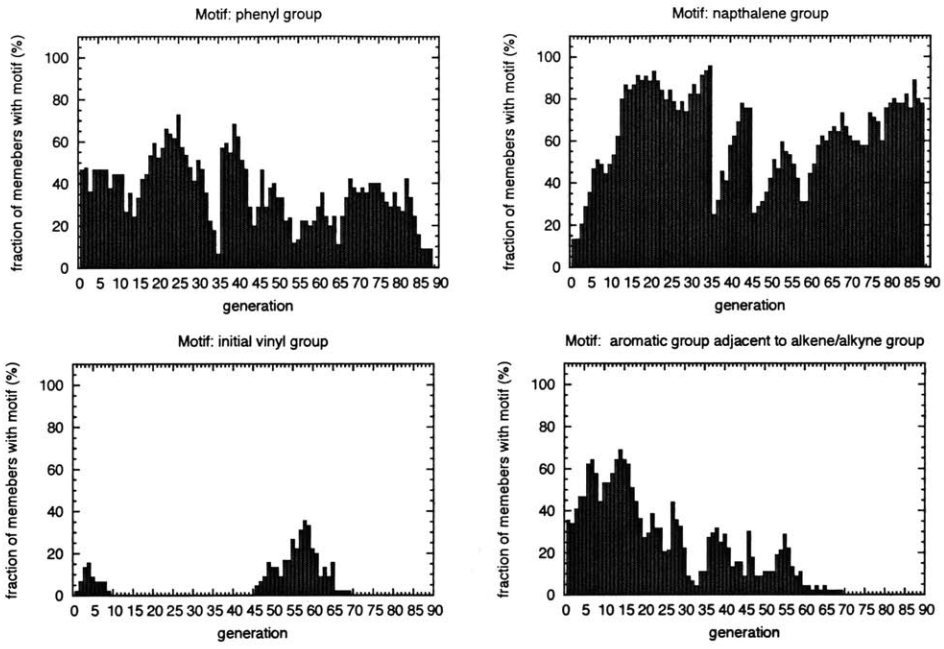


Figure B-24. Prevalence of motifs involving unsaturated/aromatic groups in generations 1 to 75 in Experiment III. Motif descriptions are listed in the title of each graph. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

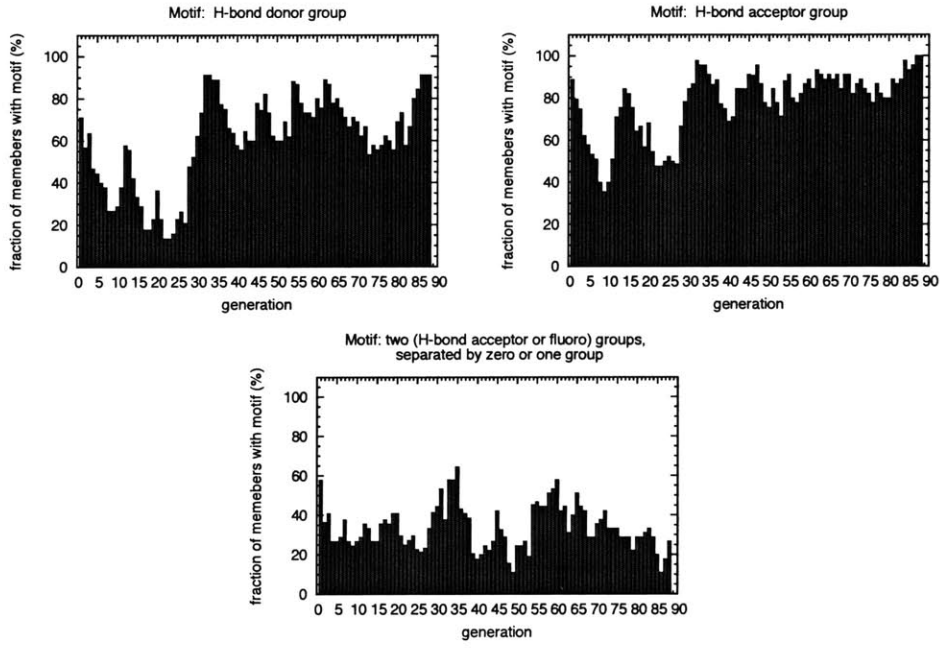


Figure B-25. Prevalence of motifs involving hydrogen-bond donors and acceptors in generations 1 to 75 of Experiment III. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

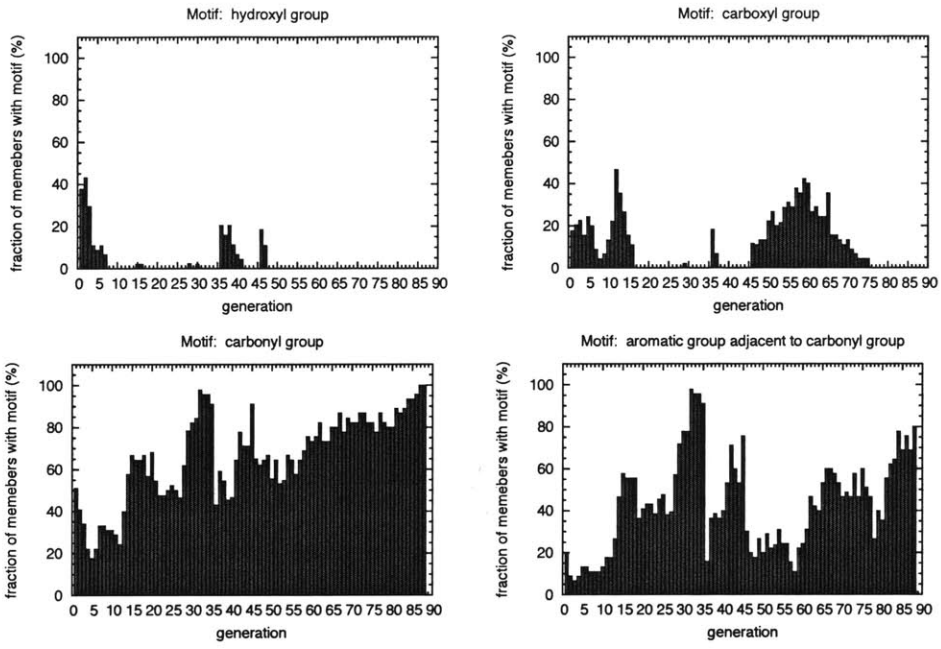


Figure B-26. Prevalence of motifs involving oxygen- or sulfur-containing groups in generations 1 to 75 in Experiment III. Note that the "carbonyl groups" do not include the carbonyl carbon within carboyxl groups. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

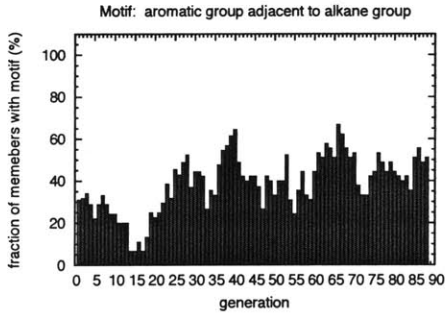


Figure B-27. Prevalence of other motifs in generations 1 to 75 in Experiment III. Note that ether groups can consist of either oxygen-based ether groups or thioether groups. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

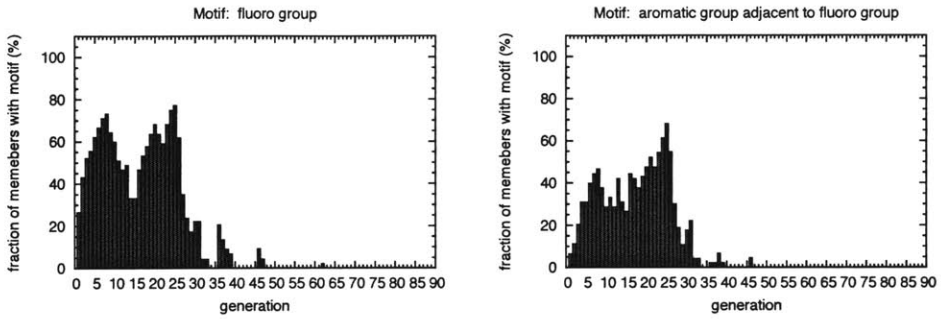


Figure B-28. Prevalence of motifs involving halide and amino groups in generations 1 to 75 in Experiment III. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

Table C-8. Top-scoring ligand designs from generations 1 through 88 in Experiment III, *for which at least 9.0 ns of production MD* (corresponding to two ligand evaluations) was performed. Listed scores are averages of all evaluations for each design, and inlcude the length penalty for each ligand. All values are in kcal/mol.

