Journal of
CHEMICAL INFORMATION
AND MODELING

Article

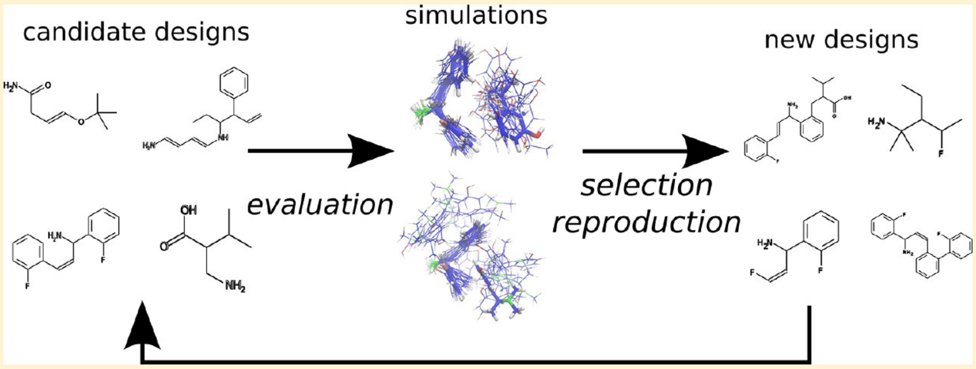
pubs.acs.org/jcim

Design of Linear Ligands for Selective Separation Using a Genetic Algorithm Applied to Molecular Architecture

Erik E. Santiso,[†] Nicholas Musolino, and Bernhardt L. Trout*

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02144, United States

Ⓢ *Supporting Information*



ABSTRACT: Continuous purification of chemical reaction products through adsorption-based operations during workup may present advantages over batch chromatography or crystallization. In pharmaceutical syntheses, however, the desired product is often structurally similar to byproducts or unconverted reactant, so that identifying a suitable adsorption medium is challenging. 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 such ligands through selection and reproduction steps. This procedure identified chemical designs of the surface-bound ligands that were outside the set we 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 0.8−1.6 kcal/mol in single-ligand, solvent-free simulations, if entropic contributions to the relative selectivity are neglected. We believe this molecular evolution technique presents a useful method for the directed exploration of chemical space or for molecular design, when the chemical properties of interest can be efficiently evaluated through simulations.

1. INTRODUCTION AND OBJECTIVES

1.1. Background. With increasing computer power and availability, it is now possible to carry out extensive molecular-level computer simulations to aid in molecular design. Computer simulations allow both rapid evaluation (compared with laboratory experiments) of thermodynamic and transport properties and provide an understanding of physical processes at a molecular level of detail which is often inaccessible to experiments. These capabilities lead to two paradigms for molecular design.

The first approach to molecular engineering is to carry out computer simulations to gain a detailed, mechanistic understanding of the physical phenomenon of interest, 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,[1] 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[2]—presents a vast domain. The space of all small, organic molecules has been estimated to contain up to 10^{60} members,[3] 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

Received: January 20, 2013
Published: July 11, 2013

Table 1. Previous Approaches to Molecular Evolution

program or author	purpose of molecular design	chromosome encoding	fitness function
Weininger[16]	fit to pharmacophore or resemble a given molecule	contemplated: 3D structure, connectivity graph, or SMILES string	contemplated: Tanimoto similarity[17] to a given molecule; presence of features such as rings, cationic sites; steric fit into 3D binding site; or experimental measurement
Chemical Genesis[18]	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[19]	fit to pharmacophore or mimic a given molecule	3D structure	presence of features at specific locations
Nachbar[20]	mimic structure of a given molecule	hierarchical text expressions specifying topology	atom-pair similarity or Dice similarity[21] to a given molecule
TOPAS[22]	mimic structure of a given molecule	graph of enumerated functional groups	Tanimoto similarity to target molecule, or 2D topological similarity to pharmacophore
ADAPT[23]	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,[24] plus penalty functions for violating QSAR constraints
LEA3D[25]	design ligand to bind to a particular site on a protein target	linear string of enumerated functional groups	FlexX 1.13.1[26] docking score
Molecular Evoluator[27]	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
Mandal et al.[28]	prioritize compounds for subsequent stages of drug discovery/screening process	positions in a scaffold and groups at those positions	weighted geometric mean of "desirability functions" computed from QSPRs
Dey and Caflisch[29]	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 $\Delta\Delta F_{\text{ads}}$ 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.

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"[4−6] or synthesizable.[7] But such databases also present a fixed subset of candidate solutions and, to this point, have been focused on potential pharmaceutical leads.

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[11] and other optimization approaches[12,13] have previously been used for molecular design; the use of GAs in the context of drug design was reviewed by Gillet[14] and more briefly by Terfloth and Gasteiger.[15] Examples of such work are summarized in Table 1.

Our approach differs from previous work, in that we employ as an objective function properties estimated 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 estimate an array of thermodynamic and transport properties. Other evaluation techniques could include properties calculated using density functional theory (DFT) or ab initio methods[30] or simpler quantitative structure−activity (or property) relationships (QSPR/QSARs).[31,32] Within the Harvard Clean Energy Project, for example, DFT calculations are used as a screening technique in the evaluation of novel organic photovoltaics.[33,34] 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[35−37] for molecular mechanics force fields and have been used with some success to evaluate the binding affinities of small molecules to other small molecules[38] or to proteins.[39−41] (In the cited examples, the GAFF force field[35,36] was applied to a small set of molecules identified a priori by researchers.)

In this particular paper, we seek to illustrate how this combination of GAs and MD simulations is useful in identifying potentially useful molecular structures and identify some of the issues that arise in combining the two approaches.

1.2. 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.[42] Its purpose is to selectively remove unconverted reactant from a solution also containing a reaction product, which should remain in the solution for further processing.

Such a material could be used to separate undesired solution components (while leaving the desired intermediate in solution) in a continuous fashion using a simulated moving bed (SMB) unit.[43−45] 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,[46−53] as has been reviewed elsewhere,[54−58] although nonenantiomeric separations are also possible.[43,59,60] Preliminary economic evaluations have shown that SMB-based separations[61] and continuous manufacturing more generally[62] have the ability to reduce overall process costs in pharmaceutical manufacturing.

The particular separation task in this study is the adsorption of 3-[1-(hydroxyl)ethyl]phenol (which we call "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. The surface we aim to design thus must simultaneously satisfy two design criteria: to adsorb E2 as strongly as possible, while adsorbing E6 as minimally as possible.

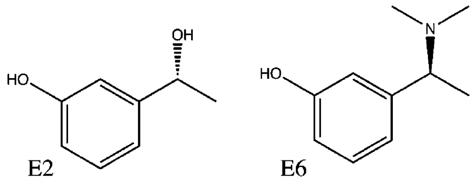


Figure 1. Structure of pharmaceutical intermediates designated E2 and E6.

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",[43] 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 2.2 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.[63] 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 water to the organic solvent after the separation.

Because of the multistep nature of pharmaceutical syntheses, and the frequent similarity of reactants and byproducts' chemical structures to those of desired products, we expect that identifying a suitable adsorption medium to effect continuous, SMB-based separations would often present similar challenges.

Overall, in this study we seek 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. The rest of this manuscript is organized as follows: In section 2, we describe the details of the molecular evolution algorithm used, as well as the fitness function evaluations. In section 3, we present and discuss the results of the ligand screening using different molecular evolution parameters. Section 4 contains some concluding remarks and possible directions for future work in this area.

2. METHODS

2.1. 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.[42] To optimize such a material, we have chosen to focus on the chemical architecture of the attached organic ligand. Since quasi-linear ligands are

Table 2. 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≡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≡C—
11	tert-butyl	$—C(CH_3)_3$	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	$—C_{10}H_6—$

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 2. 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 (Figure 2).

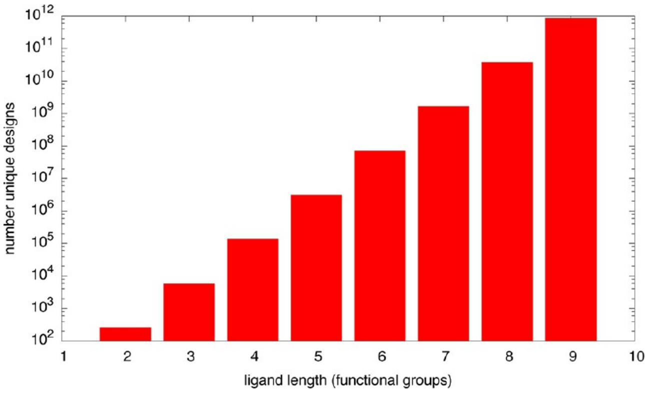


Figure 2. Number of ligand designs that can be created with functional groups used in this study, neglecting functional group combinations that are prohibited (see Table 3 in the Supporting Information).

In describing molecules in this way, it is useful to designate certain combinations of functional groups as *forbidden*.[8] 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 3 and also include combinations that would lead to incorrect atomtype designations under the GAFF force field. 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.

Table 3. 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 analogue
1001	13	−O−SH	thio-peroxide analogue
1003	1013	−COO−SH	thio-peroxide analogue
1001	1013	−O−S−	thio-peroxide analogue
1013	1001	−S−O−	thio-peroxide analogue
1001	0	−O−H	would create hydroxyl group with incorrect gaff atomtypes
1003	0	−COO−H	would create carboxyl group with incorrect GAFF atomtypes

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.

To evaluate potential ligand designs, we have employed molecular dynamics simulations. As noted in Table 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 mechanism(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,[35] 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 3.

