J-CAMD 271

Compass: A shape-based machine learning tool for drug design

Ajay N. Jain[a,*], Thomas G. Dietterich[a,b], Richard H. Lathrop[a,c], David Chapman[a,*], Roger E. Critchlow Jr.[a], Barr E. Bauer[a], Teresa A. Webster[a,d] and Tomas Lozano-Perez[a,c]

[a]*Arris Pharmaceutical Corporation, 385 Oyster Pt. Boulevard, South San Francisco, CA 94080, U.S.A.*
[b]*Computer Science Department, Oregon State University, Corvallis, OR 97330, U.S.A.*
[c]*Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.*
[d]*Computer Science Department, Stanford University, Stanford, CA 94305, U.S.A.*

Received 28 April 1994
Accepted 11 August 1994

Key words: Automated prediction; QSAR; Molecular shape; Ligand binding; Molecular recognition

SUMMARY

Building predictive models for iterative drug design in the absence of a known target protein structure is an important challenge. We present a novel technique, Compass, that removes a major obstacle to accurate prediction by automatically selecting conformations and alignments of molecules without the benefit of a characterized active site. The technique combines explicit representation of molecular shape with neural network learning methods to produce highly predictive models, even across chemically distinct classes of molecules. We apply the method to predicting human perception of musk odor and show how the resulting models can provide graphical guidance for chemical modifications.

INTRODUCTION

Analysis of biophysical data, such as X-ray crystal structures of drug–target complexes, has allowed researchers to rapidly improve characteristics of drug molecules by iterative design, synthesis, and structure determination [1,2]. However, in the absence of such data, rational drug design must rely upon predictive models derived solely from observed biological activity. In many drug discovery efforts one may have a functional biological assay, but detailed knowledge of the biochemical target or its structure is unavailable. We present a novel modeling approach, called Compass, that is applicable in such circumstances.

Compass combines three primary features. It employs a computationally efficient molecular representation that characterizes surface shape such that structurally diverse molecules exhibiting

*To whom correspondence should be addressed.

similar surface characteristics are treated as being similar. Several other methods characterize molecular shape [3–11], but Compass differs from some by computing features based exclusively on external molecular surfaces. The representation differs from pharmacophoric models [7–9] in that it is an unbiased and fine surface sampling and is not limited to a small number of specific atom-based features. Second, Compass's machine learning algorithm analyzes and selects from multiple conformations and orientations of molecules. Without the benefit of structural data, this is a critical issue. Drug design methods that consider only single conformations and alignments are of limited utility. Third, Compass improves predictive performance through an adaptive alignment process that automatically generates molecular orientations.

As a development task for the system we have used the prediction of musk odor, for which there is a functional, reproducible biological assay, and no structural data are available as to the biochemical target involved. Odor perception is believed to be mediated by selective binding of odorants to G-protein-coupled receptors in the olfactory neuroepithelium [12]. A wealth of data is available in the area, and the problem has been the focus of many modeling efforts [5,6,13–18]. We present the machine learning procedure, experimental results on two types of prediction tasks, and examples of graphical model interpretation for guidance in structural modification.

METHODS

This section begins with an overview of the algorithm, then elaborates on the three major features, and lastly gives details of the implementation and the test system. Figure 1 shows a flow

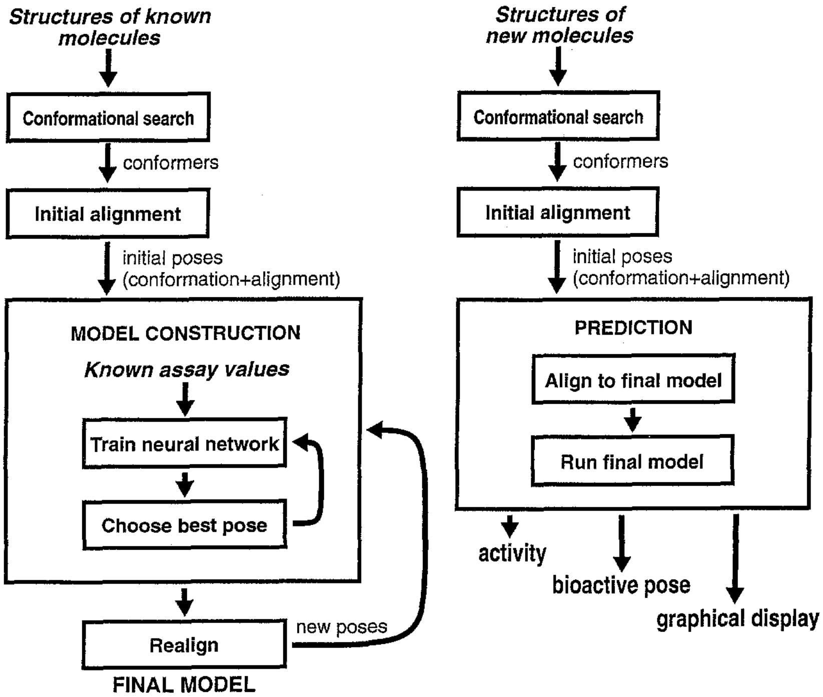


Fig. 1. Flow diagrams of the machine learning and prediction procedures of Compass.

diagram of the procedure. It begins by conducting a search for low-energy conformers of the training molecules. This provides a pool of energetically accessible shapes for each molecule. They are then placed in rough initial orientations by using a simple pharmacophoric alignment. A specific conformation and orientation is called a *pose*. From these starting poses, we construct an initial model of activity by training a neural network model. A neural network is simply a sum of nonlinear functions, having trainable weights, which can be cascaded. The model is trained with a numerical characterization of surface shape as input and measured activity as output. Choice among poses is interleaved with incremental neural network training to allow the network to examine all the available poses during training. To improve predictive performance, we use the model to generate new molecular poses for each molecule. The model is refined using the new poses, and the process iterates until it converges on a best pose for each molecule within the model. Activity predictions for new molecules are obtained by applying the resulting model. As in the training process, the model automatically selects the conformation and orientation for each molecule. It can be visualized in three dimensions to identify favorable and unfavorable interactions for a candidate molecule.

Shape-based molecular representation

The pose of a molecule is parameterized by several values. The conformation varies with the internal torsion angles of the rotatable bonds and the alignment varies with three rotational and three translational parameters. From a molecular pose, we generate a high-dimensional vector of features for purposes of activity prediction. Each element of the feature vector is associated with a single reference point in a fixed three-dimensional coordinate system. The minimum distance from the reference point to a molecule's van der Waals surface in a particular pose is the value of the feature. These features are piecewise differentiable functions of molecular pose.

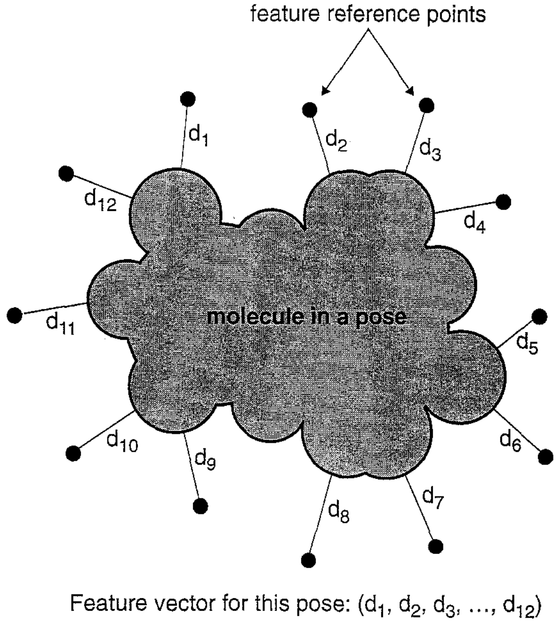


Fig. 2. Diagram of the feature computation in two dimensions.

