Simple but Highly Effective Three-Dimensional Chemical-Feature-Based Pharmacophore Model for Diketo Acid Derivatives as Hepatitis C Virus RNA-Dependent RNA Polymerase Inhibitors

Roberto Di Santo,*,† Maurizio Fermeglia,‡ Marco Ferrone,‡ Maria Silvia Paneni,‡ Roberta Costi,† Marino Artico,† Alessandra Roux,† Mirko Gabriele,† Keith D. Tardif,§ Aleem Siddiqui,|| and Sabrina Pricl*,‡

"Istituto Pasteur-Fondazione Cenci Bolognetti"—Dipartimento di Studi Farmaceutici, University of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Roma, Molecular Simulation Engineering (MOSE), Department of Chemical Engineering, University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy, Department of Microbiology and Program in Molecular Biology, Health Sciences Center, University of Colorado, Denver, Colorado 80262, and Department of Medicine, Moores UCSD Cancer Center, University of California, San Diego, 3855 Health Sciences Drive 0803, La Jolla, California 92093

Received May 10, 2005

A molecular modeling strategy using aryl diketo acid (ADK) derivatives recently reported in the literature as hepatitis C virus (HCV) polymerase inhibitors was designed. A 3D chemical-feature-based pharmacophore model was developed using Catalyst software, which produced 10 pharmacophore hypotheses. The top-ranked one (Hypo 1), characterized by a high correlation coefficient ($r = 0.965$), consisted of two hydrogen bond acceptors, one negative ionizable moiety, and two hydrophobic aromatics. This model was used to predict the anti-RNA-dependent RNA polymerase (anti-RdRp) activity of 6-(1-arylmethylpyrrol-2-yl)-1,4-dioxo-5-hexenoic acids and other ADK derivatives previously synthesized in our laboratories as HIV-1 integrase inhibitors. Furthermore, the experimental IC_{50} values of 9 compounds, tested in vitro against recombinant HCV polymerase, were compared with the corresponding values predicted using Hypo1. A good agreement between experimental and simulated data was obtained. The results demonstrate that the hypothesis derived in this study can be considered to be a useful tool in designing new leads based on ADK scaffolds as HCV RdRp inhibitors.

Introduction

The *hepatitis C virus* (HCV) is an RNA virus with a genome size of 9.4 kb. Since the advent of serological assays for HCV in 1990, it has been shown to be the major etiological agent for post-transfusion and sporadic non-A, non-B hepatitis worldwide.[1] In the USA alone, there are 175 000 cases of newly documented hepatitis C per year, and the number of HCV carriers has been estimated to be 4 million. Further, it is estimated that 3% of the world's population, or about 170 million people, are seropositive for HCV.[2]

Unlike hepatitis B, which is associated with chronicity in approximately 5% of adult infections, more than 80% of HCV-infected individuals develop chronic hepatitis, with either normal or higher levels of alanine aminotransferases (ALTs). Chronic hepatitis C, whether symptomatic or asymptomatic, can lead to cirrhosis and end-stage liver disease in 20−30% of patients, and among these, 1−4% may develop hepatocellular carcinoma.[3−6] Because of the spectrum of liver diseases, since its identification in 1989, HCV infection has been viewed as a serious threat to human health worldwide. The currently approved therapeutic protocol includes a combination of pegylated interferon and ribavirin.[7] However, these regimens have limited efficacy (10−40% of patients) and significant side effects,[8] causing up to 20% of patients to discontinue the therapy.[9] As a result, there is an urgent need to develop safe and effective antiviral agents.

The HCV belongs to the family Flaviviridae and is classified under the genus *Hepacivirus*.[10,11] The viral genome is a positive sense single-stranded RNA molecule encoding a unique open reading frame which translates into a polyprotein of about 3000 amino acids. The RNA genome is flanked by untranslated regions at both its 5′ and 3′ ends. Apart from differences in length, HCV genotypes show nucleotide heterogeneity of around 30% in their genomes, and a comprehensive analysis of these sequences has revealed the existence of at least six genotypes and more than 30 subtypes throughout the world. The importance of the genomic heterogeneity lies in the fact that some genotypes appear to be associated with more severe pathologies and are more refractory to treatment by current interferon-based therapies.