2.2. Evaluation of Ligand Candidates through Molecular Dynamics Simulation.

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

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

But free energy differences are expensive to obtain computationally, so we have employed an energetic-only fitness score:

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

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.

$$\Delta E_{\text{ads,E2}} = \langle E_{\text{lig layer+E2}} \rangle - \langle E_{\text{lig layer}} \rangle - \langle E_{\text{E2}} \rangle$$

$$\Delta E_{\text{ads,E6}} = \langle E_{\text{lig layer+E6}} \rangle - \langle E_{\text{lig layer}} \rangle - \langle E_{\text{E6}} \rangle$$

where $\langle \cdots \rangle$ indicates ensemble averaging. $\langle E_{\text{E2}} \rangle$ and $\langle E_{\text{E6}} \rangle$ are the average energies of E2-only and E6-only simulations, which were performed one time. In separate quantum calculations,[64] the gas-phase dipole moments of the E2 and E6 molecules were obtained as 2.6 and 1.7 D (each averaged from two different configurations), respectively, indicating 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.

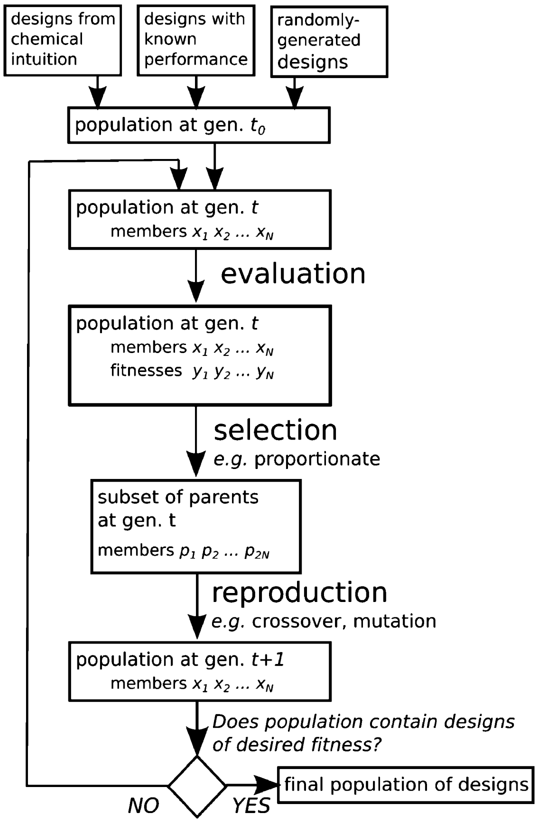

designs from chemical intuition
designs with known performance
randomly-generated designs
population at gen. t_0
population at gen. t members $x_1\ x_2 \dots x_N$
evaluation
population at gen. t members $x_1\ x_2 \dots x_N$ fitnesses $y_1\ y_2 \dots y_N$
selection
e.g. proportionate
subset of parents at gen. t members $p_1\ p_2 \dots p_{2N}$
reproduction
e.g. crossover, mutation
population at gen. $t+1$ members $x_1\ x_2 \dots x_N$
Does population contain designs of desired fitness?
NO
YES
final population of designs

Figure 3. Schematic of iterative evaluation/evolution process. The $\{x_i\}$ and $\{p_i\}$ represent sets of genomes, while $\{y_i\}$ are sets of fitness values.

This (arbitrary) length restraint is used to prevent ligands from increasing their fitness simply by virtue of accreting more atoms from other candidate ligands, thus increasing both of their interaction energies with E2 and E6 molecules.

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 2. 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.[65] 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.

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. More details can be found in section A.1 in the Supporting Information. 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.

After a reasonable 3D structure for the molecule is obtained from the FORM2GEOM program, a topology file is prepared using the ANTECHAMBER suite[36] and the GAFF force field.[35] Because atoms in the FORM2GEOM fragment library are already described by their GAFF atomtype, and bond types are likewise prespecified, 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 semiempirical technique[66,67] 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.

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 4.

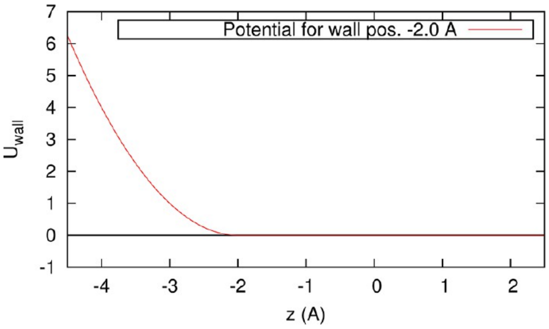


Figure 4. Illustration of half-well potential representing the solid surface to which ligands are attached. In this example, z_{wall} = −2.0 Å.

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[68] conjugate 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[68] 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.[69]

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 4.

Table 4. Summary of Genetic Algorithm and Evaluation Function Parameters in Four in silico Evolution Experiments

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

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.[70]

2.3. Molecular Evolution Procedure. As depicted in Figure 3, 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; (iii) and 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.[71,72] 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 4, 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 $P_{\text{sel}}(x_i) = f_i^s/\Sigma_j f_i^s$ 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.[71] The scaled fitness value f^s is 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:[71] it obviates the need to rescale 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 property estimations made from such simulations include statistical errors, due to limited sampling. To address the statistical error in the estimation of each ligand's adsorption selectivity, we developed a selection scheme that accounts for these uncertainties, 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 calculated 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 4. 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 $1/3\ (1 - p_{\text{crossover}})$, 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).

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,[17,73] 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 properties by their standard deviations in a reference population, as is done in the ChemGPS system,[74,75] the diversity metric was defined as the sum of pairwise differences

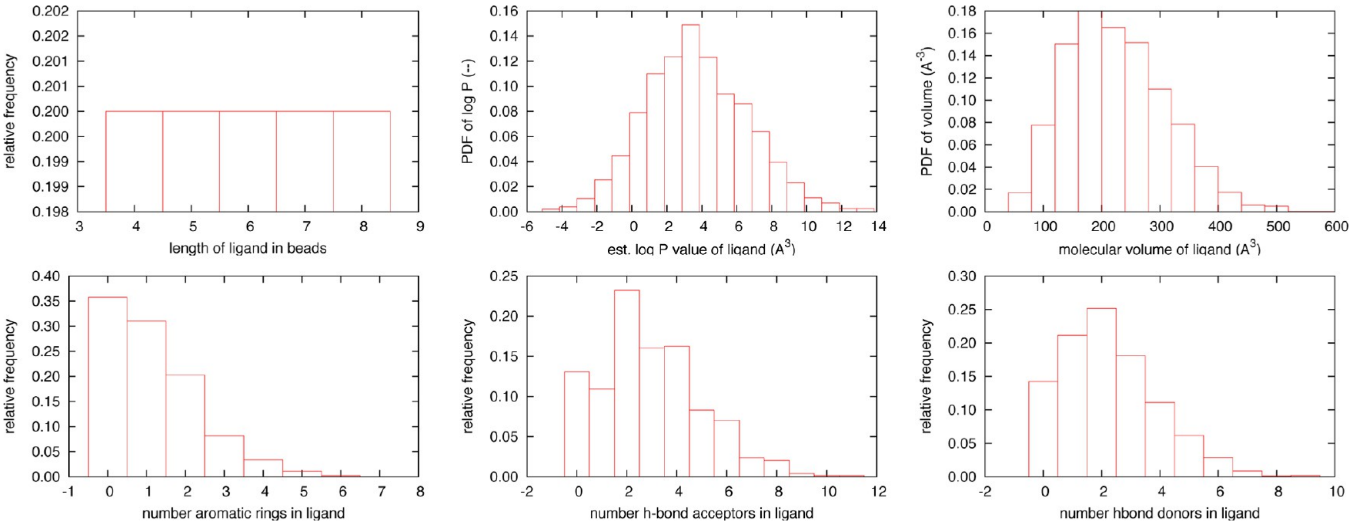


Figure 5. Properties of constituent ligand designs in a reference population of 2000 randomly generated ligands with uniform length distribution.

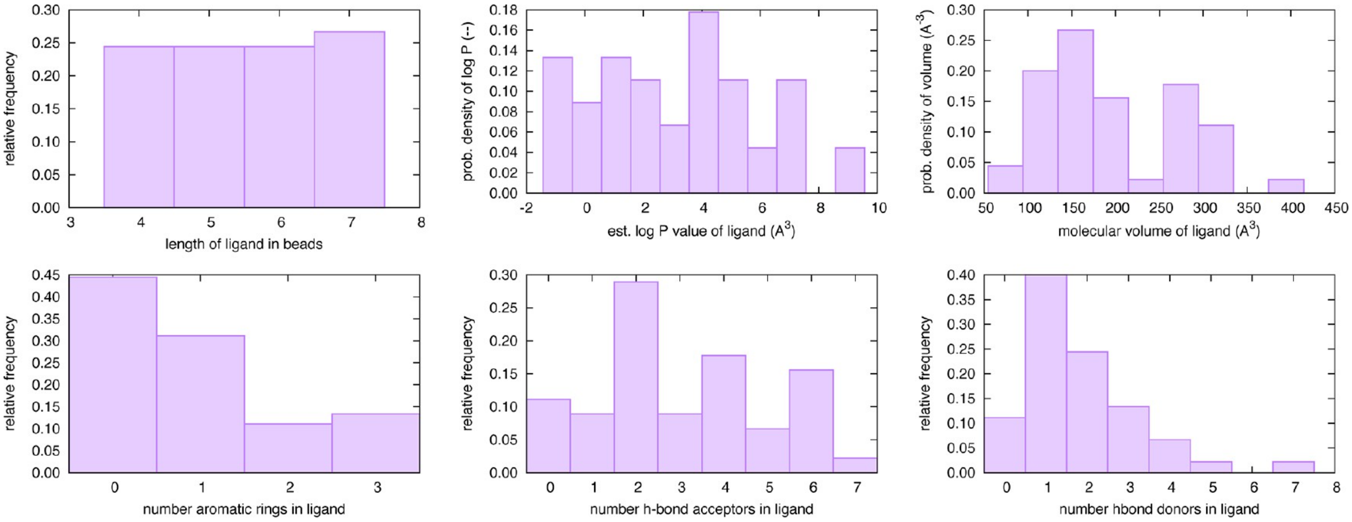


Figure 6. Properties of constituent ligand designs in generation 1 of experiment II.

in property space. Details can be found in section A.3 of the Supporting Information.