Essentially, the molecule is 'measured' from the outside, so molecules with similar shapes that have different underlying scaffolds will have similar computed feature values. Figure 2 illustrates the feature computation.

Neural network model

We construct a model that predicts the activity of a molecule as a function of its feature vector, but a complication arises since feature vectors vary with molecular pose. We define the predicted activity for a molecule to be the *maximum* value of the model over all possible poses that correspond to energetically accessible molecular conformations. In chemical terms, this is analogous to permitting a molecule to rotate, translate and alter its conformation to achieve the best possible fit to a binding site. The learning algorithm interleaves two processes: neural network construction given molecular poses, and generation of maximally active molecular poses (described in the next section).

Given a set of poses, a neural network model is trained to produce the desired mapping from the maximally active pose of each molecule to the molecule's observed biological activity. The weights of the functions that comprise the neural network are computed by a procedure that minimizes the difference between activity computed by the evolving neural network and observed biological activity. This is done by repeatedly making small adjustments to the weights [19]. For each molecule, the pose giving the highest predicted activity (according to the current model) is used to update the weights. The resulting neural network model is a continuous, differentiable function of the feature vectors, which are themselves functions of molecular pose.

Given a receptor with a cavity of some unknown shape, parameter estimation by training the neural network produces a function that mimics the effects of molecules interacting with the walls of the cavity. A favorable contact between a molecular surface and the receptor is learned from active molecules sharing similar distances to a reference point. The function that the neural network learns will have a positive contribution in that part of the feature's space. If the receptor surface is relatively rigid and there are less active molecules that have shorter distances to the reference point (i.e., they protrude toward the receptor), a negative contribution will be learned in that part of the feature's space. Note that even in the simple case of a rigid walled receptor, the function of a single feature should be nonlinear. If the surface falls short, there should be *zero* contribution; if it makes contact, there should be a *positive* contribution; and if the surface protrudes too far, there should be a *negative* contribution. Linear PLS models (indeed, any monotonic functions of features) cannot capture such effects.

The initial neural network model will capture some of the significant geometric binding determinants, but these are derived from only a coarse initial pose set. Improved molecular poses are chosen by applying the model.

Adaptive molecular pose generation

Using gradient-based search techniques, the learning algorithm adaptively varies the six pose parameters of rotation and translation for the conformations of each molecule and selects the best combination of conformation and orientation. For each molecule, the gradient of the neural network model with respect to the pose parameters is computed. This gradient guides the modification of the pose parameters to maximize the model's output. Poses computed in this fashion for the active molecules cause the molecules to align more tightly with each other along those portions of the molecular surface that are important for activity prediction. The poses computed for inactive molecules provide examples of how molecules can best 'fool' the model and thus have discriminative value. The process of model refinement and adaptive alignment iterates until it converges on a model whose predictions are not affected by additional adaptive alignment. The space of possible poses is infinite, but only a small finite portion needs to be explored during training, and the resulting model is able to choose poses of new molecules robustly. Gradient methods, although sensitive to local minima, were sufficient in the study reported here. More complex search methods did not improve the performance. The adaptive pose generation process generalizes trivially to adaptive variation of conformational parameters, but this has not been tested.

Program implementation

The implementation of the program can be described with reference to the pseudocode given below and the flow chart in Fig. 1. The detailed process of generating the initial poses for this task is described later, so the discussion will begin with the model construction process, given the initial set of poses.

```
1.  construct_model {
2.      create initial neural network with random weights
3.      until converged {
4.          train_network_from_poses
5.          realign_molecules
6.      }
7.      output final neural network
8.  }
```

As shown in Fig. 1, the main construct_model function interleaves neural network training (line 4) with molecular realignment (line 5). These two functions are given below:

```
1.  train_network_from_poses {
2.      while (iterations < MAX) {
3.          for each molecule {
4.              for each pose {
5.                  run_network(pose)
6.              }
7.              mark the molecule's best pose
8.          }
9.          for each molecule {
10.             run_network(best_pose)
11.             compute_error
12.             modify_weights
13.         }
14.     }
15. }
```

The train_network_from_poses function iterates over the whole data set (line 10) using the current set of existing poses (either from the initial conformational search and alignment or from subsequent realignment). Within that loop, the system first marks the current best poses for each molecule (lines 11–16), then runs one 'epoch' of neural network learning (lines 17–22) [19]. The process results in a neural network that performs well on the current set of poses, but which may require refinement given newly generated ones.

```
realign_molecules {
    for each molecule {
        for each pose {
            new_pose = pose
            for 20 iterations {
                run_network(new_pose)
                network_gradient = compute_network_gradient
                feature_gradient = compute_feature_gradient
                jacobian = compute_jacobian(network_gradient,feature_gradient)
              new_pose = generate_pose(new_pose, jacobian) /* step size 0.03 */
          }
          add_to_pose_set(new_pose)
      }
    }
}
```

The realign_molecules function realigns each molecule to the current neural network. It goes through each pose and performs a gradient-based pose modification by following the computed Jacobian. The function generate_pose computes a new pose and feature vector, based on an incremental adjustment of the parameters of pose.

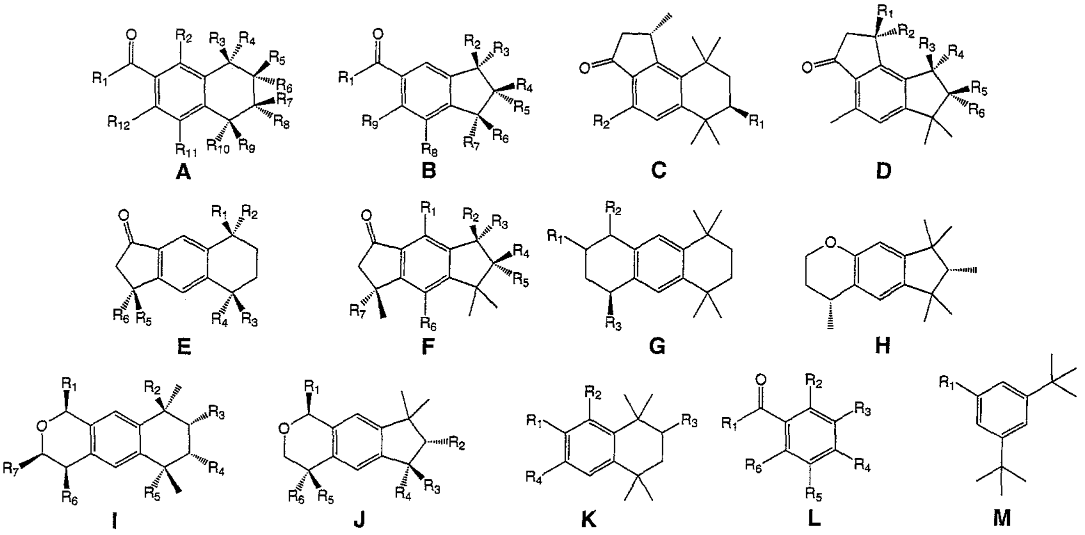


Fig. 3. Body types of the 102 molecules in the test system.