AVG SCOR	PEN	$\Delta E_{ads,\mathrm{E2}}$	$\Delta E_{ads,\mathrm{E6}}$	prod. MD (ns)	SEQUENCE
1.82 ± 0.13	0.0	-7.7	-5.7	9.0	CH_2=C=CH−$C_{10}H_6$−CF_2−O−C(Ph)H−$C(CH_3)_2$−CHO
1.10 ± 0.15	0.0	-7.9	-6.9	9.0	CH_2=C=CH−$C_{10}H_6$−CF_2−CH_3
1.05 ± 0.10	0.0	-8.8	-7.7	13.5	CH_2=C=CH−$C_{10}H_6$−CF_2−O−$C(CH_3)_2$−CHOH−Cl
1.01 ± 0.10	0.0	-8.5	-7.3	18.0	CH_2=C=CH−$C_{10}H_6$−CF_2−O−C(Ph)H−$CH(CH_2CH_3)$−CH_3
0.96 ± 0.11	0.0	-6.2	-5.3	9.0	F−CH=C=CH−(o)Ph−CF_2−COOH
0.95 ± 0.11	0.0	-7.4	-6.5	9.0	COOH−C≡C−O−$C_{10}H_6$−Cl
0.82 ± 0.14	-0.1	-7.1	-5.9	9.0	$CONH_2$−CF_2−$C_{10}H_6$−CF_2−O−C(Ph)H−$C(CH_3)_2$−CHO
0.82 ± 0.10	0.0	-5.7	-5.0	9.0	Cl−S−(p)Ph−CH=C=CH−CO−Cl
0.67 ± 0.09	0.0	-4.3	-3.7	13.5	$C(CH_3)_3$−CH_2−S−$CH=CH_2$
0.62 ± 0.12	0.0	-5.6	-5.0	9.0	CH_2=CH−NH−O−C(Ph)H−$CH(CH_2CH_3)$−CH_3
0.61 ± 0.13	0.0	-5.8	-5.2	9.0	H−CH(iBut)−CHOH−C≡C−CF_2−$C(CH_3)_2$−$C(CH_3)_3$
0.59 ± 0.08	0.0	-6.6	-6.1	18.0	F−CH=C=CH−(trans)CH=CH−CO−(o)Ph−CH=C=CH_2
0.57 ± 0.14	0.0	-6.1	-5.4	18.0	F−CH=C=CH−(o)Ph−CH_3
0.54 ± 0.09	0.0	-3.5	-2.9	9.0	F−(trans)CH=CH−CO−Cl
0.52 ± 0.09	0.0	-5.4	-4.9	13.5	$C(CH_3)_3$−CH(COOH)−S−$CH=CH_2$

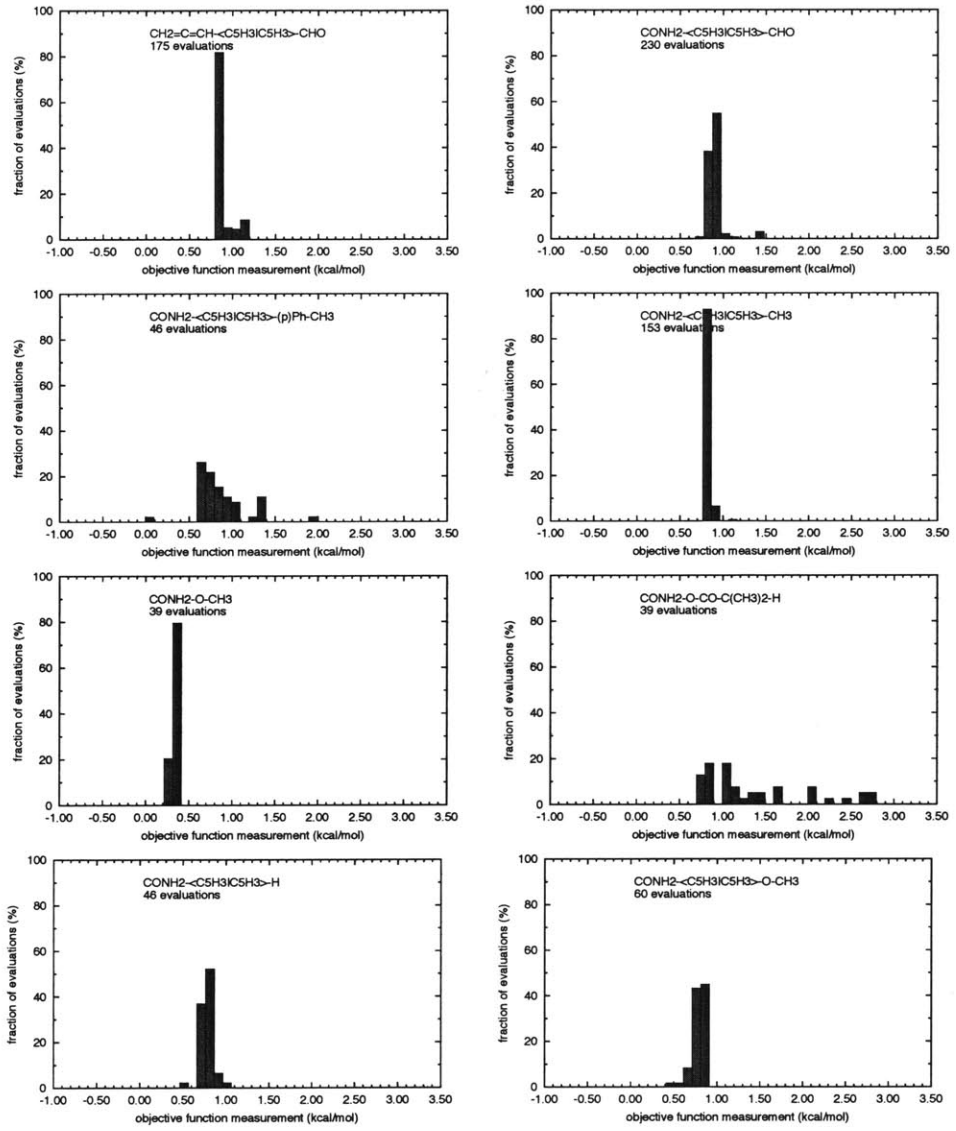
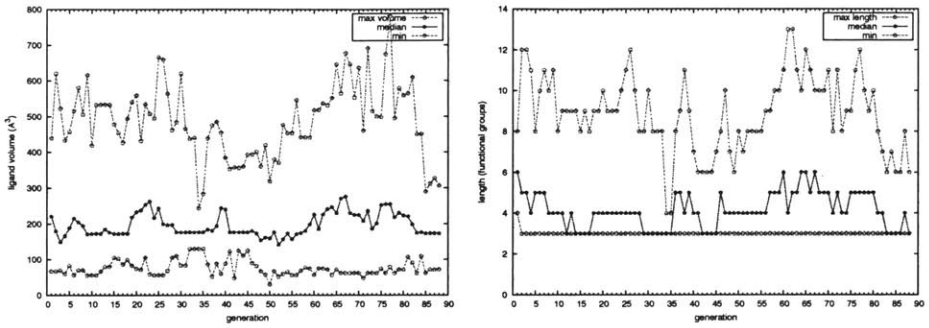
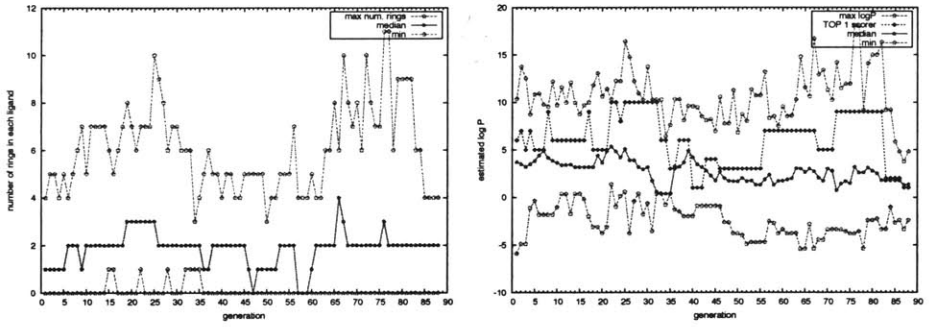


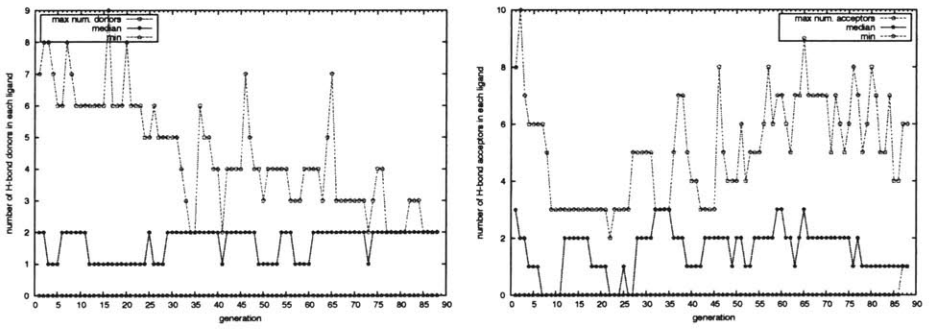
Figure B-29. Histograms of measured objective function values (fitness score plus length penalty) for frequently-occurring ligand designs in Experiment III. This experiment used 4.5 ns production MD, with *accumulative scoring* for evaluation. The latter explains why the fitness scores are generally more consistent than those in Experiment II (shown in Figures B-18 and B-19).



(a) Molecular volume and length in functional groups of each ligand.



(b) Number of rings and log of partition coefficient $\log P$ of each ligand.



(c) Number of hydrogen bond donors and acceptors in each ligand.