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 4.

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 (exps III and IV).

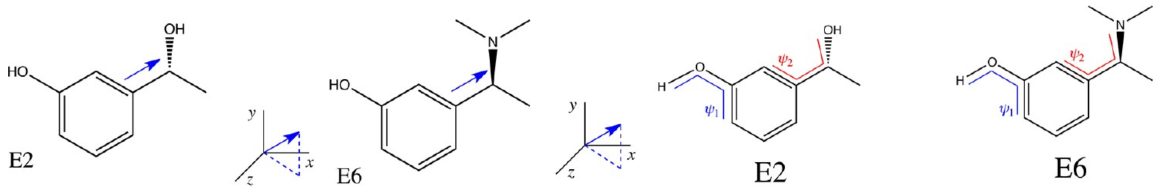
3. RESULTS AND DISCUSSION

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 section. Because experiment II began with a randomly generated set of ligand designs, it will be used to illustrate results obtained from evaluation (in section 3.1) and molecular evolution (in section 3.2).

3.1. 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 2000 ligands having uniform distribution of length from four to eight functional groups are shown in Figure 5; 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 (Figure 6), with near-uniform distribution of length between four and seven functional groups.

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.

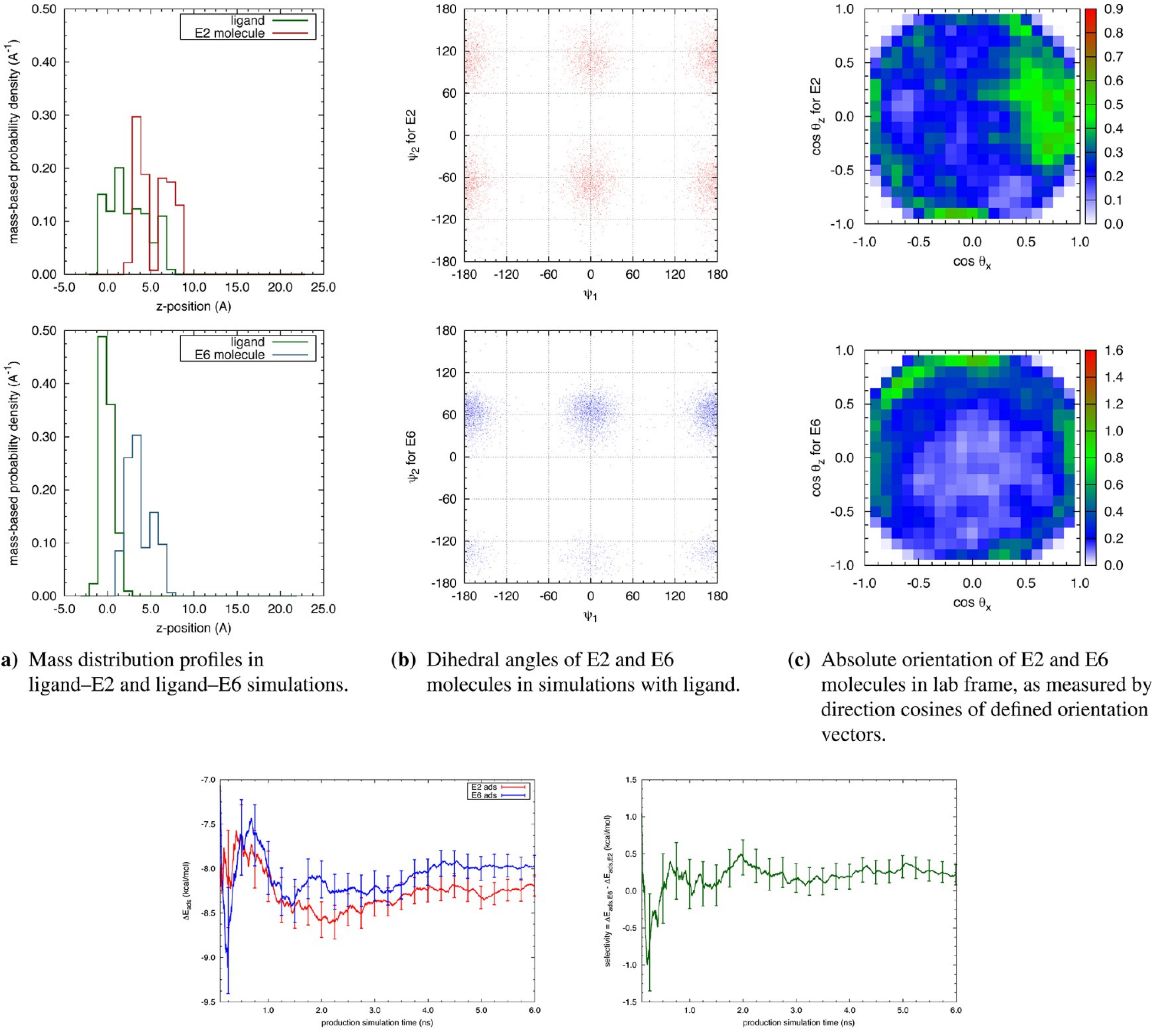
We sought to confirm that these simulations of adsorption/binding broadly sampled an energetically relevant set of ligand−target conformations. In these simulations, the bound atom in the ligand is anchored to the surface, restricting the



(a) Bond vector used to define absolute orientation.

(b) Dihedral angles.

Figure 7. Definitions of absolute orientation and internal degrees of freedom for E2 and E6.



(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 $\Delta E_{ads,\text{E2}}$ and $\Delta E_{ads,\text{E6}}$ over course of simulations.

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

Figure 8. Evaluation of ligand candidate 25 of generation 1 in experiment II, having sequence 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.

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 7) and its absolute orientation (measured by direction cosines of the vector in Figure 7a).

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 8. 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 8a 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 shown).

In examining the sampling that took place in evaluation simulations, Figure 8b 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 8c. Finally, the convergence of $\Delta\Delta E_{\text{ads}}$ and its two component adsorption energies are shown in Figure 8d and 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 10−15 ns, equivalent to two to three 4.5- or 6.0-ns evaluations. Figure 9 is one such example, and others can be found in the Supporting 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 10, and others can be found in Figures B-16 and B-17 of the Supporting Information.

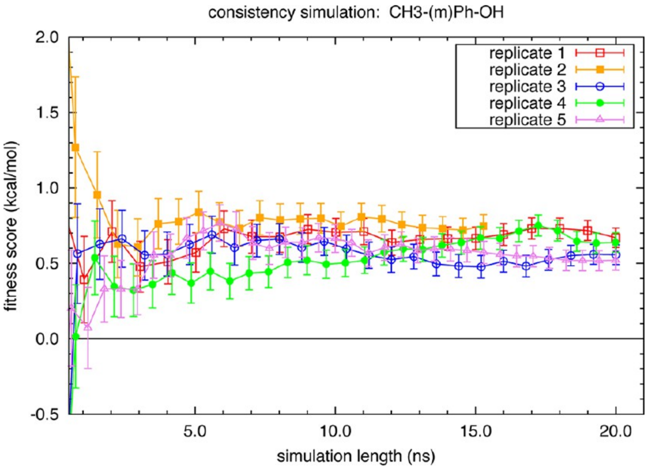


Figure 9. Convergence of fitness score of ligands with structure CH_3−(m)Ph−OH. Error bars are ±1 standard error. Similar results for six other ligand designs are shown in Figure A-3 in the Supporting Information.

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 11. The magnitude of the standard errors, calculated using the method of Allen and Tildesley,[70] were confirmed using bootstrap sampling.

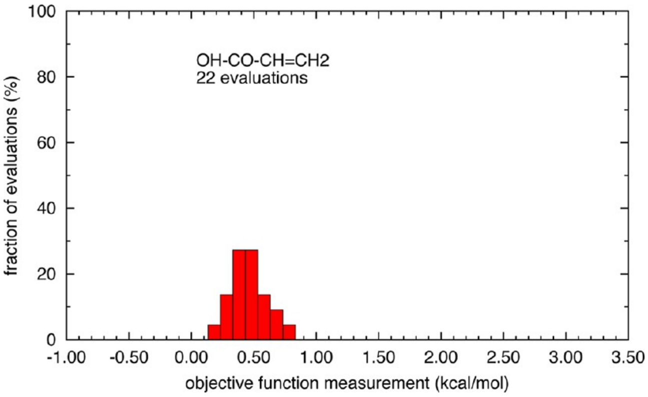


Figure 10. Histogram of fitness score evaluations for the indicated ligand design in experiment II. In that experiment, all fitness evaluations were independent.

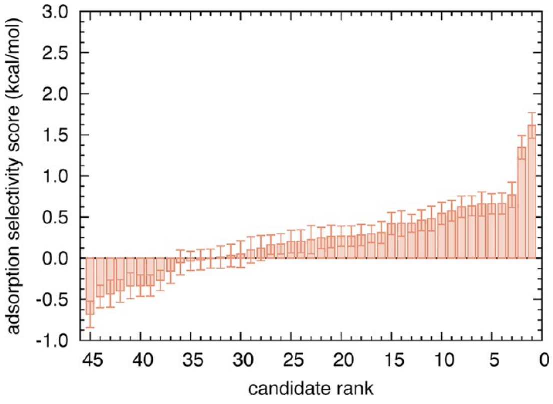


Figure 11. Fitness scores of members of initial population ($N = 45$) in experiment II, in rank order.