TABLE 1
MOLECULES USED IN CROSS-VALIDATION AND CLASS-HOLDOUT EXPERIMENTS[a]

Molecule	Type	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	Act.	Pred.
1	A	Me		Me	Me					Me	Me			0	0.41
2	A	Me		Me	Me					Me	Me		Et	1	0.64
3	A	iPr		Me	Me					Me	Me		Et	0	0.36
4	A			Me	Me					Me	Me			0	0.20
5	A	Me		Me	Me					Me	Me		Me	1	0.99
6[b]	A	Me		Me	Me			Me		Me	Me			1	0.87
7	A	Me		Me	Me			Me		Me	Me		iPr	0	0.41
8	A	Me		Me	Me					Et	Me		Me	1	0.52
9	A	Me		Me	Et					Me	Et		Me	0	0.45
10	A	Me		Me	Me								Et	0	0.01
11	A	Me		Me	Me					iPr			Me	0	0.33
12	A	Me		Et	Me					Me	Et		Et	0	0.34
13	A	Me		Me	Et					Me	Et		iPr	0	0.01
14	A	Me		Me	Me								Me	0	0.02
15	A	iPr		Me	Me					Me	Me		Me	0	0.00
16	A			Me	Me					Me	Me	Me	Me	1	0.80
17	A		Me	Me	Me					Me	Me		Me	1	0.86
18	A	Me	Me	Me	Me					Me	Me		Me	0	0.14
19	A		Me	Me	Me					Me	Me		Me	1	0.96
20	A		Me	Me	Me	Me				Me	Me		Me	1	0.99
21	A			Me	Me	Me			Me	Me	Me		Me	1	0.93
22	A	Me		Me	Me		Me		Me	Me	Me		Me	1	0.93
23	A			Me	Me			Me		Me	Me		Me	1	0.92
24	A	Me		CH_2–	Me					CH_2–	Me		Me	0	0.02
25	A	Me		Me	Me					Me	iPr		Me	0	0.03
26	A			Me	Me					Me	CH_2–	CH_2–	Me	1	0.95
27	A	Me		Me	Me					Me	CH_2–	CH_2–	Me	1	0.85
28	B	Me	Me	Me	Me		Me	Me		Me				1	0.83
29	B	Me	Me	Me		Me		Me		Me				0	0.32
30	B	Me	Me	Me			Me	nPr		Me				0	0.57
31	B		Me	Me		Me	Me	Me		Me				1	0.93
32	B	Me		iPr	Me		Me	Me		Me				1	0.10
33	B	Me		iPr			Me	Me		Me				0	0.12
34	B	Me	Me	Me			Me	Me		iPr				0	0.01
35	B	Me		iPr			Me	Me		Et				0	0.04
36	B	Me		iPr		Et	Me	Me		Et				0	0.08
37	B	Me	Me	Me					tBt					1	0.82
38	B	Et	Me	Me					tBt					1	0.84
39	B	Me	Me	Me					tBt-Me					1	0.94
40	B						Me	Me						0	0.00
41	B	Me	Me	Me										0	0.00
42	C	Me	Me											1	0.98
43	C		Et											1	0.95
44	D			Me	Me	Me	Me							1	0.94
45	D	Me	Me	Me	Me									0	0.84
46	D	Me	Me											0	0.00
47	E	Me	Me	Me	Me	Me	Me							0	0.05
48	E	Me	Me	Me	Me									1	0.77
49	E					Me	Me							0	0.00
50	F		Me	Me		Me								1	0.72
51	F		Me	Me				Me						0	0.01
52	F		Me	Me	Me		Me							0	0.04
53	F	Me					Me							0	0.00
54	F			Me	Me	Me								0	0.58

[a] The data are for body types A–G, as given in Fig. 3. Abbreviations: Me = methyl; Et = ethyl; iPr = isopropyl; nPr = *n*-propyl; tBt = *t*-butyl; tBt-Me = $C(CH_2)_2C_2H_5$. An activity of 1 = musk and 0 = non-musk. The predicted values are from a 20-fold cross-validation. R groups with a '–' are connected. The isomer predicted to be most active is shown.

[b] This is known to be the active stereoisomer and was predicted to be so from a mixture.

TABLE 2
MOLECULES USED IN CROSS-VALIDATION AND CLASS-HOLDOUT EXPERIMENTS[a]

Molecule	Type	R_1	R_2	R_3	R_4	R_5	R_6	R_7	Act.	Pred.
55	G		=O	Me					0	0.37
56	G		=O						0	0.40
57	G	=O							0	0.00
58	G	=O		Me					0	0.08
59	H								0	0.00
60	I		Me			Me			1	0.81
61	I		Me	Me	Me	Me			1	0.98
62	I		Me			Me	Me		1	0.97
63	I		Et			Et			0	0.62
64	I		Me			Me		Me	0	0.49
65	I	Me	Me			Me			0	0.08
66	I		Me	Me		Me			1	0.97
67	J		Me	Me	Me				1	0.85
68	J			Me	Me	Me			1	0.23
69	J			Me	Me	Me	Me		0	0.10
70	J		Me	Me	Me	Me			1	0.99
71	J					Me	Me		0	0.00
72	J	Me	Me	Me	Me				0	0.78
73	K			=O					0	0.00
74	K	CH_2–	$COCH_2$–		Me				0	0.00
75	K	O–	C_3H_6–						0	0.00
76	K	CH_2–O–	CH_2–						1	0.80
77	K	Me		=O					0	0.00
78	L	iPr		tBt		tBt			0	0.05
79	L	Me			tBt		tBt		0	0.00
80	L			tBt			tBt		0	0.00
81	L	Me	iPr		iPr		iPr		0	0.00
82	L	Me	CH_2CH_2–	CH_2–		tBt			0	0.89
83	L	Me	CH_2–	$C(CH_3)_2CH_2$–		tBt			0	0.00
84	L	Me		tBt		$C(CH_3)_2CH_2$–	CH_2–		1	0.92
85	L			tBt		$C(CH_3)_2CH_2$–	CH_2–		1	0.92
86	L	Et		tBt		$C(CH_3)_2CH_2$–	CH_2–		1	0.65
87	L	iPr		tBt		$C(CH_3)_2CH_2$–	CH_2–		0	0.83
88	L	Me		tBt		$C(CH_3)(C_2H_5)$–	CH_2– (*S* isomer)		1	0.80
89	L	Me		iPr		$C(CH_3)_2CH_2$–	CH_2–		1	0.94
90	L	Me		$C(CH_2)_2CH_2$–**4**	CH_2–**3**	$C(CH_3)_2CH_2$–**6**	CH_2–**5**		1	0.92
91	L	Me	CH_2–**3**	$C(CH_2)_2CH_2$–**2**		CH_2–**6**	$C(CH_3)_2CH_2$–**5**		0	0.00
92	L			$C(CH_2)_2C_2H_5$	$C(CH_3)_2CH_2$–	CH_2–			1	0.91
93	L	Me				$C(CH_3)_2CH_2$–	CH_2–		0	0.00
94	L	Me	$C(CH_3)_2CH_2$–	CH_2–					0	0.00
95	L	Me		tBt		C_3H_6–	CH_2–		0	0.01
96	L	CH_2–			tBt		$C(CH_2)_2$–		0	0.00
97	L	CH_2–	CH_2–		tBt		tBt		0	0.00
98	L	CH_2–**6**	$C(CH_3)_2CH_2$–**3**	$C(CH_3)_2$–**2**			CH_2–**1**		0	0.00
99	L	CH_2–**6**	$C(CH_3)_2CH_2$–**3**	$C(CH_2)_2$–**2**		Me	CH_2–**1**		0	0.00
100	L	CH_2–			tBt		C_2H_4–		0	0.00
101	M	$CH{=}CHCOCH_3$ (*trans*)							0	0.13
102	M	C_2H_4CHO							0	0.45