Figure B-30. Property distribution evolution in generations 1 through 88 in experiment III. In each figure, the filled points with solid lines represents the median value of that property, while the open points connected with dotted lines are the maximum and minimum value in that generation. The population of ligand designs in Experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

B.3.4 Experiment IV: Sixty-eight generations of "fuzzy" tournament selection, using 6-ns production MD and accumulative scoring for evaluation

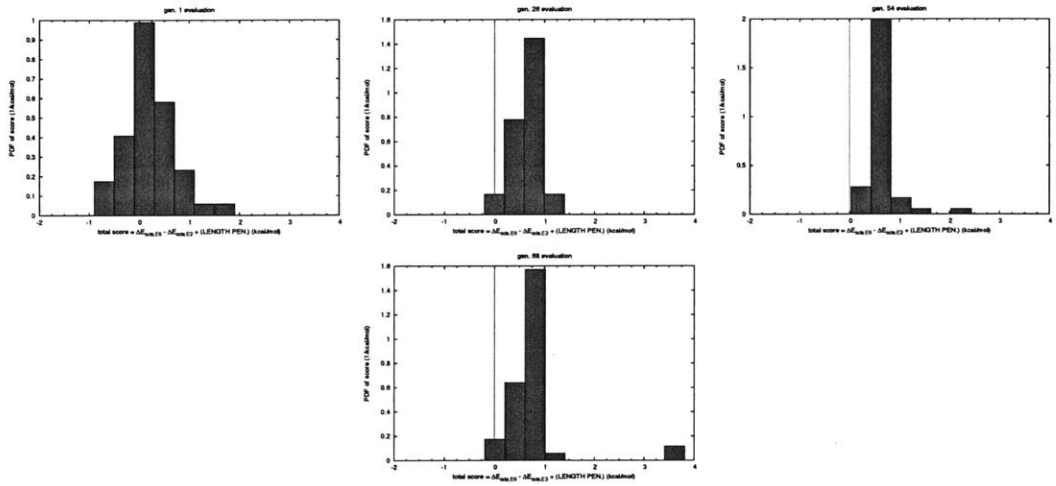
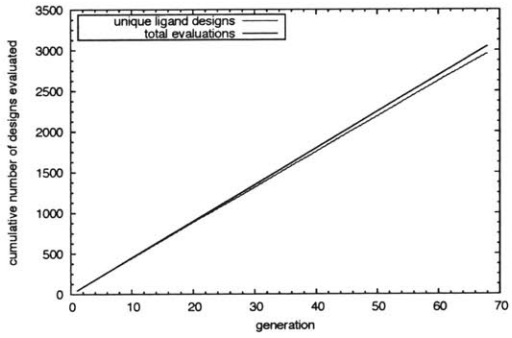


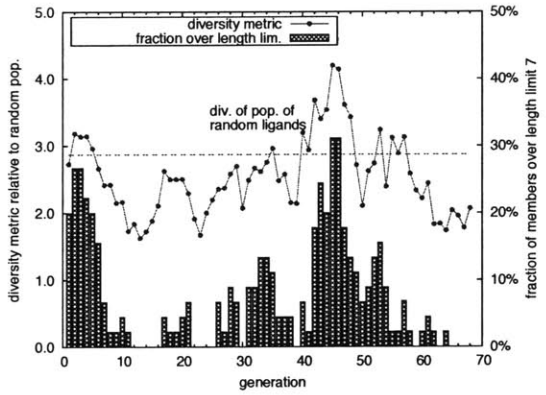
Figure B-31. Distribution of fitness scores (including length penalties) of members of generations 1, 26, 54, and 68 in experiment IV.

Table C-9. Prevalence of structural motifs in three different populations in Experiment IV. The "top 15" population are the top-scoring ligand candidates, ranked by average score, for which *at least* 18 ns of production MD was performed (as in Table C-10). These 90 top-scoring members had selectivity scores ranging from 3.7 to 0.84 kcal/mol $\approx 1.4kT$. Motif entries are listed in order of prevalence in the top-15 scorers.

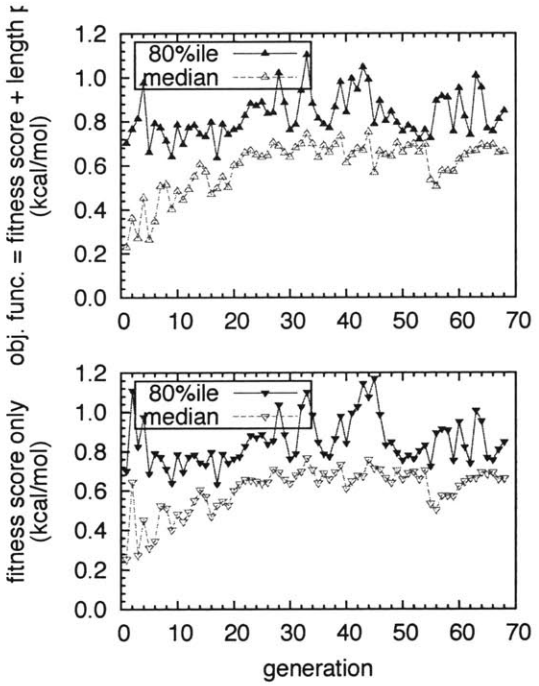
	in gen. 1		in gen. 75		in top 15 scorers	
MOTIF	count (out of 45)	fraction (%)	count (out of 45)	fraction (%)	count (out of 15)	fraction (%)
phenyl group	17	40	42	95	10	67
aromatic group: phenyl or napthalene group	20	47	42	95	10	67
H-bond donor or acceptor group	40	93	22	50	7	47
H-bond acceptor group	40	93	22	50	7	47
soft, bulky group	26	60	5	11	5	33
H-bond donor group	37	86	12	27	4	27
carbonyl group	17	40	13	30	4	27
isopropyl group	9	21	5	11	3	20
aromatic group adjacent to H-bonding group	11	26	20	45	3	20
aromatic group adjacent to H-bond acceptor	11	26	20	45	3	20
thiol group	7	16	0	0	2	13
ether or thioether group	9	21	6	14	2	13
carboxyl group	13	30	11	25	2	13
aromatic group adjacent to H-bond donor	9	21	12	27	2	13
aromatic group adjacent to ether/thioether group	1	2	6	14	2	13
aromatic group adjacent to alkane group	11	26	21	48	2	13
H-bond donor within 2 groups of acceptor group	18	42	3	7	1	7
fluoro group	13	30	3	7	1	7
aromatic group adjacent to thiol group	1	2	0	0	1	7
aromatic group adjacent to carboxyl group	1	2	11	25	1	7
aromatic group adjacent to carbonyl group	2	5	11	25	1	7
aromatic group adjacent to bulky, soft group	6	14	0	0	1	7
terminal phenyl ring	6	14	0	0	0	0
terminal halide (F or Cl)	0	0	0	0	0	0
napthalene group	3	7	0	0	0	0
ligand is a saturated alkane	0	0	0	0	0	0
initial vinyl group	4	9	0	0	0	0
hydroxyl group within 2 units of internal phenyl ring	4	9	1	2	0	0
hydroxyl group	14	33	1	2	0	0
H-bond donor adjacent to acceptor group	15	35	0	0	0	0
E6-like: amino–any group–int. phenyl ring–hydroxyl	0	0	0	0	0	0
E2-like: hydroxyl–any group–int. phenyl ring–hydroxyl	0	0	0	0	0	0
aromatic group adjacent to isopropyl group	1	2	0	0	0	0
aromatic group adjacent to hydroxyl group	3	7	1	2	0	0
aromatic group adjacent to fluoro group	2	5	3	7	0	0
aromatic group adjacent to amino group	5	12	0	0	0	0
aromatic group adjacent to alkene/alkyne group	10	23	1	2	0	0
amino group within 2 units of internal phenyl ring	4	9	0	0	0	0
amine group	17	40	0	0	0	0
alkene/alkyne group	35	81	1	2	0	0



(a) Cumulative number of ligand evaluations.

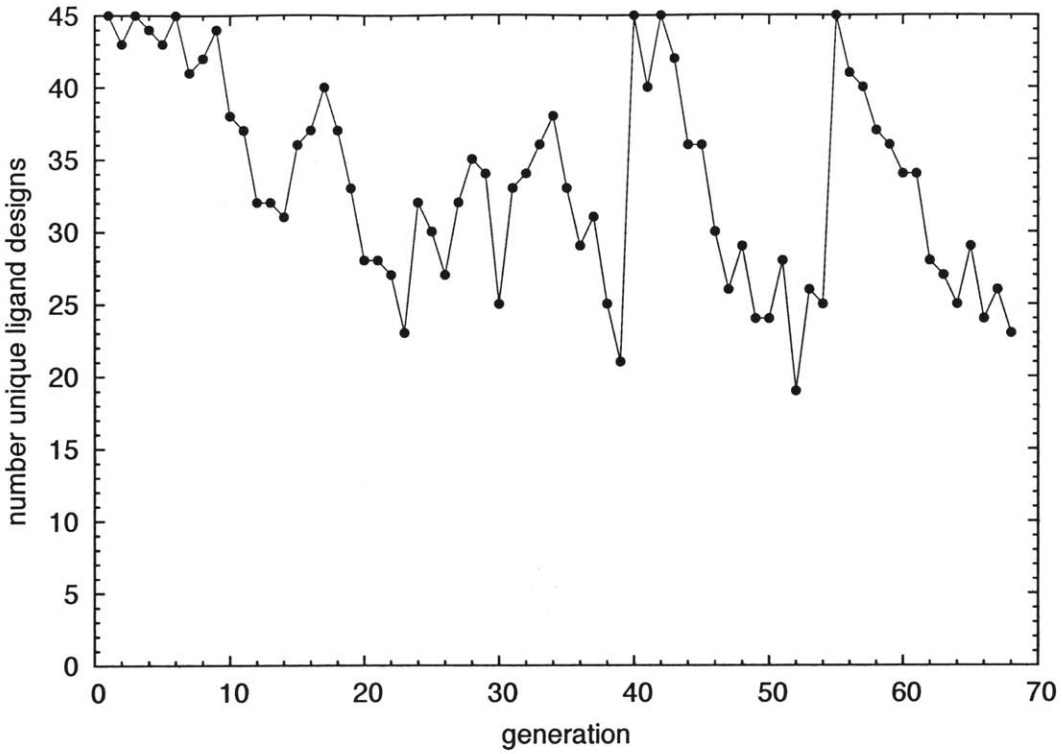


(b) Phenotypic diversity of evolving population of ligands. Bars show the fraction of ligands which exceed the length limit of 7 functional groups.