3.2. Molecular Evolution Outcomes.

The molecular evolution processes observed in experiments I−IV are summarized in Figures 12−15. 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 80th 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.[76]

As the GA was applied to the ligand population, the score distribution generally shifted toward higher scores, as suggested by Figures 12−15. 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 12−15) increases at a rate less than 45 per generation, especially toward 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

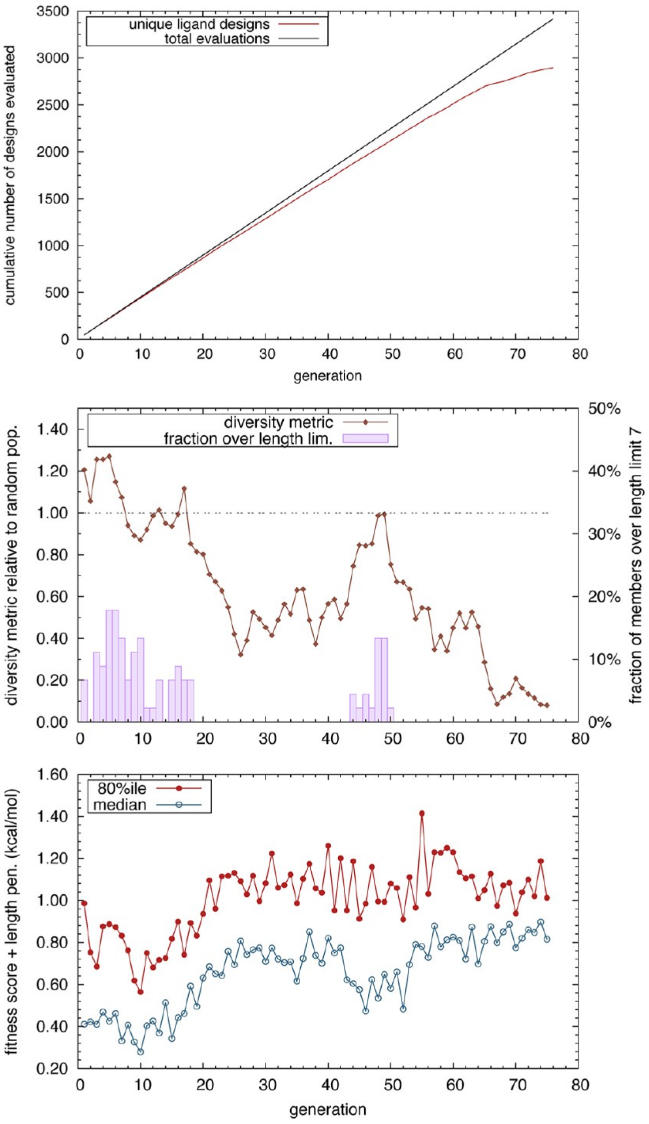


Figure 12. Characterization of evolution over generations 1−76 in experiment I, which featured N = 45 ligands, roulette wheel selection after window-based scaling, and 3.0-ns production MD in each evaluation.

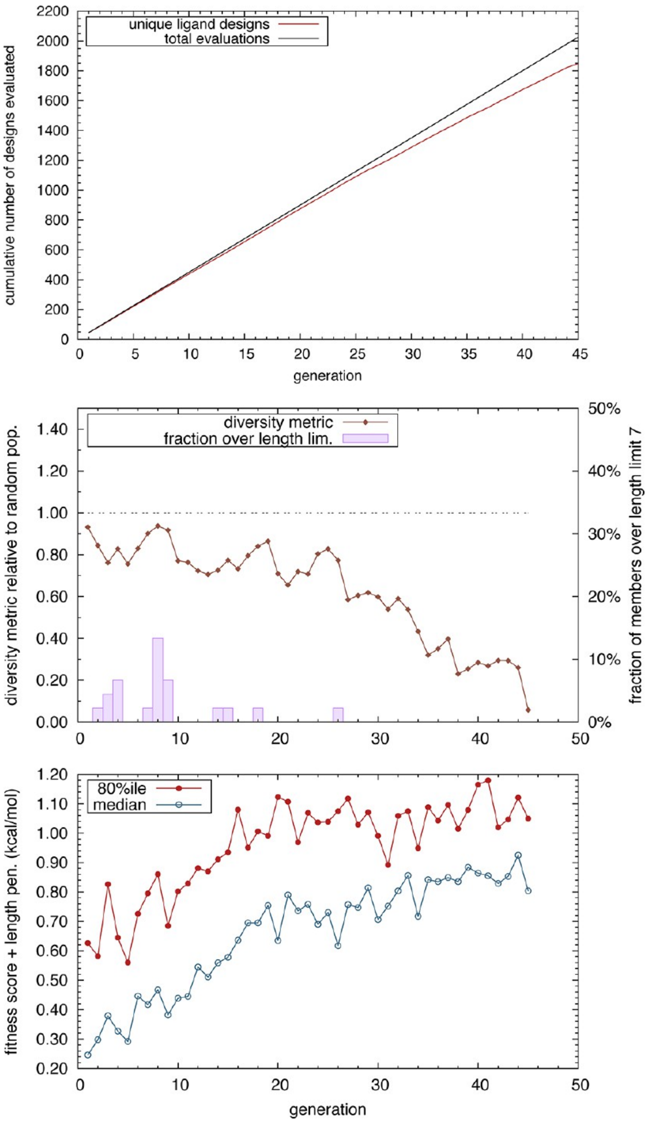


Figure 13. Characterization of evolution over generations 1−45 in experiment II, which featured N = 45 ligands, roulette wheel selection after window-based scaling, and 4.5-ns production MD in each evaluation.

design from one generation to 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 successive generation.

3.2.1. Ligand Design. The top-scoring ligand designs from each experiment are listed in Table 5. As noted in the caption, only ligand designs 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 5, showing results from exp I, 8 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 nonselection, after a one-time fluctuation 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-16, B-17, and B-27 in the Supporting Information and would tend to mitigate this problem.

In the ligand designs in Table 5, certain functional groups appear with greater-than-random frequency, namely phenyl, naphthalene, 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

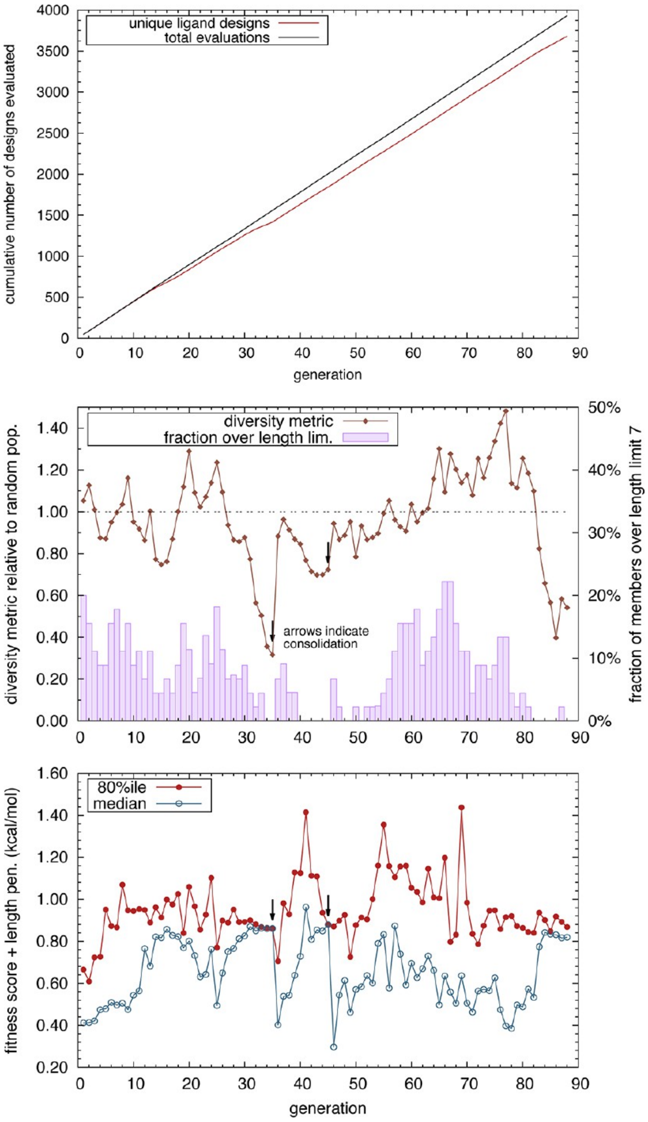


Figure 14. Characterization of evolution over generations 1−88 in experiment III, which featured $N = 45$ ligands, 2-member tournament selection, and 4.5-ns ligand evaluation. The population of ligand designs in experiment III was subjected to "consolidation" before generations 35 and 45 (see text).

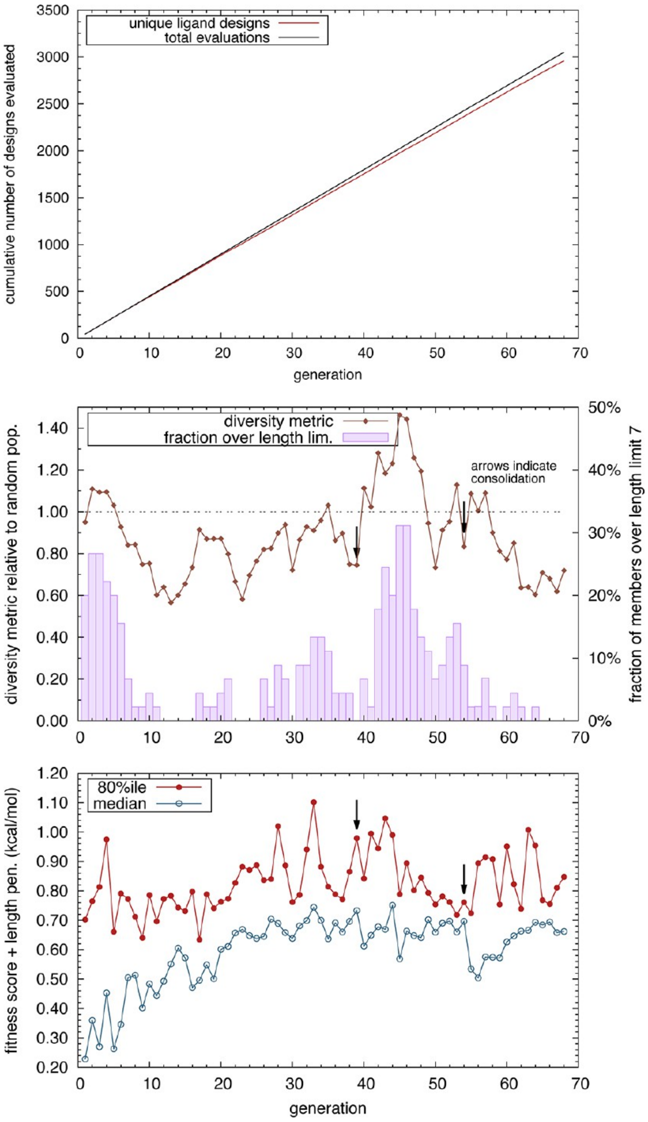


Figure 15. Characterization of evolution over generations 1−68 in experiment IV, which featured $N = 45$ ligands, 2-member fuzzy tournament selection, and 4.5-ns ligand evaluation. The population of ligand designs in experiment IV was subjected to "consolidation" before generations 40 and 55 (see text).