[a] The data are for body types H–M, as given in Fig. 3. Abbreviations: Me = methyl; Et = ethyl; iPr = isopropyl; tBt = *t*-butyl. An activity of 1 = musk, 0 = non-musk. The predicted values are from a 20-fold cross-validation. R groups with a '–' are connected. R groups with a '–**N**' are connected to R_N. The isomer predicted to be most active is shown.

TABLE 3
ACTIVE CONFORMATION OF MOLECULE 4[a]

No.	Atom	X	Y	Z	B_1	B_2	B_3	B_4
1	C	−1.678933	−1.876324	−0.002684	2 [2]	6	15	
2	C	−2.280481	−0.593324	−0.030263	1 [2]	3	17	
3	C	−1.441443	0.539712	−0.038728	2	4 [2]	19	
4	C	−0.036315	0.446614	−0.011419	3 [2]	5	7	
5	C	0.560204	−0.834212	−0.002464	4	6 [2]	8	
6	C	−0.274069	−1.967928	0.016510	5 [2]	1	20	
7	C	0.799726	1.724469	−0.004569	4	13	10	14
8	C	2.075701	−1.018575	−0.000270	5	12	9	11
9	C	2.834269	0.266685	−0.371400	8	10	21	22
10	C	2.270442	1.475569	0.370240	7	9	23	24
11	C	2.556357	−2.062139	−1.026628	8	25	26	27
12	C	2.572593	−1.507757	1.371501	8	28	29	30
13	C	0.750654	2.426121	−1.373244	7	31	32	33
14	C	0.308148	2.759765	1.025141	7	34	35	36
15	C	−2.457163	−3.144277	0.011825	16 [2]	1	37	
16	O	−3.643372	−3.200949	−0.265693	15 [2]			
17	C	−3.796466	−0.419206	−0.040024	2	18	38	39
18	C	−4.351220	1.005847	0.067456	17	40	41	42
19	H	−1.881816	1.520456	−0.069796	3			
20	H	0.179481	−2.945843	0.045496	6			
21	H	3.892443	0.164545	−0.126802	9			
22	H	2.757462	0.443428	−1.443855	9			
23	H	2.874803	2.351023	0.128721	10			
24	H	2.353205	1.301723	1.442701	10			
25	H	3.645672	−2.057980	−1.083064	11			
26	H	2.242010	−3.068680	−0.753735	11			
27	H	2.155502	−1.823984	−2.012790	11			
28	H	3.659677	−1.593624	1.378921	12			
29	H	2.266754	−0.819289	2.158523	12			
30	H	2.148634	−2.486808	1.596009	12			
31	H	1.383684	3.314037	−1.373692	13			
32	H	1.084776	1.752913	−2.161956	13			
33	H	−0.270876	2.731884	−1.600782	13			
34	H	1.013874	3.589005	1.089052	14			
35	H	−0.662506	3.170663	0.750362	14			
36	H	0.226929	2.294599	2.008463	14			
37	H	−1.949947	−4.075077	0.219376	15			
38	H	−4.184413	−0.843070	−0.967197	17			
39	H	−4.209112	−0.978211	0.800850	17			
40	H	−5.440460	0.959782	0.092989	18			
41	H	−4.002126	1.480617	0.985215	18			
42	H	−4.052277	1.597810	−0.798279	18			

[a] X,Y,Z give the coordinates of the atom. B_{1-4} give the identities of the bonded atoms (and the order if > 1 (between square brackets)).

Test system: Musk odor prediction

We collected a set of 102 diverse structures in several chemical classes from published studies [5,13,18,20]. The data set contained 39 aromatic, oxygen-containing molecules with musk odor and 63 homologs that lacked musk odor (see Fig. 3 and Tables 1 and 2). A musk molecule requires a hydrogen-bond acceptor on a roughly ellipsoidal hydrocarbon, and its odor is strongly

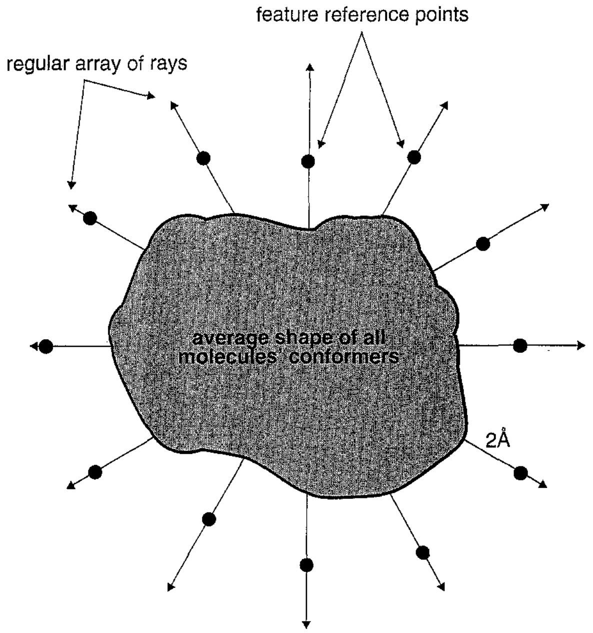


Fig. 4. Diagram of feature reference point placement.

dependent on subtle steric interactions [20–23]. Musk odor is a highly specific property [24], and the addition or deletion of a single methyl group can convert an odorless compound into a strong musk. Musk odor is assessed by human experts and as such is somewhat subjective. We included as active molecules only those which have been labeled in at least two references to have musk odor. Molecules which in any reference were marked as medium in strength or on which references disagreed were not included in the set.

Each molecule was conformationally searched using the Monte Carlo procedure of Still [25]. All rings and torsions were searched; structures were energy-minimized within MacroModel [26] using the AMBER force field [27] in vacuo to less than 0.05 kJ/Å gradient. All local minima within 20 kJ/mol of the global minimum were retained. Some molecules possessed flexible side chains and exhibited over 250 conformations, many of which significantly changed the overall shape of

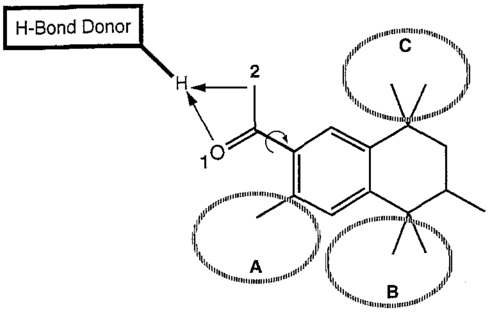


Fig. 5. Schematic diagram of learned requirements for musk odor activity. The learned model is sensitive to about 50 specific surface patches, mostly in regions A, B and C.

the molecule. A further complication was that nearly all chiral molecules were assayed as mixtures, so all stereoisomers were included. The final data set contained 6953 conformations of the 102 molecules. Table 3 gives the coordinates of the predicted active conformation of molecule 5*.