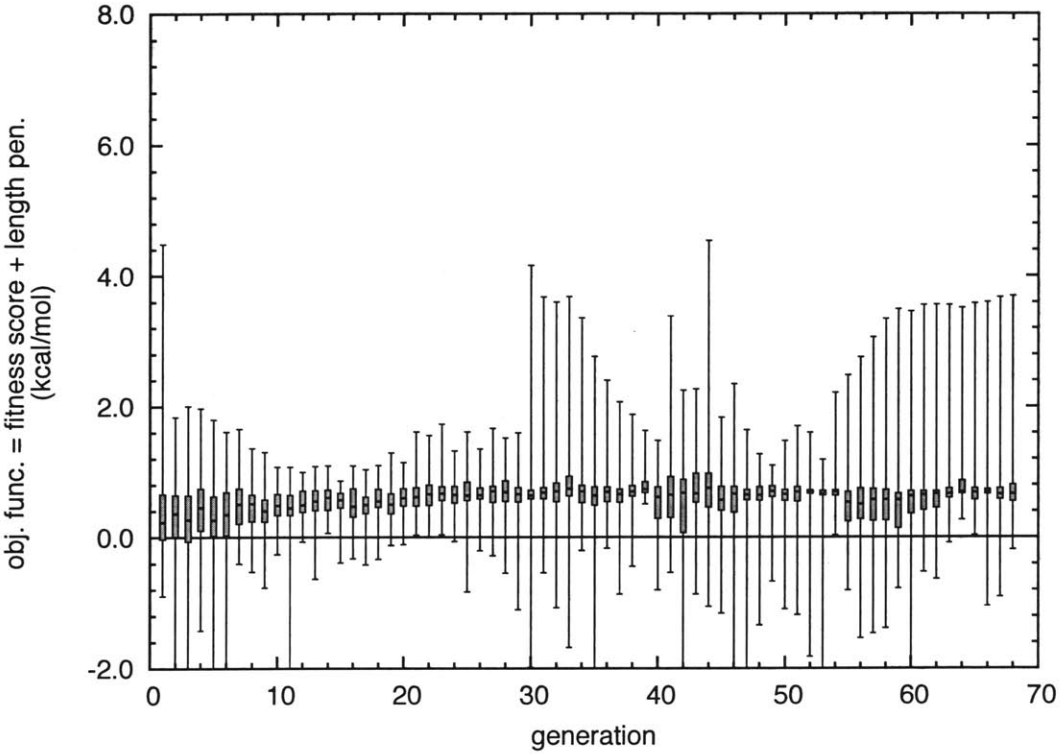


(c) Value of objective function at 80th and 50th percentiles in each generation. Depicted are the fitness score plus length penalty (*top*), and the finess score alone (*bottom*).

Figure B-32. Characterization of evolution over generations 1 to 68 in Experiment IV. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).



(a) Number of unique ligand designs (out of a total of 45) within each generation.



(b) Box-and-whisker plots of objective function (selection score + length penalty) measurements for members of each generation.

Figure B-33. Evolution dynamics in Experiment IV. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

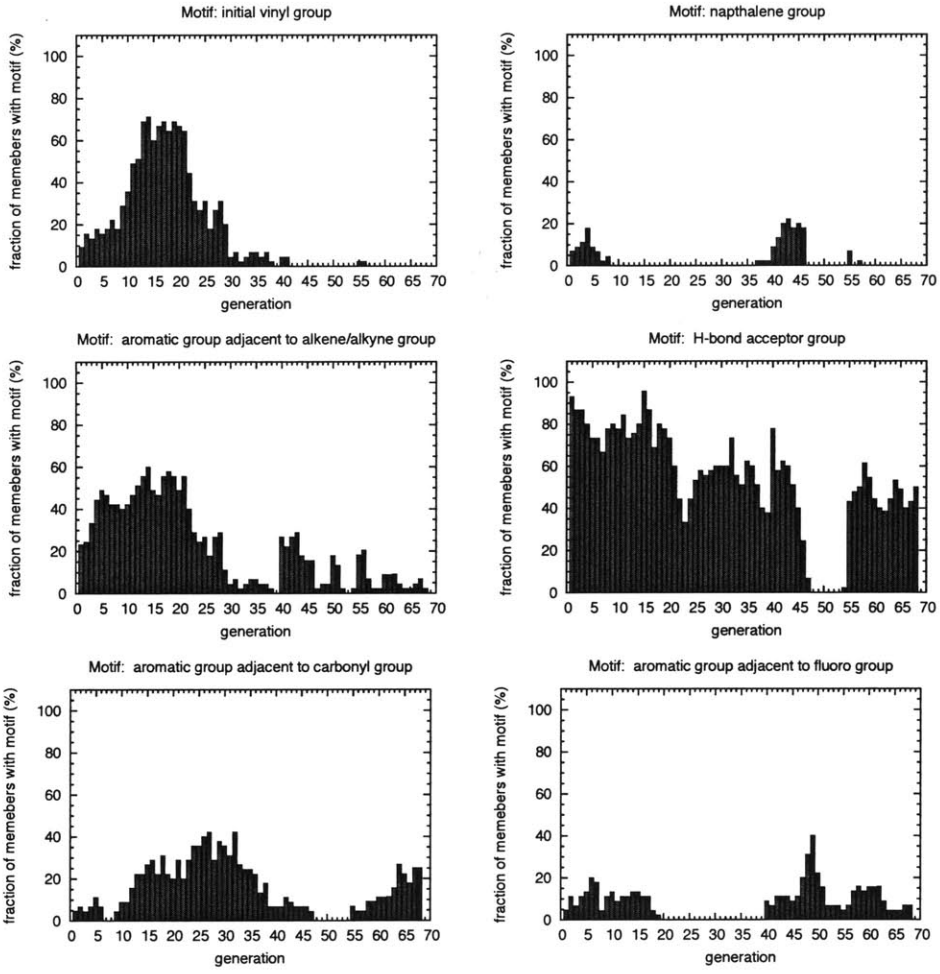


Figure B-34. Prevalence of motifs in generations 1 to 68 of experiment IV. Motif descriptions are listed in the title of each graph. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

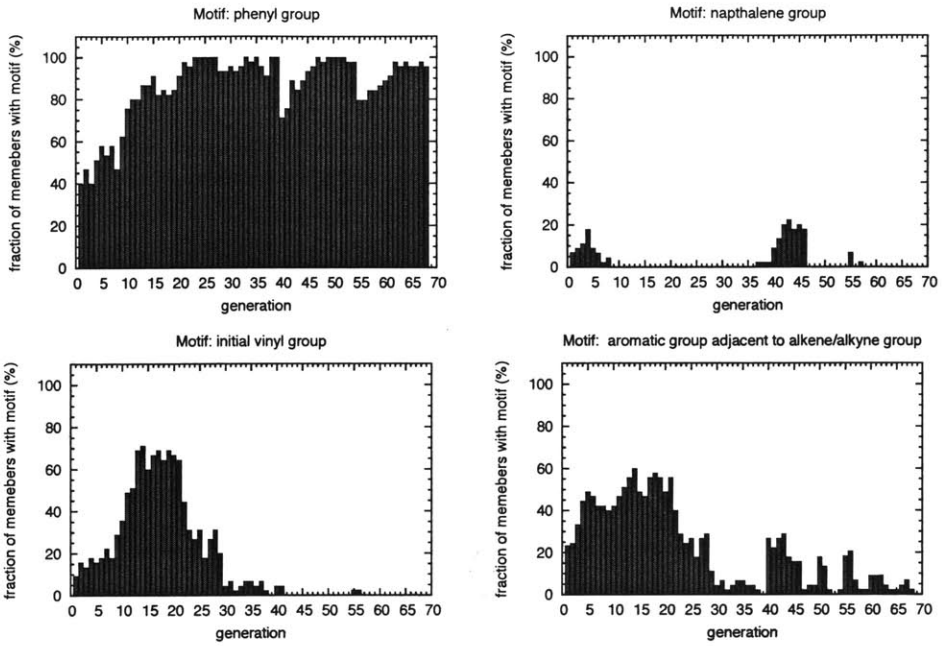


Figure B-35. Prevalence of motifs involving unsaturated/aromatic groups in generations 1 to 68 in Experiment IV. Motif descriptions are listed in the title of each graph. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