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 3.3 below.

A selection of relevant chemical motifs from each experiment is shown in Figure 16. 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. Figure 16a and b 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, Figure 16b and c show that a successful motif can be germinated during the evolution process and, then, successfully emerge to be present in a significant portion of the population's 45 members.

3.3. 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 developed 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 16, ligand designs that emerged from molecular evolution, including many of the highest-scoring designs in Table 5, contained aryl (phenyl or naphthalene) groups, alkenyl/alkynyl groups, or carbonyl groups, especially in close proximity to each other. 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

Table 5. Top-Scoring Ligand Designs in Each Experiment, for Which at Least Two Ligand Evaluations Were Performed (Except Experiment IV, Which Has a Threshold of Four Evaluations)[a]

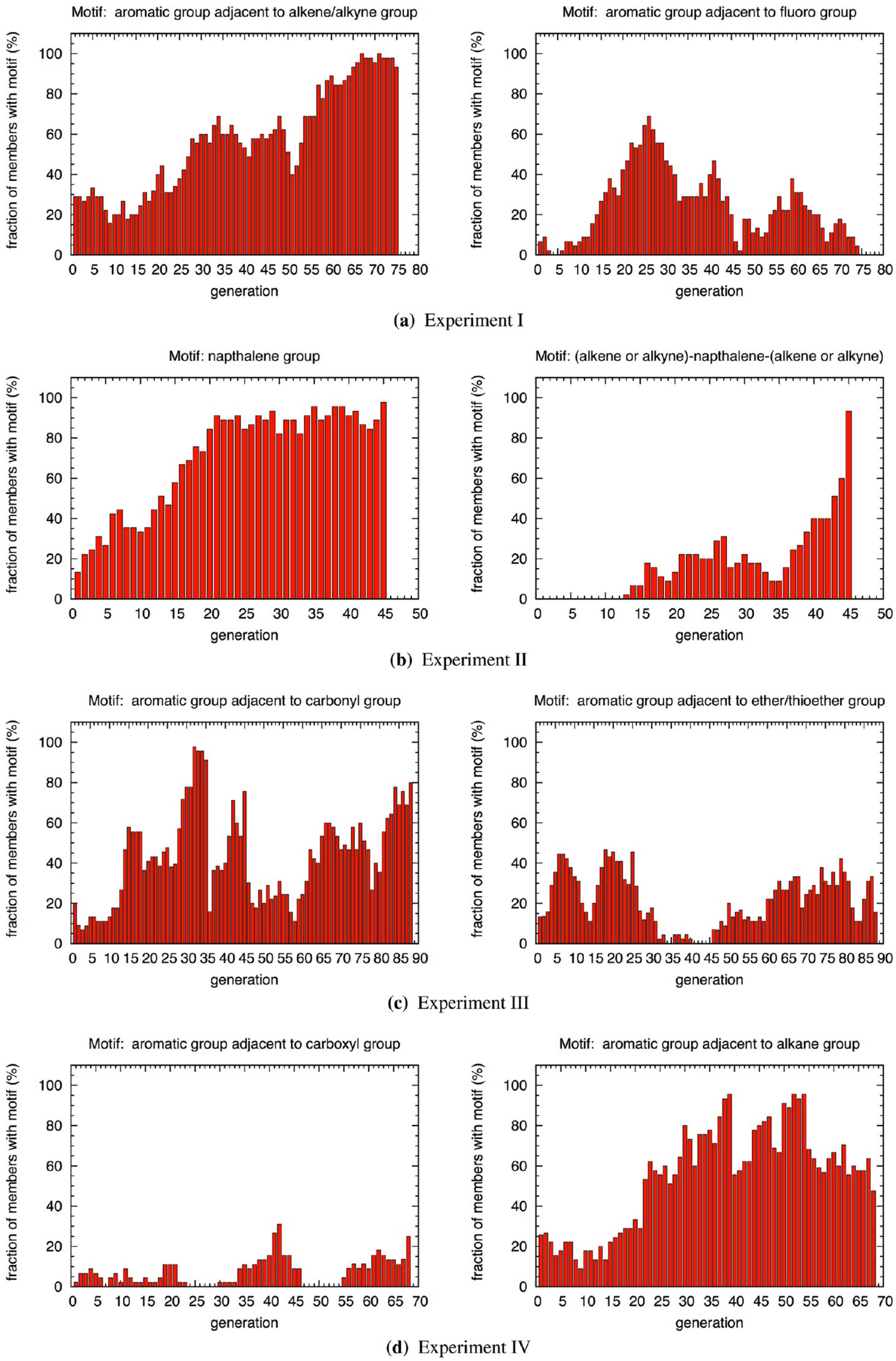
	avg score	pen	$\Delta E_{ads,E2}$	$\Delta E_{ads,E6}$	prod (ns)	sequence
exp I	1.60 ± 0.13	0.0	−7.2	−5.6	6.0	F—$C_{10}H_8$—CH=C=CH—(trans)CH=CH—H
	1.50 ± 0.14		−10.3	−8.8	9.0	Cl—C(Ph)H—$C_{10}H_8$—CH=C=CH—CH(COOH)—CH(NH_2)—C(CH_3)$_3$
	1.44 ± 0.15	−0.1	−7.4	−5.9	6.0	CH_2=C=CH—(trans)CH=CH—CH(NH_2)—(p)Ph—C≡C—(trans)CH=CH—CF_2—CH_3
	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

[a] Listed scores are averages of all evaluations for each design and include the length penalty each ligand. All score and ΔE_{ads} values are in kilocalories per mole.

interact favorably with E2's differentiating hydroxyl group. We also checked to see whether the presence of an sp^2 terminal group (e.g., CH_2=CH—) at the simulated wall kept the ligand's principle 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 17, in which each subfigure includes the minimum-potential-energy frames from all three simulations.

To quantify this difference and compare selective and nonselective 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 only accounts for $\pi-\pi$ stacking interactions



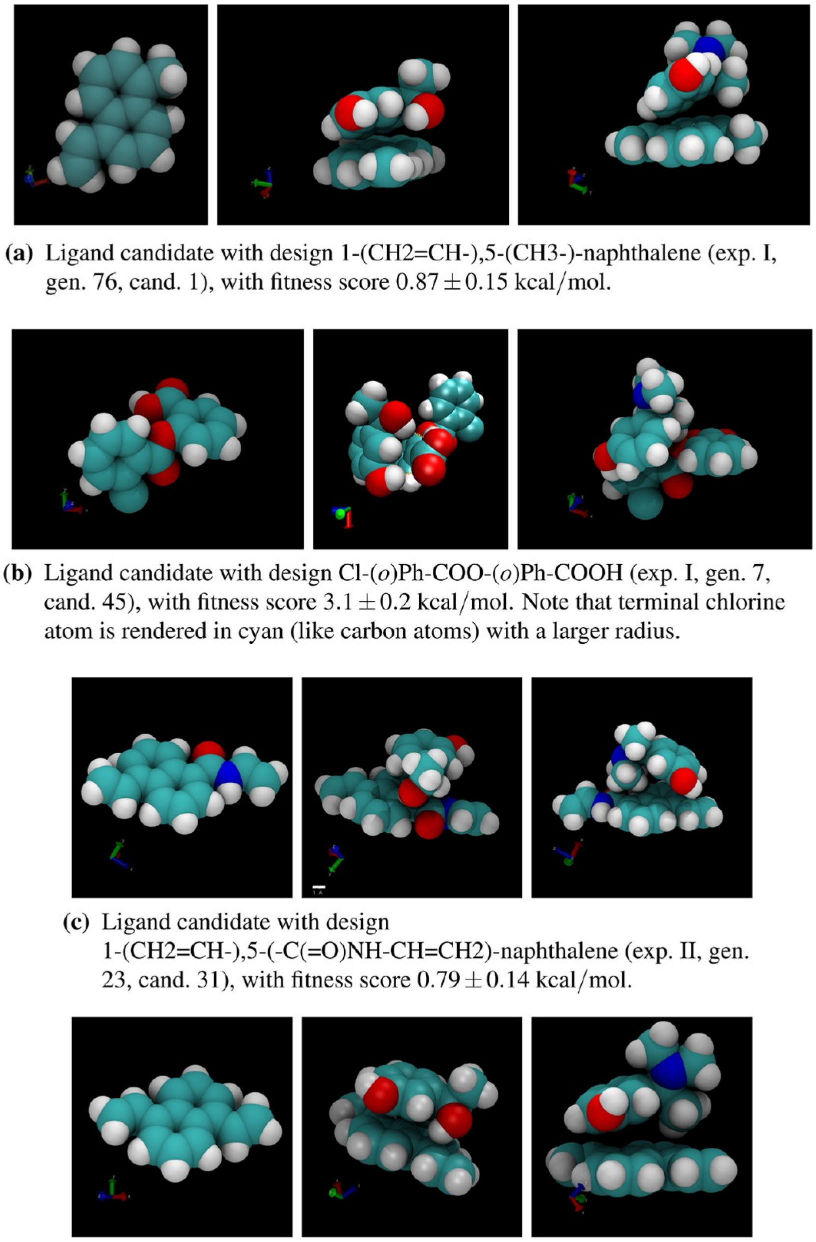
(a) Experiment I

(b) Experiment II

(c) Experiment III

(d) Experiment IV

Figure 16. Prevalence of motifs in experiments I–IV. Motif descriptions are listed in the title of each graph. An "aromatic group" is a phenyl or naphthalene group.