All conformations of all molecules were placed in an initial alignment by rms fit to the six carbons of the arene of an active reference molecule (molecule 5), placed at the center of the coordinate system. The correspondence of atoms was chosen to best match the oxygen-containing substituent of the reference molecule. This brought the backbones of the molecules into tight alignment and placed their oxygens to allow a hydrogen bond with an assumed hydrogen-bond donor atom [28,29]. The initial alignment must be sufficiently good to allow Compass to detect some areas of regularity that are predictive of activity. However, in our work to date, pharmacophoric substructure-based alignments have been sufficient for Compass to build an initial model that is sufficiently meaningful to allow the adaptive pose generation process to move productively forward.

Feature reference points were placed an average of 2 Å from the molecules in their initial poses along a set of 162 rays emanating in a regular pattern from the origin of the coordinate system (see Fig. 4). Placing the points with respect to the starting poses ensures that the molecules will be adequately sampled. The distance of the reference points from the molecular surface affects the effective shape resolution of the trained neural network. Functions constructed from points placed further away from the molecules produce spatially smoother fields. The value of 2 Å is approximately the radius of a methyl group and biases the network to learn fields that have that level of surface detail. Placement of reference points need only guarantee that molecules are ‘covered’. In cases where a new molecule has no reference points that ‘measure’ part of its surface, the molecule is probing areas of space not explored by existing data and cannot be predicted with confidence.

The neural network had 162 input units (one for each feature), three hidden units, and one output unit that produced a value between 0.0 and 1.0. The output activity of active molecules was encoded as 1.0, and the activity of inactives was encoded as 0.0. A molecule was predicted to be active if the model computed its activity to be greater than or equal to 0.5. The learning rate was 0.05, and no momentum term or acceleration methods were used. Each round of learning consisted of 150 repetitions of the training set. This was followed by adaptive molecular alignment using 20 gradient steps per conformation with a step size of 0.03. For this problem to attain convergence, at most five iterations of model construction and pose generation were required.

We performed our experiments on a Silicon Graphics Iris Indigo with an R4000 processor and 96 MB of main memory. The run time of the initial pose generation process is dominated by conformational search, which requires several hours per molecule. However, this needs to be done only once for a set of molecules. Model building takes about a minute per molecule, and the resulting model can be applied to predict a new molecule in seconds.

RESULTS AND DISCUSSION

We conducted three sets of experiments that measured the learning algorithm’s ability to make biological activity predictions within mixed chemical classes, predictions across chemical classes, and to provide guidance for structural modification.

*The molecules and conformations can be obtained from the authors.

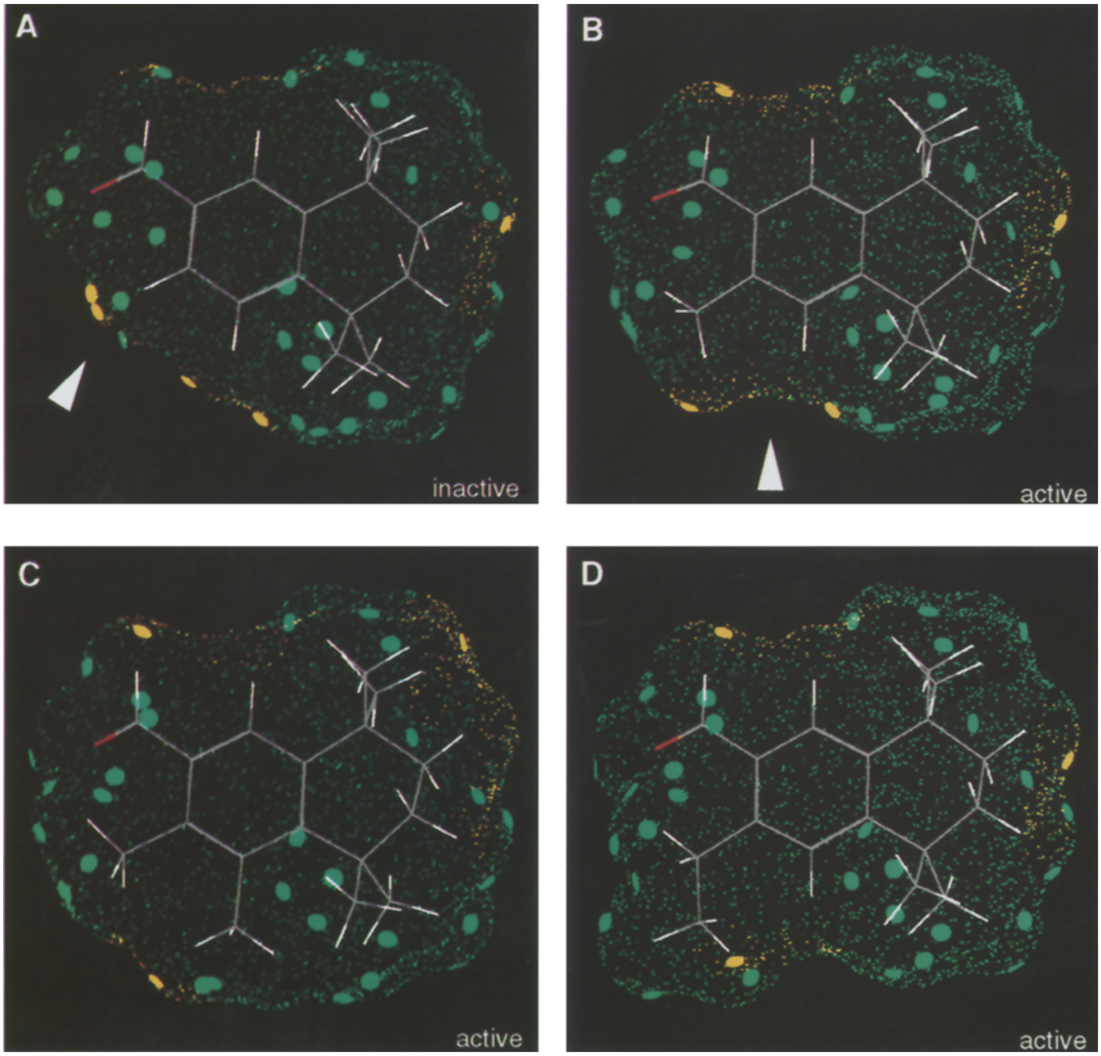


Fig. 6. Application of shape-based models to molecular modification. (A) Non-musk lacking sufficient bulk; (B) moderate-activity analog of molecule A; (C and D) analogs of molecule B with strong activity.