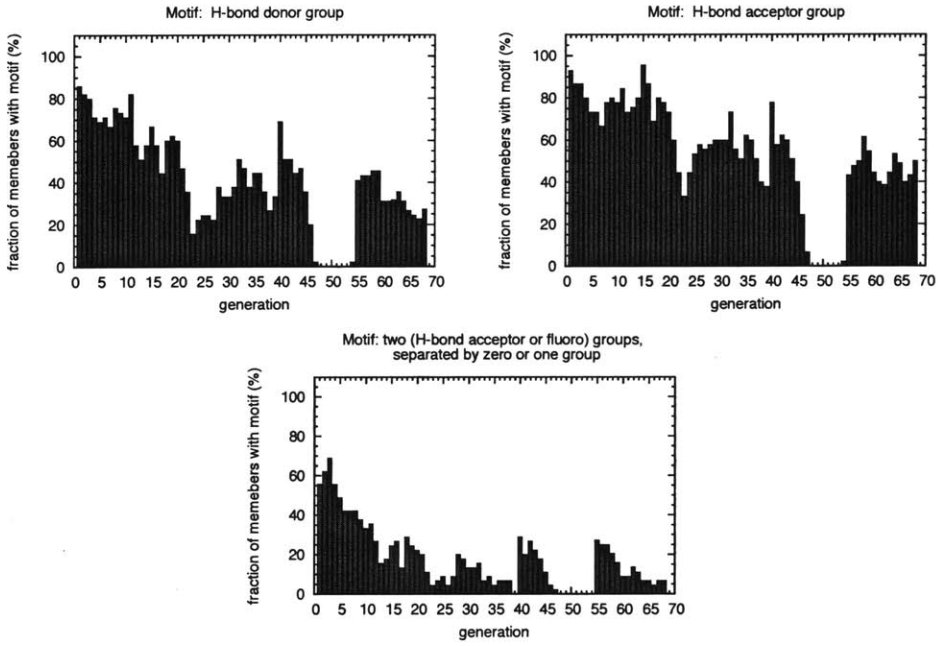


Figure B-36. Prevalence of motifs involving hydrogen-bond donors and acceptors in generations 1 to 68 in Experiment IV. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

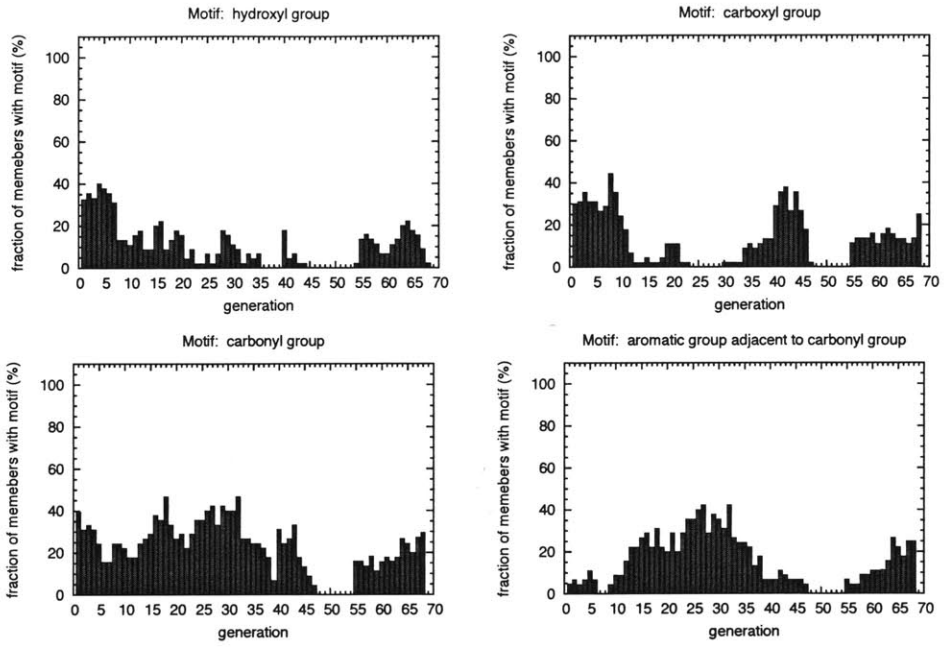


Figure B-37. Prevalence of motifs involving oxygen- or sulfur-containing groups in generations 1 to 68 in Experiment IV. Note that the "carbonyl groups" do not include the carbonyl carbon within carboyxl groups. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

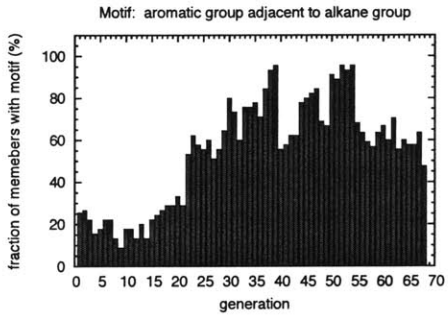


Figure B-38. Prevalence of other motifs in generations 1 to 68 in Experiment IV. Note that ether groups can consist of either oxygen-based ether groups or thioether groups. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

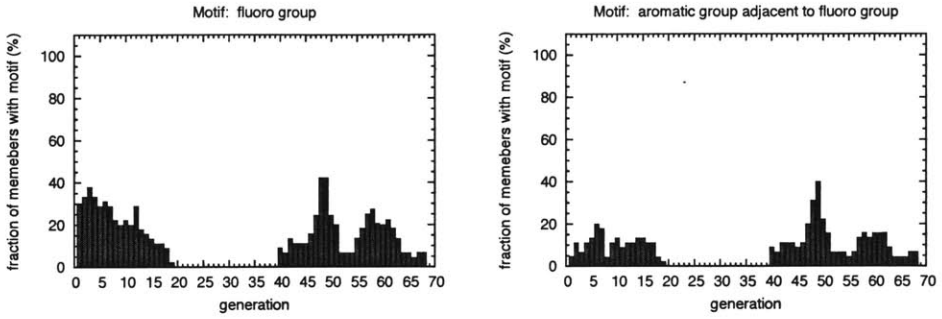


Figure B-39. Prevalence of motifs involving halide and amino groups in generations 1 to 68 in Experiment IV. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

Table C-10. Top-scoring ligand designs from generations 1 through 68 in Experiment IV, *for which at least 18 ns of production MD* (corresponding to four ligand evaluations) was performed. Listed scores are averages of all evaluations for each design, and inlcude the length penalty for each ligand. All values are in kcal/mol.

AVG SCOR	PEN	$\Delta E_{ads,\mathrm{E2}}$	$\Delta E_{ads,\mathrm{E6}}$	prod. MD (ns)	SEQUENCE
3.67 ± 0.04	0.0	-9.5	-7.0	72.0	COOH−(m)Ph−(m)Ph−Ph
2.00 ± 0.10	-0.4	-11.3	-8.4	18.0	CH_3−(m)Ph−CH(COOH)−(m)Ph−(m)Ph−(m)Ph−O−CH(CH_2CH_3)−COOH
1.42 ± 0.09	0.0	-8.6	-7.1	18.0	CH_3−(m)Ph−CF_2−O−(m)Ph−Ph
1.25 ± 0.16	0.0	-5.0	-4.4	202.5	CH_3−(m)Ph−SH
1.19 ± 0.25	0.0	-8.6	-8.4	18.0	CH_2=CH−(m)Ph−(m)Ph−(m)Ph−(m)Ph−Ph
1.12 ± 0.06	0.0	-7.0	-5.7	31.5	CH_3−(m)Ph−CH(CH_2CH_3)−Ph
1.03 ± 0.10	0.0	-9.3	-8.3	18.0	CH_3−CO−(m)Ph−(m)Ph−(m)Ph−CH_2−Ph
1.01 ± 0.06	0.0	-6.6	-5.1	31.5	F−CHOH−CO−CH(iBut)−COOH
0.93 ± 0.08	0.0	-8.3	-7.4	18.0	COOH−(m)Ph−COO−CH(CH_3)$_2$
0.92 ± 0.06	0.0	-6.6	-5.8	31.5	COOH−(m)Ph−Ph
0.86 ± 0.20	0.0	-6.9	-6.3	49.5	CH_2=CH−(m)Ph−CO−CH(CH_3)$_2$
0.86 ± 0.05	0.0	-7.9	-7.0	63.0	CH_3−(m)Ph−(m)Ph−(m)Ph−SH
0.84 ± 0.13	0.0	-8.3	-7.7	22.5	COOH−(m)Ph−(m)Ph−(m)Ph−Ph
0.84 ± 0.12	0.0	-7.1	-6.5	27.0	CH_2=CH−(m)Ph−CH_3
0.84 ± 0.09	0.0	-7.6	-6.9	18.0	CH_2=CH−(m)Ph−CO−(m)Ph−Ph