(a) Ligand candidate with design 1-(CH2=CH-),5-(CH3-)-naphthalene (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)-naphthalene (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-)-naphthalene (exp. II, gen. 45, cand. 1), with fitness score 0.88 ± 0.13 kcal/mol.

Figure 17. 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).

in an effective manner through nonbonded terms, and second, there is no natural way to apply the 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_r, 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 18 below.

The values of ϕ_q and h were then recorded and histogrammed for all configurations in which $\phi_r < \phi_{\text{cutoff}} = 60°$. This cutoff was imposed to avoid counting occurrences of

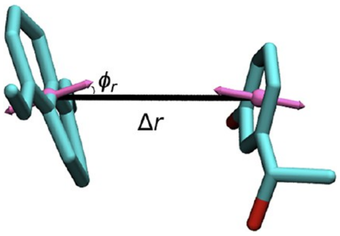
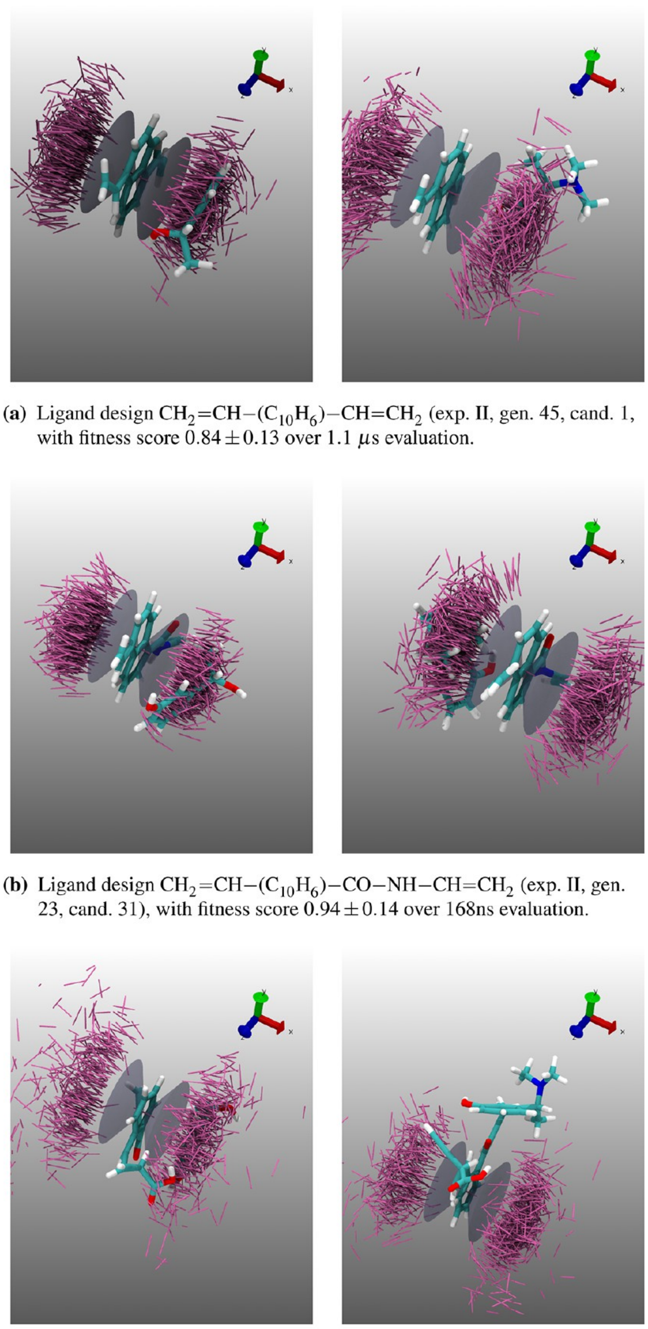


Figure 18. Illustration of the bond orientation angle ϕ_r for two aryl rings. The relative orientation angle ϕ_q is the (smaller) angle between the two normal vectors in purple. This notation is from ref 77. For clarity, hydrogen atoms are not shown.

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 Figure 19a and 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).



(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 19. Alignment of E2 (left) and E6 (right) molecules with three different ligand designs, for states in which $\phi_r < \phi_{\mathrm{cutoff}} = 60°$. 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 part c is a nonpositive control. For clarity, only 1000 such arrows are shown in each rendering.

Histograms of the aryl normal/aryl normal angle ϕ_q (Figure 19, 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 a 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 nonpositive control. In this case, the design in generation 3 and 20 in experiment II had a poor selectivity score of −0.36 kcal/mol over 9 ns of production. The nonselective control does not exhibit such noticeable differences between E2−ligand and E6−ligand adsorption configurations, as shown in Figure 20c. 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.

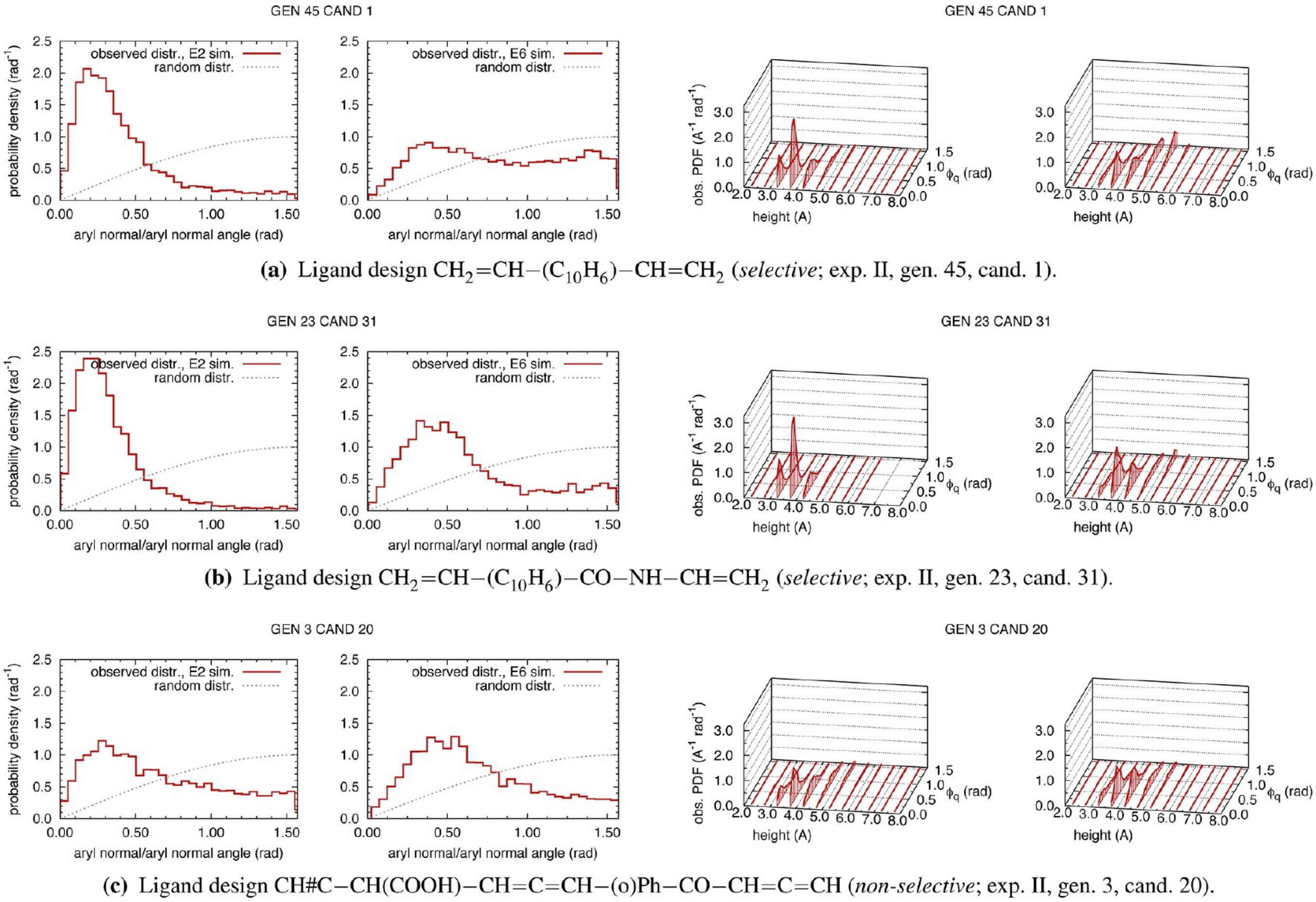
3.4. Top-Scoring Ligand Designs. Experiments III and IV were designed to take advantage of several practices we believed would improve the molecular evolution procedure: use of final coordinates as initial coordinates in repeated evaluations; accumulative scoring for ligand designs; population consolidation, 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 80th percentile scores in Figures 12−15. But these latter two experiments did identify several ligand designs that made up the top-scoring ligands overall, including the top 12 scorers (Table 5).

These ligands, like those discussed above, contained internal naphthyl groups (experiment III) or multiple phenyl rings (experiment IV).[78] 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) and alignment analyses of the kind shown above (data in the Supporting Information).

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, 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.

In fact, this hydrogen-bond acceptor motif was in agreement with our initial chemical reasoning: because the E2 molecule



(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 20. 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 $\phi_r < \phi_{\mathrm{cutoff}} = 60°$. The ligand design in part c is a nonpositive control.