Overall predictive performance

To test predictive performance on a structurally mixed pool, the molecules were randomly partitioned into 20 subsets. We measured the effects of conformational selection and adaptive alignment in four experiments. In each experiment 20 models were trained, with one of these subsets excluded from the training data during each execution. The model constructed in each execution was then tested to see how well it could predict whether the withheld molecules were musks or non-musks, and the results were added (this process is called cross-validation). Table 4 summarizes the results. A model constructed from the initial fixed molecular alignments and the single lowest energy conformer for each molecule resulted in a predictive performance of just

75%, with errors roughly split between false negatives and false positives. By allowing the model to choose from *multiple conformations*, the performance jumped to 81%, with a large reduction in false negatives. We hypothesize that the lowest energy conformers are often not those that are biologically active. With multiple conformations *and* adaptive alignment of molecules, the predictive performance was 91%, with just three musks predicted to be inactive and six non-musks predicted to be active (Tables 1 and 2 show the predictions from this cross-validation). Adaptive alignment allowed the model to find molecular orientations that aided in discriminating inactive molecules. In the single-conformer case, however, adaptive alignment made no significant improvement – the algorithm could not overcome a lack of conformational choices. Conformational selection, coupled with adaptive alignment, substantially improved the performance.

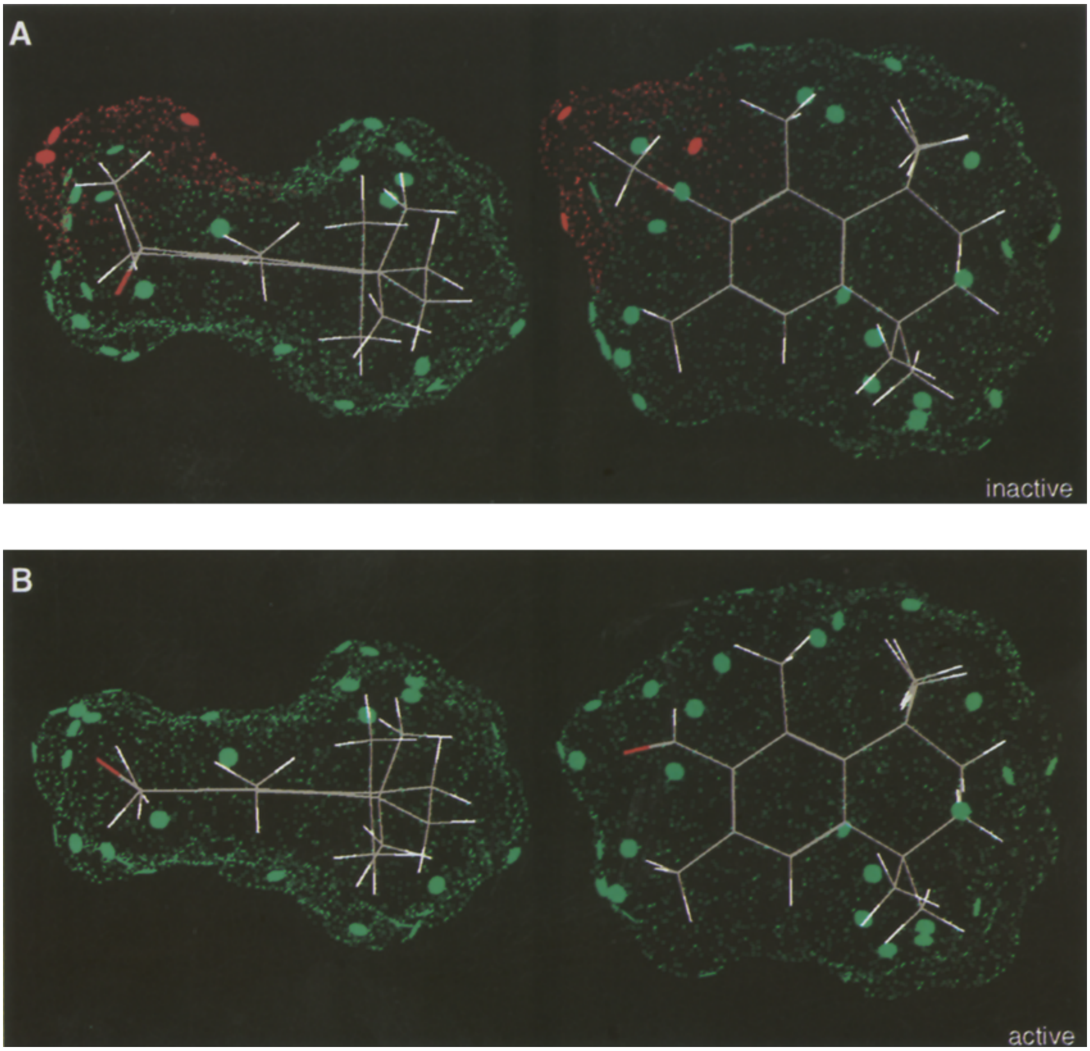


Fig. 7. Remediation of unfavorable steric interaction. (A) Non-musk with protruding methyl group; (B) corresponding musk analog with methyl group removed.

TABLE 4
EFFECTS OF AUTOMATIC CONFORMATIONAL SELECTION AND ADAPTIVE ALIGNMENT ON PREDICTIVE PERFORMANCE[a]

	Performance (%)	
	Single conformers	Multiple conformers
Without adaptive alignment	75 (4.3)	81 (3.9)
With adaptive alignment	77 (4.2)	91 (2.8)

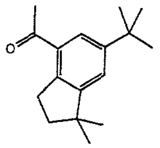
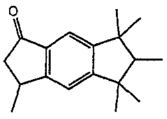
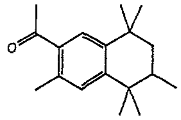
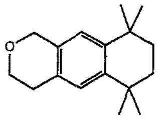
[a] Standard errors are shown in parentheses.

Figure 5 schematically illustrates the primary requirements for musk activity embodied in the learned model. Molecules must have a hydrogen-bond acceptor with the appropriate geometry (positions 1 or 2) and the appropriate surface shape in regions A, B and C. The actual learned model was sensitive to the location of about 50 specific surface patches in the A, B and C regions. This model is consistent with other models of musk odor activity [5,6,13–18], but it was learned exclusively from a general surface-based representation of shape. The learning algorithm automatically selected and weighted important shape characteristics, selected specific stereoisomers, and chose conformations and orientations of the molecules.

Interclass predictive performance

The surface-based molecular representation facilitates generalization across chemical structural classes, an ability crucial for accurate extrapolation to novel active chemotypes. We conducted a series of experiments in which all molecules of a given structural class were withheld during training and then evaluated during testing. This simulates a situation in which one wishes to apply a learned model to guide the synthesis of a new class of compounds or to discover new chemotypes from a database search. Table 5 shows four structurally different classes. Class 1 has a substantially different arrangement of hydrophobic bulk than the other classes, and classes 2 and 4 have molecules with different hydrogen-bonding geometries. Class 3 has complementary characteristics. Without adaptive alignment, the cross-class predictive performance ranged from 57 to 85%. With adaptive alignment, the performance ranged from 71 to 100%; the error rate dropped by more than half. In all cases the performance improved substantially by using adaptive

TABLE 5
PREDICTIVE ACCURACY OF MUSK MODEL ACROSS STRUCTURAL CLASSES[a]

Structural class	(1) 4-substituted dihydroindanes	(2) 1-indanones	(3) 6-substituted tetrahydronapthalenes	(4) benzopyrans
Representative molecule				
Number of molecules	13	20	27	14
% correct (fixed alignment)	85 (9.9)	75 (9.3)	74 (8.4)	57 (13.2)
% correct (adaptive alignment)	100 (0.0)	90 (6.5)	85 (6.8)	71 (12.1)

[a] Standard errors are shown in parentheses. Class 1 = 82–94; class 2 = 41–56, 74, 96–99; class 3 = 1–27; class 4 = 59–73.

alignment. The performance on class 4 was least good; we hypothesize that this is related to the different geometry of the ether component of these molecules with respect to the carbonyl groups of the other molecules. However, the *ranking* of the withheld molecules was quite good. Only two of seven inactive molecules were assigned higher predicted activities than any of the seven active molecules.