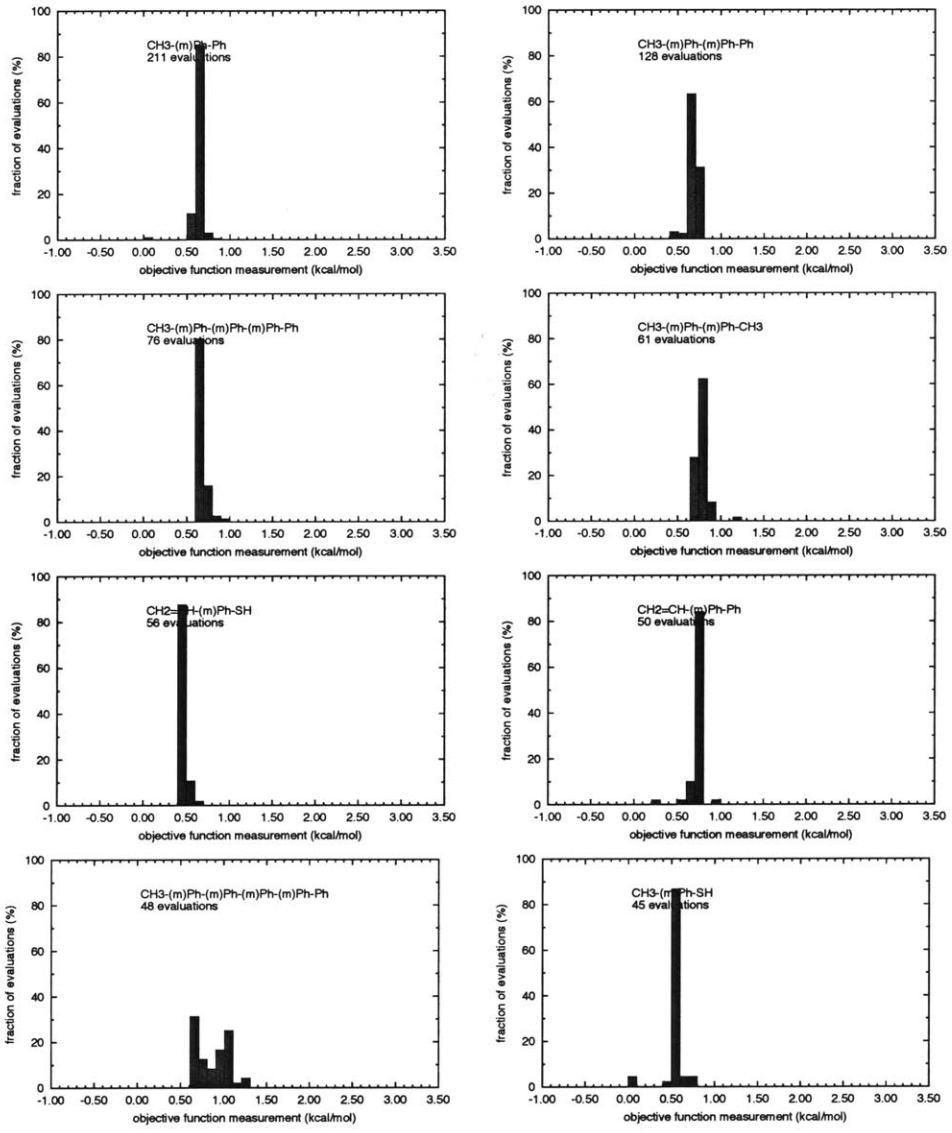


Figure B-40. Histograms of measured objective function values (fitness score plus length penalty) for frequently-occurring ligand designs in Experiment IV, which used 4.5 ns production MD and lookback scoring for evaluation.

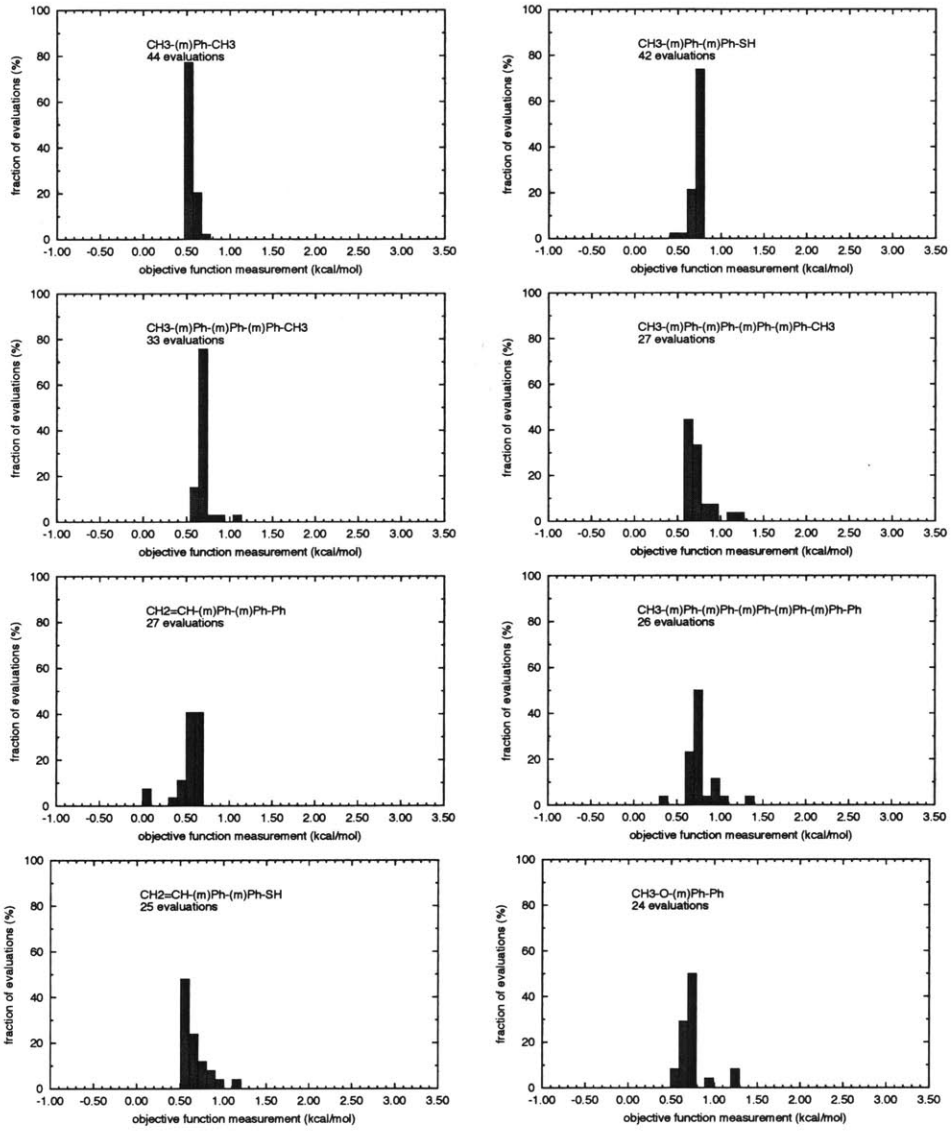
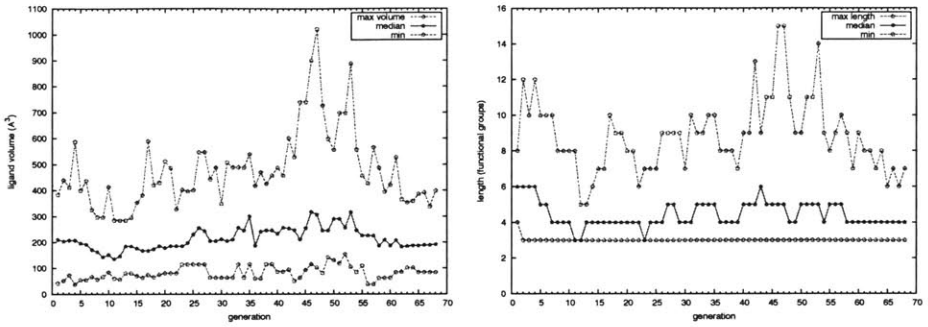
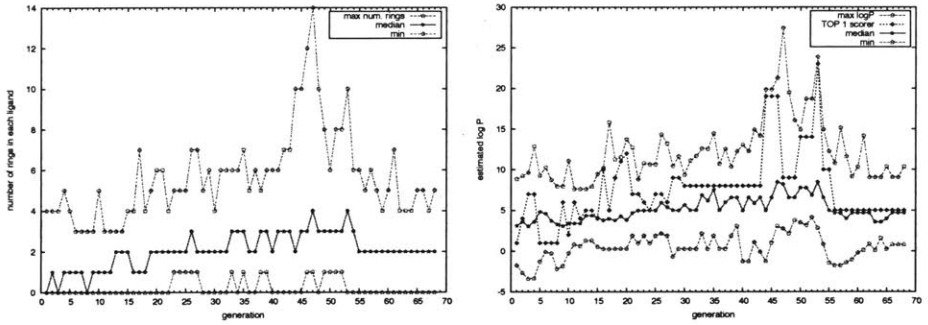


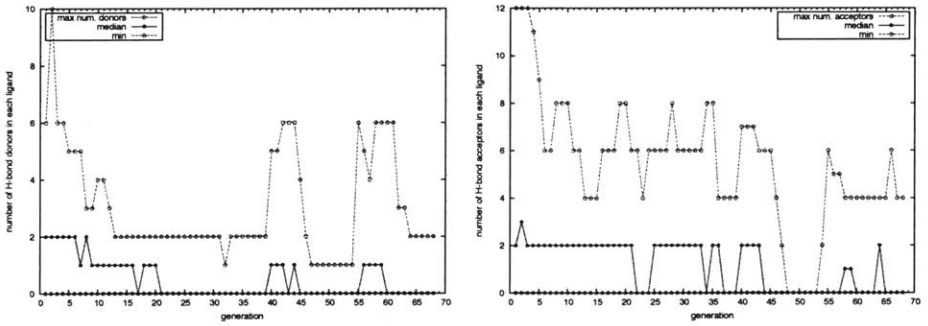
Figure B-41. Histograms of measured objective function values (fitness score plus length penalty) for frequently-occurring ligand designs in Experiment IV, which used 4.5 ns production MD and lookback scoring for evaluation.