Table 6. Observed Alignment and Hydrogen Bonding Behavior of Selected High-Scoring Ligands[a]

		$\langle \phi_q \rangle_{h=3.25\ \text{Å}}$ (deg)[c]		H-bonds observed[d]	
sequence	score[b] (kcal/mol)	E2	E6	E2	E6
COOH—(m)Ph—(m)Ph—Ph	3.67 ± 0.04	14	25	120 (100)	61 (61)
$CONH_2$—$C(CH_3)_2$—$C_{10}H_6$—$C_{10}H_6$—$C(CH_3)_2$—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)
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]Entries below the dotted line are nonselective controls containing aromatic groups. [b]Including score penalty. [c]Statistical errors of these average values were typically about 1°. [d]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.

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

3.5. Evolution Dynamics and Effect of Fitness Uncertainty.

To understand the dynamics of molecular evolution, we employed the "selection intensity" paradigm of Muhlenbein and co-workers. 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 $N(0,1)$".[72,79] 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 7 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 $\mu_t^{\mathrm{P}} = \mu_t + I\sigma_t$, and the subsequent generation will have the same fitness level as the parental subset.[72]

Table 7. Selection Intensity of Selection Schemes Used in This Work[a]

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

[a] Adapted from ref 72. [b] The order statistic $\mu_{n:k}$ is the expected value of the kth ordered (largest) of n values sampled from a unit normal distribution. According to a standard reference,[80] the value of $\mu_{2:2}$ for a two-member tournament is 0.564.

3.5.1. 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 12 and 13), *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 21a 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 13. 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 21a, 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 $1/(10) \times 1/2 \approx 5\%$ of its original value.

The function fitted to values of σ_t/μ_t in the middle panel of Figure 21a was used to predict the expected fitness score improvement $\mu_{t+1} - \mu_t = (\sigma_t/\mu_t)\sigma_t = (\sigma_t/\mu_t)^2\mu_t$. 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

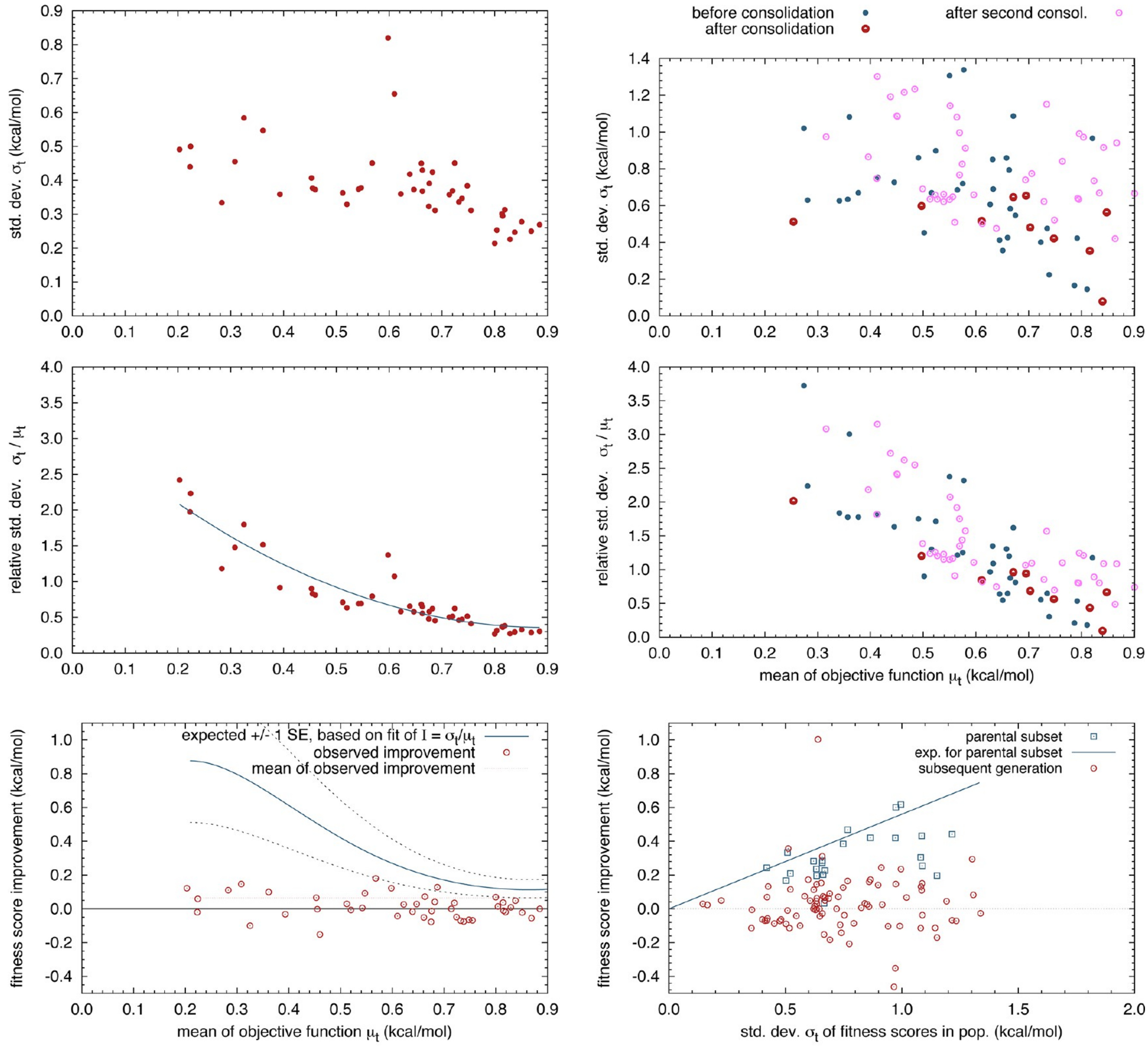
Diminishing fitness variance presents a similar, albeit less pressing, problem for evolutionary procedures employing tournament selection, like experiments III and IV. In these cases, however, the value of the selection pressure I is constant, so that $I\sigma_t$ decreases less severely (as μ_t increases) 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 0.35−0.70 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, the initial rate of fitness score improvement for experiments III and IV was greater than that of experiments I and II.

3.5.2. Noise in Fitness Function Evaluation.

Another complication when performing MD using automated molecular dynamics simulations is that estimated thermodynamic properties obtained from such simulations include statistical errors. In this work, the typical estimates of statistical error for the fitness function, due to finite sampling, are 0.15−0.20 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[72,81] 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:

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



(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 21. 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.

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 the following:

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

In this noisy case, the selection-based improvement in evolution (usually equal to $I\sigma_t$) in each generation is diminished by a factor $(\sigma_t/(\sigma_t^2 + \sigma_N^2)^{1/2})$ 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 21), 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.

3.5.3. Genetic Algorithm Performance. The four experiments performed in this study 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) ranks 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 4). 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 *consolidation*, 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.[82]

The effects of consolidation in experiment III can be seen in Figure 21b. Before consolidation, the value of and σ_t/μ_t as a function of μ_t is similar to those observed in experiment II (The quantity σ_t/μ_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 σ_t/μ_t-versus-μ_t curve was, in effect, shifted to the right, *increasing* the value of σ_t/μ_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 36−41 in Figure 14), although the median scores then returned to approximately the same level (approximately 0.88 kcal/mol) as before consolidation.

The bottom panel of Figure 21b 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: $\mu_{t+1} = \mu_t^{\mathrm{P}}$.

The bottom panel of Figure 21b 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 21b. 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.

4. CONCLUSIONS AND OUTLOOK

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., $\Delta\Delta E_{\mathrm{ads}}$) obtained from our simplified simulations for the top-scoring molecular designs were in the range 0.60−1.60 kcal/mol, corresponding to separation factors of approximately 3−15, if the entropic component $\Delta\Delta S_{\mathrm{ads}}$ 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 successful 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.

We would like to point out that this work is far from being the last word on the use of GAs in combination with MD simulations. A number of the issues that arose in this work deserve further investigation, such as the influence of each optimization parameter on sampling and convergence, and how to best deal with the uncertainty in fitness evaluation. Additionally, a comparison of the GA approach with other optimization approaches (including a well-designed random search) would be extremely interesting. Due to the computational cost of carrying out these studies with the example considered here, we decided to leave some of this questions for future studies, perhaps using a less computationally intensive example.

On the basis of the work in this study, we believe the molecular evolution approach could be improved and applied to new problems in molecular design. First, the molecular genome could be expanded in descriptive capability: it could describe functional groups using a base-2 string and the Gray encoding,[83] 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, nontopological information, such as molecular conformations;[84] lattice spacings[85] 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 has recently been used to model mathematical functions.[86] And of course, the set of functional groups that serve as building blocks of molecules can be expanded to include heteroaromatic groups, amino acids, and other groups. Finally, 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.[8]

In addition to improving the search algorithm, a number of improvements in the evaluation of the fitness function for separation problems are worth exploring for future work. In particular, ways to estimate the free energy of adsorption, and more realistic models of the adsorbent surface, may allow the search to identify additional candidate ligands that may achieve separation via other mechanisms than the ones observed in this work. We have developed a way to generate multiligand surfaces to reflect arrangement in a self-assembled monolayer, which we describe in the Supporting Information. We plan to study this type of arrangement in the future.

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 nonlinear 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 include (i) solution additives for viscosity modification, solubility enhancement,[87,88] or other property modification;[89−91] (ii) organic molecules/materials with desirable electronic properties like small band gap or nonlinear optical response;[33,34] (iii) chelation or binding with small molecules; (iv) protein docking and drug design; (v) novel materials for drug delivery;[92] and (vi) catalysis.[93−95]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

More results from the four computational experiments in this study—such as fitness score consistency, motif emergence and frequency, and score distributions under evolution. Also included are procedural details for the evaluation process, for both single- and multiligand simulations. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: trout@mit.edu.