Molecular design

The method's ability to provide detailed guidance in molecular design is demonstrated in Figs. 6 and 7. To visualize a molecule interacting with a Compass neural network, one can analyze the function of molecular features that the neural network has constructed. For each feature, which corresponds to some part of the molecular surface, one can assess its contribution to the activity of the molecule: (i) a favorable contact (positive contribution); (ii) falls short of contact (zero contribution); or (iii) an unfavorable contact (negative contribution). Figure 6 displays four molecules in their predicted poses as chosen by a model trained on the remaining 98 molecules. Each molecule is displayed along with its Connolly surface [30]. The relative musk odor strength of these four hold-out molecules is known [18]. The patches on each surface correspond to the features selected by the model during training. Each patch is located at the nearest point on the molecule's surface to the feature's reference point. The surface has a favorable steric interaction if it has a green patch at that location. Yellow patches indicate areas that should be increased in size, and red patches indicate areas whose size should be decreased. Figure 6A displays a correctly predicted inactive molecule, and the yellow patches suggest that activity could be increased by adding bulk near the arrow (corresponding to area A in Fig. 5). Figure 6B shows the molecule resulting from the addition of a methyl group at this point, correctly predicted to have musk odor. From this molecule, which has only moderate musk odor intensity, the indicated region (corresponding to area B of Fig. 5) is predicted to benefit from additional bulk. Either adding a methyl group to the aromatic ring, shown in Fig. 6C, or changing the methyl group added to the molecule in Fig. 6A to an ethyl group, Fig. 6D, achieves this result. Both the molecules in Figs. 6C and D have stronger musk odor than the molecule in Fig. 6B, as predicted.

Figure 7 shows the application of another model, constructed by withholding the pair of molecules shown. In Fig. 7A, the red patches suggest an unfavorable interaction. This can be directly remedied by removal of the corresponding methyl group. The result is a correctly predicted molecule with strong musk odor, shown in Fig. 7B. Another approach is removal of the methyl substituent on the aromatic ring that is responsible for the ketone's unfavorable orientation. This results in a molecule of medium musk strength (not shown). We observed several other examples of correctly predicted improvements, suggested by the model, on molecules in this data set from different structural classes.

Previous studies of musk odor have analyzed similar molecules using atom-based approaches [5,6]. In cross-validation studies, they have shown predictive accuracy ranging from 90% (standard error 6.7) to 93% (standard error 6.4). However, these studies have not reported predictive results across chemical classes or employed molecular properties that could easily be interpreted to guide the design of new compounds. Instead, they relied on global molecular properties or features tied to the underlying chemical scaffolding of molecules.

Figure 8 contains a superposition of molecular poses for a single molecule. Figure 8A shows

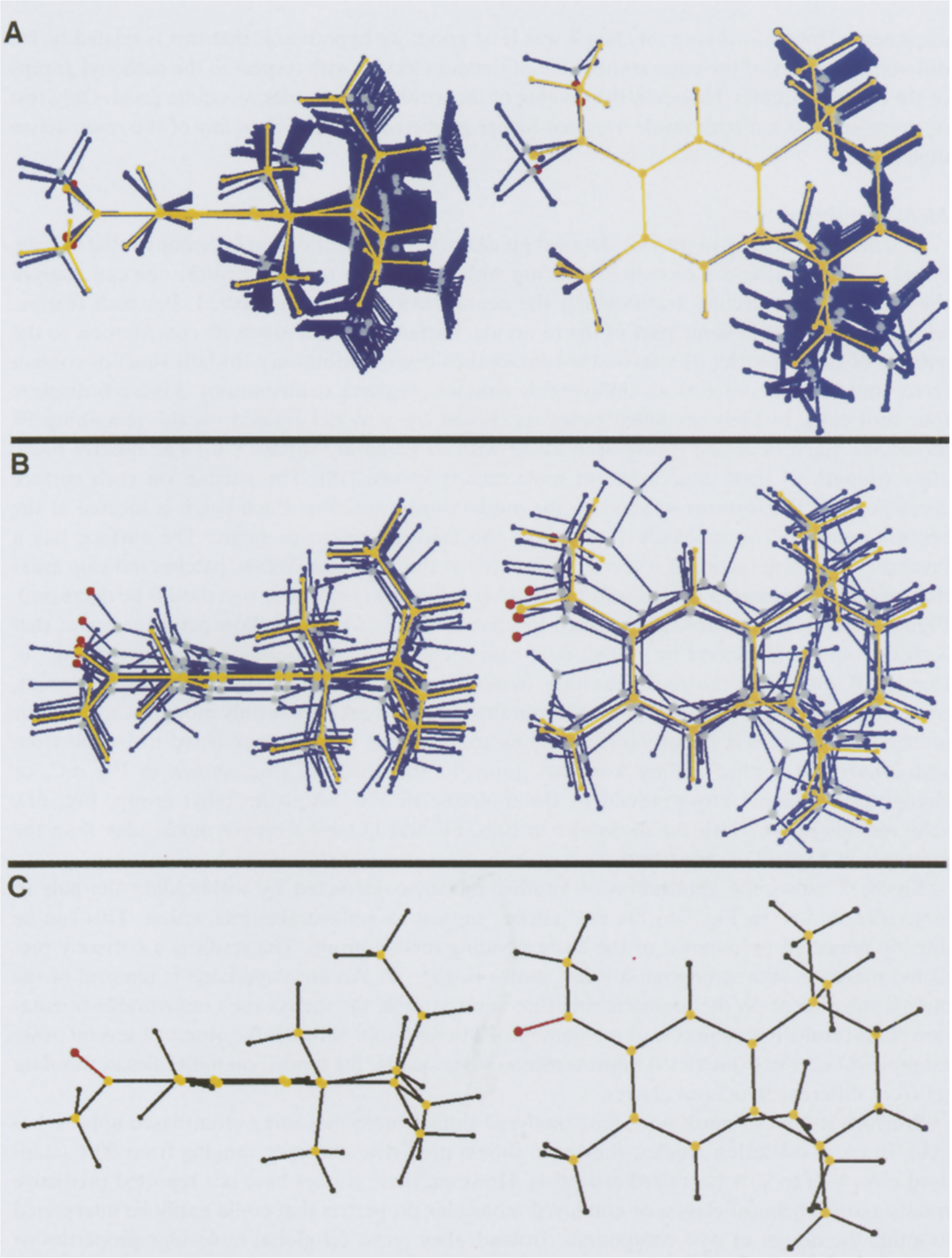


Fig. 8. Adaptive selection of molecular pose. (A) Superposition of conformers for an active molecule; (B) superposition of orientations of the selected conformer; (C) final adaptively selected pose.

its conformations, the shapes of which have substantial deviations. The yellow conformer corresponds to the best one as chosen by a trained model. Figure 8B shows this conformer in accessible orientations around its initial one and its final orientation in yellow. Figure 8C shows the molecule's final chosen pose. This adaptive approach to posing molecules is a major departure from previous methods. Methods that are sensitive to subtle shape differences among molecules must measure molecular properties that vary with pose (e.g., interatomic distances or occupancy of binding sites).