(a) Molecular volume and length in functional groups of each ligand.



(b) Number of rings and log of partition coefficient $\log P$ of each ligand.



(c) Number of hydrogen bond donors and acceptors in each ligand.

Figure B-42. Property distribution evolution in generations 1 through 68 in experiment IV. In each figure, the filled points with solid lines represents the median value of that property, while the open points connected with dotted lines are the maximum and minimum value in that generation. The population of ligand designs in Experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

B.4 Formulation of surrogate objective function

A surrogate objective function is an objective function that is much simpler to calculate than the true objective function, and is often used to carry out accelerated evolution at the beginning of a GA process, or to better understand the evolution process effected by a GA.

In our case, the purposes of describing a surrogate objective function and formulating a "toy problem" are:

1. to understand how population size affects evolution, and how long convergence takes

2. to understand influence of parameters on the evolution process

3. to understand effects of objective function noise on evolution

4. to determine whether evolution process can select for ligand spacing (future)

B.4.1 Surrogate objective function definition

To be clear, the surrogate objective function here is designed to quickly estimate the selectivity of a ligand layer towards E2 adsorption, as against E6 adsorption. Given a particular ligand's sequence (corresponding to its molecular structure), quantitative structure–property relationships (QSPRs) would allow us to infer molecular properties of the ligand; the surrogate objective function can then attempt to link selectivity to these inferred physical properties, as detailed below. The properties available for inference through QSPR are listed in Table 5-4 in Section 5.5.

To estimate the adsorption selectivity of a ligand layer (for E2 as against E6), we attempt to "build into" the surrogate objective function several physico-chemical phenomena: hydrogen bond acceptance and donation between the ligand layer and E2 or E6; hydrophobicity-like effects that make E2 or E6 adsorption from ethyl acetate more or less favorable; and epitaxial effects that may occur if E2 or E6 adsorbs onto a ligand layer in an ordered fashion.

No experimental information is available about the lattice parameters or crystal form of either E2 or E6, so the last effect cannot be considered.

The first two effects described above—hydrogen bonding and solvent effects—can be considered, albeit in an admittedly *ad hoc* way. For example, we could consider the number of hydrogen bonds that ligand molecules (either as donors or acceptors) could make with each E2 molecule. Each E2 molecule can accept two hydrogen bonds and can donate two hydrogen bonds, so adsorption of E2 will be more enegetically favorable if the ligand molecule has H-bond donors and acceptors in its outermost functional groups, and the benefit should be proportional to the energetic strength of hydrogen bonds. The same reasoning would apply for the adsorption of E6 molecules.

To quantitatively apply the reasoning above, the tendency of the E2 molecule to adsorb F_{E2} on a ligand of sequence $\mathbf{q}$ includes a term which multiplies the number of hydrogen bond donors in each functional group by 2 (*i.e.* the number of H-bond acceptors in the E2 molecule) and by a weighting factor based on its position relative to the surface of the ligand layer. This weighting factor $w(i)$ is an exponentially decaying function depicted in Figure B-43. Likewise, the F_{E2} function, measuring tendency of E2 to adsorb, also includes a term multiplies the number of hydrogen bond acceptors in each functional group by 2 (*i.e.* the number of H-bond donors in the E2 molecule) and by the weighting factor.

Similar terms are present in the function F_{E6} estimating the tendency of E6 to adsorb on the ligand layer, as shown in the equation below. In all cases the terms include a hydrogen bond energy $E_{hb} = 2.0$ kcal/mol.

The overall surrogate objective function is then the difference between these two estimates of adsorption favorability, plus a length penalty to discourage overly-long

ligands (discussed below).

$$F_{\mathrm{E6}}(\mathbf{q}) = \sum_{\text{genes } i} [2N_{hb,don}w(i)E_{hb} + N_{hb,acc}w(i)E_{hb}]$$
$$+ A_{hp}g_{\mathrm{E2}}(\log P_{lig})$$
$$F_{\mathrm{E2}}(\mathbf{q}) = \sum_{\text{genes } i} [2N_{hb,don}w(i)E_{hb} + 2N_{hb,acc}w(i)E_{hb}]$$
$$+ A_{hp}g_{\mathrm{E6}}(\log P_{lig})$$
$$F_{obj}(\mathbf{q}) = F_{\mathrm{E2}}(\mathbf{q}) - F_{\mathrm{E6}}(\mathbf{q}) + p_{len}(\mathbf{q})$$

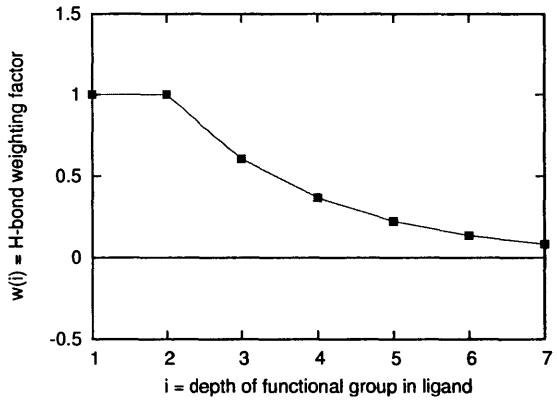


Figure B-43. H-bond donor or acceptor weighting function. The "depth" of outermost functional group is 1.

The other physico-chemical effect mentioned above is a "solvent-like" effect: during reverse-phase chromatography, solutes do not behave like their thermodynamic idealizations, present on a two-dimensional Gibbs dividing surface. Instead, the adsorbed molecules can be present within the ligand layer, or at the interface between the ligand layer and the solvent molecules. In the former case, the ligand layer provides the chemical environment surrounding the E2 or E6 molecule, and the tendency of this environment to promote adsorption depends on the thermodynamic stability of the E2 or E6 molecule in this environment, compared to its state in solvent (in this case, ethyl acetate).

To account for these effects, we have put forward a simple stepwise function to compare the hydrophobicity of the ligand to the hydrophobicity of the E2 and E6

molecules, based on the idea that "like dissolves like." More specifically, if the solvent environment provided by the ligand layer is more *like* to E2 (or E6) than the solvent environment provided by ethyl acetate, we ascribe to the ligand a greater tendency for adsorption of E2 (or E6). This stepwise function examines the logarithm of the water-octanol partition coefficient of the ligand, $\log P_{lig}$ which is a measure of hydrophobicity estimated using a QSPR[152] (see section 5.5.1 above).

The function accounting for these effects is denoted $g_{\mathrm{E2}}(\log P_{lig})$ or $g_{\mathrm{E6}}(\log P_{lig})$, and pictured in Figure B-44. The step-changes in value occur at the points $\log P_{EtOAc} = 1.32$; $\log P_{E2} = 2.73$; and $\log P_{E6} = 2.48$. The coefficient in the equation above is $A_{hp} = 0.6$ kcal/mol.

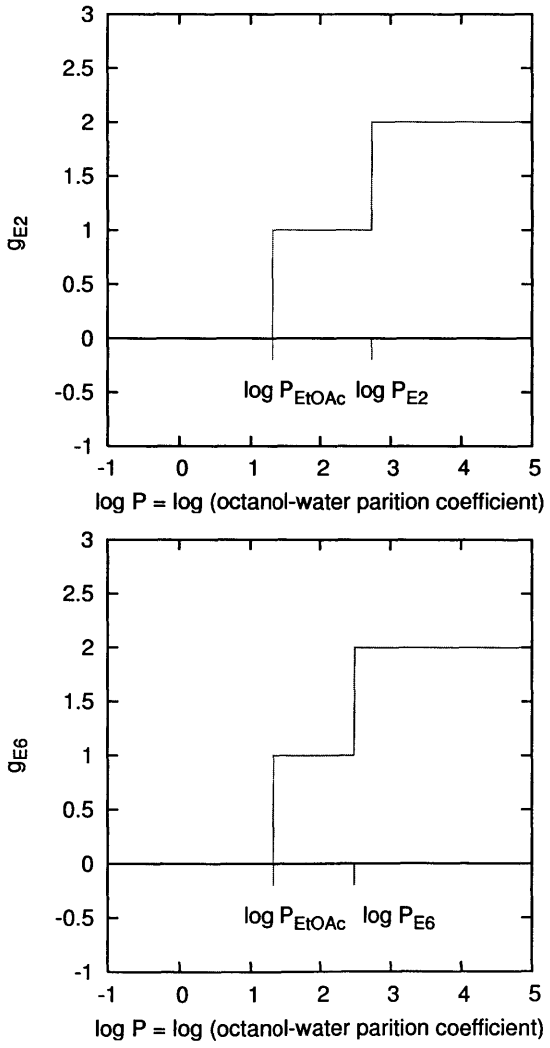


Figure B-44. Hydrophobicity-matching similarity functions for E2 and E6.

Finally, a length penalty is imposed within the overall surrogate objective function to discourage long ligands. Rather than an absolute cap on ligand length (which is always measured in functional groups), a "soft" penalty function is imposed at a length of 13 functional groups, as shown in Figure B-45.

Finally, to simulate the effect on he evolution process of random, statistical error in objective function, we tested a "noisy" surrogate objective function. In the noisy cases, random noise $\epsilon = A\, n(0,1)$, with $A = 2.0$ or 4.0 kcal/mol, is added to objective function F_{obj}. The designation $n(0,1)$ indicates a Gaussian random variable with mean zero and standard deviation 1.