Present Address

†E.E.S.: Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the Novartis−MIT Center for Continuous Manufacturing for supporting this work. The authors also kindly acknowledge several helpful discussions with Geoffrey P. F. Wood on encodings for and performance of genetic algorithms. Molecular graphics within this paper were produced using VMD[96] and OpenBabel.[97]

■ REFERENCES

(1) Li, L.; Bum-Erdene, K.; Baenziger, P. H.; Rosen, J. J.; Hemmert, J. R.; Nellis, J. A.; Pierce, M. E.; Meroueh, S. O. BioDrugScreen: a computational drug design resource for ranking molecules docked to the human proteome. *Nucleic Acids Res.* **2010**, *38*, D765−D773.

(2) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.

(3) 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.

(4) 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.

(5) 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.

(6) Matter, H.; Baringhaus, K. H.; Naumann, T.; Klabunde, T.; Pirard, B. Computational approaches towards the rational design of drug-like compound libraries. *Comb. Chem. High Throughput Screen* **2001**, *4*, 453−75.

(7) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. LEAP into the Pfizer Global Virtual Library (PGVL) space: creation of readily synthesizable design ideas automatically. In *Chemical Library Design*; Methods in Molecular Biology; Zhou, J. Z., Ed.; Springer: New York, 2011; Vol. *685*; pp 253−276 and references therein.

(8) Joback, K. G.; Stephanopoulos, G. Searching spaces of discrete solutions: the design of molecules possessing desired physical properties. In *Intelligent Systems in Process Engineering*; Advances in Chemical Engineering; Academic Press: Waltham, MA, 1995; Vol. *21*; pp 257−311.

(9) Patkar, P. R.; Venkatasubramanian, V. Genetic algorithms based CAMD. In *Computer Aided Molecular Design: Theory and Practice*; Computer-Aided Chemical Engineering; Elsevier: New York, 2003; Vol. *12*; pp 95−128.

(10) Eslick, J. C.; Shulda, S. M.; Spencer, P.; Camarda, K. V. Optimization-Based Approaches to Computational Molecular Design. In *Molecular Systems Engineering*; Process Systems Engineering; Wiley-VCH: New York, 2010; Vol. *6*; pp 173−194.

(11) Refs 8−10 and references therein.

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

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

(14) Gillet, V. J. De Novo Molecular Design. *Meth. Prin. Med. Chem.* **2000**, *8*, 49−66.

(15) Terfloth, L.; Gasteiger, J. Neural networks and genetic algorithms in drug design. *Drug Discov. Today* **2001**, *6*, S102−S108.

(16) Weininger, D. (Daylight Chemical Information Systems). Method and Apparatus for Designing molecules with Desired Properties by Evolving Successive Populations. US Patent 5,434,796, 1993.

(17) Rogers, D.; Tanimoto, T. A computer program for classifying plants. *Science* **1960**, *132*, 1115−1118.

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

(19) 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.

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

(21) Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297−302.

(22) 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.

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

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

(25) 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.

(26) 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.

(27) 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. Model.* **2006**, *46*, 545−552.

(28) Mandal, A.; Johnson, K.; Wu, C. F. J.; Bornemeier, D. Identifying promising compounds in drug discovery: Genetic algorithms and some new statistical techniques. *J. Chem. Inf. Model.* **2007**, *47*, 981−988.

(29) Dey, F.; Caflisch, A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* **2008**, *48*, 679−690.

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

(31) Kubinyi, H. Quantitative Structure−Activity Relationships in Drug Design. In *Encyclopedia of Computational Chemistry*; von Rague Schleyer, P., Ed.; Van Nostrand Reinhold, Wiley: New York, 1998; pp 2309−2320.

(32) Jurs, P. C. Quantitative Structure−Property Relationships (QSPR). In *Encyclopedia of Computational Chemistry*; von Rague Schleyer, P., Ed.; Van Nostrand Reinhold, Wiley: New York, 1998; pp 2320−2330.

(33) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sanchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. 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.

(34) Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sanchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **2011**, *4*, 4849−4861.

(35) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(36) 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.

(37) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. 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*, 671−690.

(38) Arbuse, A.; Anda, C.; Martinez, M. A.; Perez-Miron, J.; Jaime, C.; Parella, T.; Llobet, A. Fine-tuning ligand-receptor design for selective molecular recognition of dicarboxylic acids. *Inorg Chem* **2007**, *46*, 10632−10638.

(39) Monti, S.; Cappelli, C.; Bronco, S.; Giusti, P.; Ciardelli, G. Towards the design of highly selective recognition sites into molecular imprinting polymers: A computational approach. *Biosens. Bioelectron.* **2006**, *22*, 153−163.

(40) 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.

(41) 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.

(42) Ulman, A. Formation and structure of self-assembled monolayers. *Chem. Rev.* **1996**, *96*, 1533−1554.

(43) Gembicki, S.; Rekoske, J.; Oroskar, A.; Johnson, J. Adsorption, liquid separation. *Kirk-Othmer Encyclopedia of Chemical Technology*; John Wiley and Sons: New York, 2002; Vol. *1*; pp 678−691.

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

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

(46) Guest, D. Evaluation of simulated moving bed chromatography pharmaceutical process development. *J. Chrom. A* **1997**, *760*, 159−162.

(47) Deveant, R.; Jonas, R.; Schulte, M.; Keil, A.; Charton, F. Enantiomer Separation of a Novel Ca-Sensitizing Drug by simulated moving bed (SMB)−chromatography. *J. Prakt. Chem.* **1997**, *339*, 315−321.

(48) 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.

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

(50) 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.

(51) 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.

(52) 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.

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

(54) Juza, M.; Mazzotti, M.; Morbidelli, M. Simulated moving-bed chromatography and its application to chirotechnology. *Trends Biotechnol.* **2000**, *18*, 108−118.

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

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

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

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

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

(60) 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.

(61) 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.

(62) Schaber, S. D.; Gerogiorgis, D. I.; Ramachandran, R.; Evans, J. M. B.; Barton, P. I.; Trout, B. L. Economic Analysis of Integrated Continuous and Batch Pharmaceutical Manufacturing: A Case Study. *Ind. Eng. Chem. Res.* **2011**, *50*, 10083−10092.

(63) Centrone, A.; Santiso, E. E.; Hatton, T. A. Separation of Chemical Reaction Intermediates by Metal-Organic Frameworks. *Small* **2011**, *7*, 2356−2364.

(64) Single-point calculations from low-energy MD frames, performed using the correlation-consistent cc-pVTZ[98] basis set and RIMP2 correlation with a matching auxiliary basis set.[99,100] The dipole moments from two configurations were Boltzmann-weighted to calculate to the average values.

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

(66) 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.

(67) 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.

(68) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.

(69) Marcus, Y. *The Properties of Solvents*; Wiley: New York, 1998.

(70) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford: New York, 1987.

(71) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

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

(73) Augustson, J. G.; Minker, J. An analysis of some graph theoretical cluster techniques. *J. Assoc. Comput. Mach.* **1970**, *17*, 571−588.

(74) Oprea, T.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157−166.

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

(76) The reason for using the 80th 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 80th percentile score as illustrated in Figure B-1 in the Supporting Information.

(77) Santiso, E. E.; Trout, B. L. A general set of order parameters for molecular crystals. *J. Chem. Phys.* **2011**, *134*, 064109−064109−16.

(78) In such cases, our GAFF atom type assignments allowed directly linked aromatic groups (as in biphenyl) to assume a coplanar configuration.

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

(80) Harter, H. L.; Balakrishnan, N. *CRC Handbook of Tables for the Use of Order Statistics in Estimation*; CRC Press: Boca Raton, FL, 1996.

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

(82) 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 $I\sigma_t$.

(83) Savage, C. A Survey of Combinatorial Gray Codes. *SIAM Rev.* **1997**, *39*, 605−629.

(84) 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.

(85) Bianchi, E.; Doppelbauer, G.; Filion, L.; Dijkstra, M.; Kahl, G. Predicting patchy particle crystals: Variable box shape simulations and evolutionary algorithms. *J. Chem. Phys.* **2012**, *136*, 214102−214102−9.

(86) Bellucci, M. A.; Coker, D. F. Empirical valence bond models for reactive potential energy surfaces: A parallel multilevel genetic program approach. *J. Chem. Phys.* **2011**, *135*, 044115.

(87) Sagisaka, M.; Oike, D. K.; Mashimo, Y.; Yoda, S.; Takebayashi, Y.; Furuya, T.; Yoshizawa, A.; Sakai, H.; Abe, M.; Otake, K. Water/supercritical CO2 microemulsions with mixed surfactant systems. *Langmuir* **2008**, *24*, 10116−10122.

(88) Sagisaka, M.; Iwama, S.; Hasegawa, S.; Yoshizawa, A.; Mohamed, A.; Cummings, S.; Rogers, S. E.; Heenan, R. K.; Eastoe,

J. Super-Efficient Surfactant for Stabilizing Water-in-Carbon Dioxide Microemulsionst. *Langmuir* **2011**, *27*, 5772−5780.

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

(90) Hamada, H.; Arakawa, T.; Shiraki, K. Effect of additives on protein aggregation. *Curr. Pharm. Biotechnol.* **2009**, *10*, 400−407.

(91) Mohamed, A.; Sagisaka, M.; Guittard, F.; Cummings, S.; Paul, A.; Rogers, S. E.; Heenan, R. K.; Dyer, R.; Eastoe, J. Low Fluorine Content CO2-philic Surfactants. *Langmuir* **2011**, *27*, 10562−10569.

(92) 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*, 1031−1049.

(93) Andersson, M.; Bligaard, T.; Kustov, A.; Larsen, K.; Greeley, J.; Johannessen, T.; Christensen, C.; Norskov, J. Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts. *J. Catal.* **2006**, *239*, 501−506.

(94) Baerns, M.; Holena, M. *Combinatorial Development of Solid Catalytic Materials*; Catalytic Science Series; Imperial College Press: London, 2009; Vol. 7.

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

(96) Humphrey, W.; Dalke, A. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(97) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(98) 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.

(99) 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.

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