Previous methods assume that the correct poses of molecules can be selected before a predictive model is constructed [3–7,10,11]. Models constructed from fixed poses may not make accurate predictions, since new molecules must be placed in appropriate poses based on intuition or ad hoc procedures. Such procedures may behave poorly, especially on novel chemical structural classes. Our approach, in contrast, applies the constructed model to guide the generation of the correct poses, so that molecules are aligned along those surface regions that are most predictive of activity differences. In a separate study, we demonstrated performance superior to two fixed-pose techniques in predicting the binding affinity of a series of endogenous steroids to transport proteins [31].

CONCLUSIONS

In drug discovery efforts that are intractable from a biophysical perspective, techniques such as ours can provide an important source of data. Automatic conformation selection coupled with adaptive alignment removes a major obstacle to making accurate predictions, and the resulting models are able to resolve the effects of subtle surface changes. Interpretation of models through three-dimensional visualization can guide structural changes of molecules to enhance biological activity. The method's high predictivity, even across structural classes, should support the discovery of novel active chemotypes, a crucial capability for advancing drug design.

Our ongoing work focuses on two primary areas. The first is in broadening the set of features that Compass considers, for example, hydrogen bonding [31], surface polarizability, and formally charged interactions. The second is in developing more automated ways of generating rough initial alignments. For this work, a very simple pharmacophoric alignment was sufficiently good to allow adaptive pose generation to produce a predictive model. For other, more complex problems, we are developing algorithms with less human bias.

We have applied Compass in Arris's drug discovery projects. In one application, it helped elucidate the mode of action and basis for selectivity of a series of inhibitors targeting tryptase. This suggested a structural modification to a potent lead compound that retained nanomolar potency and showed improved selectivity.

ACKNOWLEDGEMENTS

We thank G. Whitesides for insight into ligand–receptor interactions, M. Ross for corporate support and for guidance, Kimberle Koile for programming support, and J. Park, B. Katz and J. Burns for data set construction and validation. The early stages of this work were funded by Firmenich SA (Geneva, Switzerland), and we are grateful for their support and their insight into musk odor chemistry.

REFERENCES

1 Appelt, K., Bacquet, R.J., Bartlett, C.A. Booth, C.L.J., Freer, S.T., Fuhry, M.A.M., Gehring, M.R., Herrmann, S.M., Howland, E.F., Janson, C.A., Jones, T.A., Kan, C.-C., Kathadekar, V., Lewis, K.K., Marzoni, G.P., Matthews, D.A., Mohr, C., Moomaw, E.W., Morse, C.A., Oatley, S.J., Ogden, R.C., Reddy, M.R., Reich, S.H., Schoettlin, W.S., Smith, W.W., Varney, M.D., Villafranca, J.E., Ward, R.W., Webber, S., Webber, S.E., Welsh, K.M. and White, J., J. Med. Chem., 34 (1991) 1925.

2 Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M., Science, 259 (1993) 1445.

3 Sheridan, R.P., Nilakantan, R., Dixon, J.S. and Venkataraghavan, R., J. Med. Chem., 29 (1986) 899.

4 Hopfinger, A.J., J. Am. Chem. Soc., 102 (1980) 7196.

5 Narvaez, J.N., Lavine, B.K. and Jurs, P.C., Chem. Senses, 11 (1986) 145.

6 Yoshii, F., Liu, Q., Hirono, S. and Moriguchi., I., Chem. Senses, 16 (1991) 319.

7 Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoehler, R.A. and Dunn, D.A., In Olsen, E.C. and Christoffersen, R.C. (Eds.) Computer-assisted Drug Design, American Chemical Society, Washington, DC, 1979, p. 57.

8 a. Hypotheses in Catalyst., BioCAD Corporation, Mountain View, CA, 1992.
b. Conformational Analysis in Catalyst., BioCAD Corporation, Mountain View, CA, 1992.

9 Ghose, A.K. and Crippen, G.M., J. Med. Chem., 28 (1985) 333.

10 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.

11 Good, A.G., So, S. and Richards, W.G., J. Med. Chem., 36 (1993) 433.

12 Buck, L. and Axel, R., Cell, 65 (1991) 175.

13 Bersuker, I.B., Dimoglo, A.S., Gorbachov, M.Yu., Vlad, P.F. and Pesaro, M., New J. Chem., 15 (1991) 307.

14 Brugger, W.E. and Jurs, P.C., J. Agr. Food Chem., 25 (1977) 1158.

15 Ham, C.L. and Jurs, P.C., Chem. Senses, 10 (1985) 491.

16 Chastrette, M., Zakarya, D. and Elmouaffek, A., Eur. J. Med. Chem., 21 (1986) 505.

17 Chastrette, M. and De Saint Laumer, J.-Y., Eur. J. Med. Chem., 26 (1991) 829.

18 Fehr, C., Galindo, J., Haubrichs, R. and Perret, R., Helv. Chim. Acta, 72 (1989) 1537.

19 Rumelhart, D.E., Hinton, G.E. and Williams, R.J., In Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (Eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations, MIT Press/Bradford, Cambridge, MA, 1986, p. 318.

20 Amoore, J.E., In Kugelmass, I.N. (Ed.) Molecular Basis of Odor, Ch.C. Thomas, Springfield, IL, 1970, p. 456.

21 Ohloff, G., Experientia, 42 (1986) 271.

22 Beets., M.G.J., Structure-Activity Relationships in Human Chemoreception, Applied Science Publishers, London, 1978.

23 Theimer, E.T. and Davies, J.T., J. Agr. Food Chem., 15 (1967) 6.

24 Chastrette, M., De Saint Laumer, J.-Y. and Sauvegrain, P., Chem. Senses, 16 (1991) 81.

25 Chang, G., Guida, W.C. and Still, W.C., J. Am. Chem. Soc., 111 (1989) 4379.

26 Mohamadi, F., Richards, N.G.J., Guida, W.C., Liskamp, R.M.J., Lipton, M.A., Caufield, C.E., Change, G., Hendrickson, T. and Still, W.C., J. Comput. Chem., 11 (1990) 440.

27 Weiner, S.J., Kollman, P.A., Nguyen, D. and Case, D.A., J. Comput. Chem., 1 (1986) 230.

28 Chastrette, M., In Schild, D. (Ed.) Chemosensory Information Processing, NATO ASI Series, Vol. H39, Springer, Berlin, 1990, p. 97.

29 Murray-Rust, P. and Glusker, J.P., J. Am. Chem. Soc., 106 (1984) 1018.

30 Connolly, M.J., J. Appl. Crystallogr., 16 (1983) 548.

31 Jain, A.N., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.