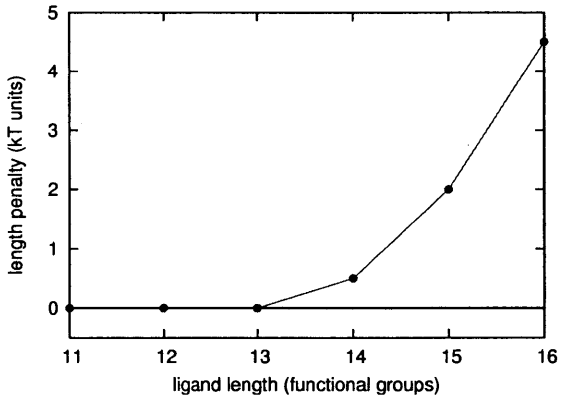


Figure B-45. Ligand length penalty function $p_{len}(\mathbf{q})$.

B.4.2 Results and conclusions from evolution with surrogate objective function

In the base case, the number of H-bond acceptors grew, as expected. The emergence of such groups was strongest in positions near the end of the ligand, since the weight-averaged number of acceptors grew to its max value.

$$\langle N_{acc} \rangle_w = \frac{1}{\sum_i w(i)} \sum_{\text{genes } i} N_{hb,acc} w(i)$$

In this case, it appears, yes: many $\log P$ values are in the advantageous range This acheived by including hydroxyl (1012), carboxyl (1019), and thiol (13) groups.

Table D-11. Summary of evolution trials using surrogate objective function. Numbers following a scaling or selection technique descibe the value of the key parameter used: "tournament 2" indicates a two-member tournament; "top 9" indicates top scaling with 9 members; and "rank 1.6" or "linear 1.6" indicates rank or linear scaling with a selection pressure of 1.6.

test	noise level[a]	pop	soft len. lim.	scaling	selection	$p_{crossover}$
1		60	9	window	roulette	0.40
2		60	9	none	tournament 2	0.40
3		60	9	top 9	roulette	0.40
4		60	9	rank 1.6	roulette	0.40
5		60	9	linear 1.6	roulette	0.40
6		60	9	none	tournament 2	0.40
7		60	9	none	tournament 2	0.40
8	2.0	60	9	none	tournament 2	0.40
9	4.0	60	9	none	tournament 2	0.40
10	8.0	60	9	none	tournament 2	0.40
11	12.0	60	9	none	tournament 2	0.40
12		60	9	none	tournament 2[b]	0.40

[a] In noisy cases, random noise $\epsilon = A\ n(0,1)$ was added to the surrogate objective function, with values of A (in kcal/mol) listed here.
[b] No elitism used.

Length limit proved effective compromise between shorter ligands, flexibility of evolution process. About 1/5 of ligands were over length limit of 13 units. Evolution can select for properties, like hydrophobicity, that are secondary in magnitude to the major interactions (H-bonds).

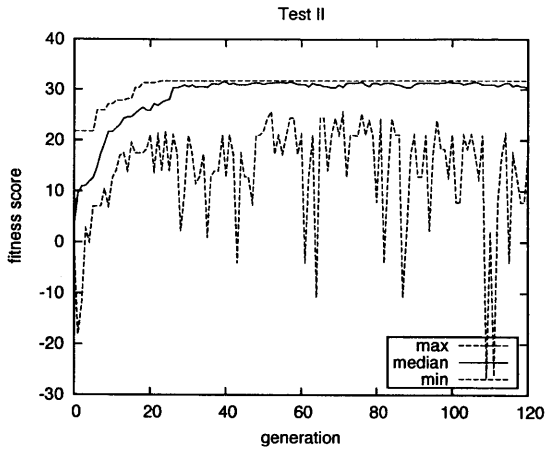
1. The genetic algorithm software and procedures used here does indeed change a population of ligands in conformance with the objective function used.

2. As expected, a larger population leads to faster convergence as measured by number of evolution steps; but in this case, some of the larger populations took more function evaluations and longer wall time than smaller ones. That is, there appears to be an optimum population size for use in the GA.

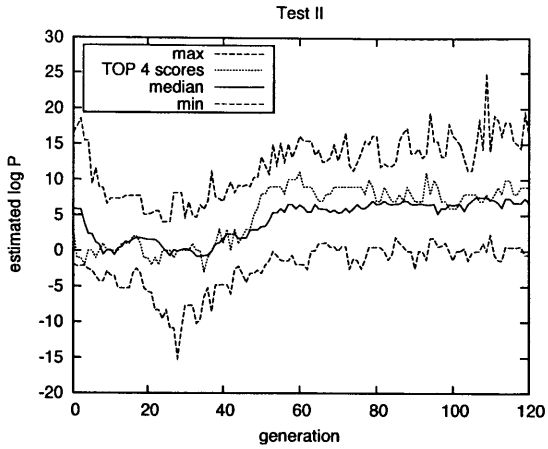
3. Evolution can lead to optimization of second-order effects, namely the hydrophobicity match between ligand and E2, as measured by $\log P$ estimates for the ligand and E2 molecules.

4. Noise in evaluation function seems to delay convergence, and also leads to suboptimal candidates in population. In cases with a "noisy" surrogate objective function, the top candidates do seem to reach same score levels as members evolved using the non-noisy evolution procedure. There was no dramatic difference in the application of the GA at the two noise levels used.

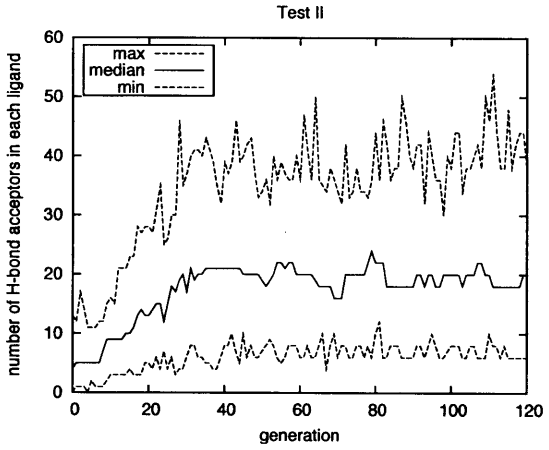
5. In this GA testing with a surrogate objective function, we did not examine the influence of evolution parameters on population diversity (as defined by similarity among sequences or among properties).



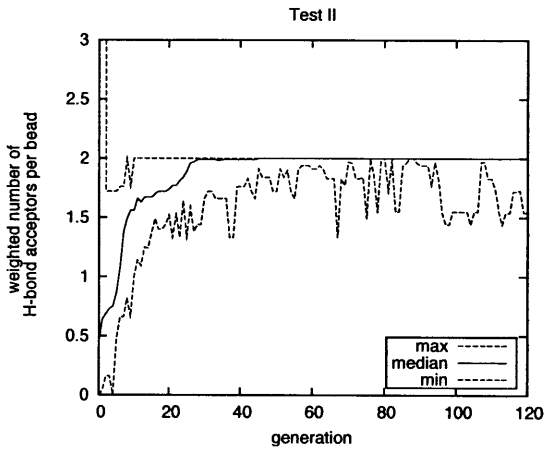
(a) Score evolution by generation.



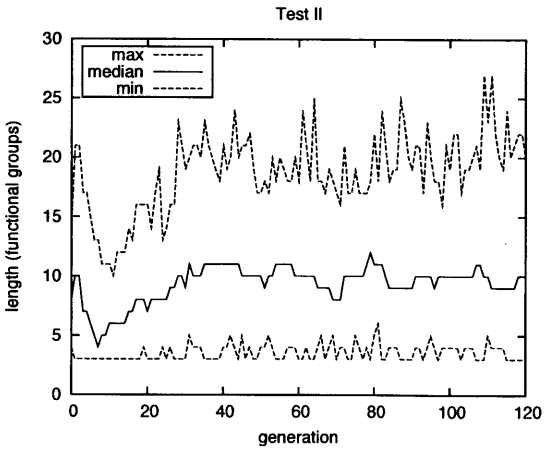
(b) log P values by generation. The "Top 4" series denotes the average of the four *highest-scoring* ligands.



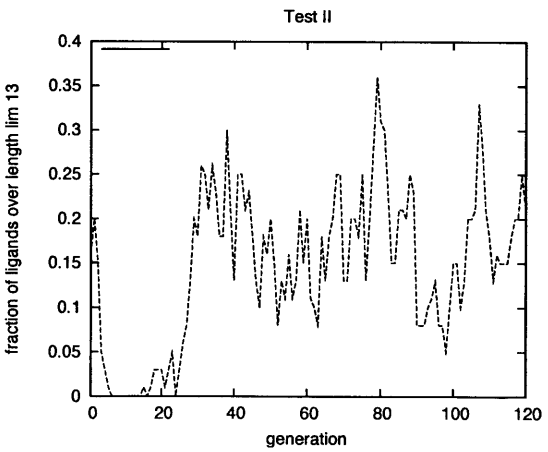
(c) Average number of hydrogen bond acceptors in each ligand.



(d) Weighted average of hydrogen bond acceptors per bead.



(e) Evolution of distribution of ligand length.



(f) Number of ligands longer than the "soft" limit of 13 functional groups.

Figure B-46. Overview of evolution process using a surrogate objective function in "base case": evolution carried out with population 60, length limit 13, and window scaling approach. In all cases, the "min," "med," and "max" designations refer to the minimum, median, and maximum values of *the plotted quantity* within the population of ligands.