The NS5B gene product is the key component in the replication of the HCV. It was first recognized as the virally encoded RNA-dependent RNA polymerase (RdRp) by the GDD motif at residues 317−319 in the polymerase domain, which is totally conserved not only between different strains of HCV but also among all flaviviruses and all other RNA viruses.[12]

As evidenced by X-ray analysis, the enzyme has the typical right-ended fingers−palm−thumb domains of the polymerases but, in the case of HCV, extensions of the fingers and thumb lead to a more fully enclosed active site tunnel (Figure 1).[13−15]

* To whom correspondence should be addressed. Phone & fax: +39-6-49913150. E-mail: roberto.disanto@uniroma1.it (R.D.S.). Phone: +39-040-5583750. Fax: +39-040-569823. E-mail: sabrina.pricl@dicamp.units.it (S.P.).
† University of Rome "La Sapienza".
‡ University of Trieste.
§ University of Colorado.
|| University of California, San Diego.

10.1021/jm0504454 CCC: $30.25
Published on Web 09/09/2005

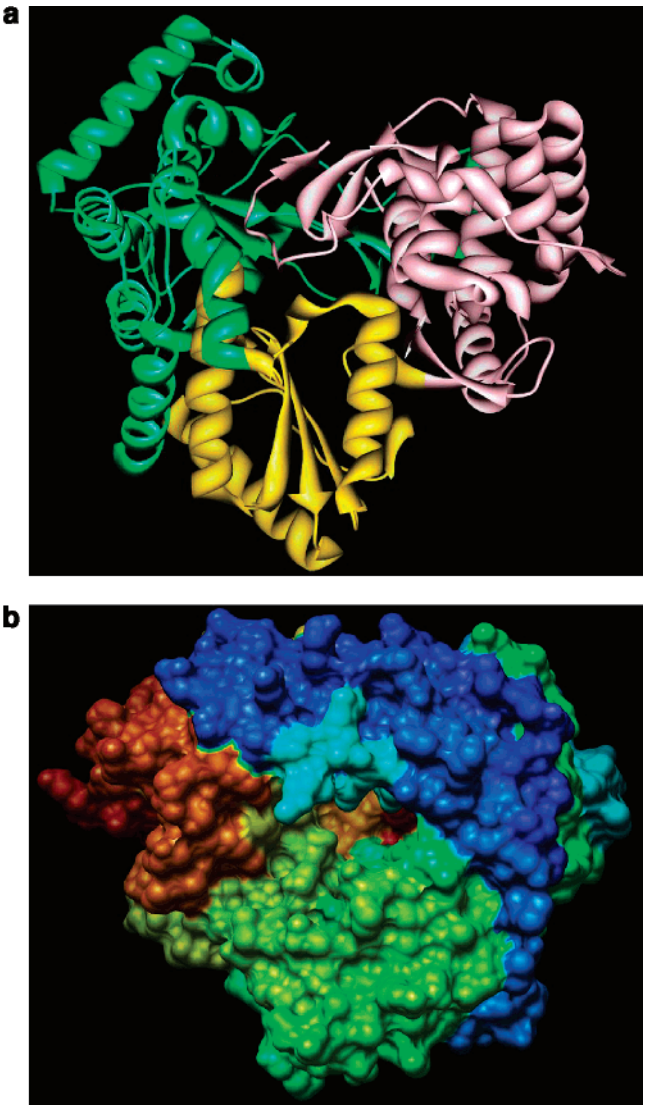


Figure 1. (a) Ribbon representation of HCV RNA-dependent RNA polymerase (RdRp) with domains colored as follows: thumb (pink, residues 371−563), palm (yellow, residues 188−227 and 287−370), and fingers (green, residues 1−187 and 228−286). (b) Molecular surface of HCV RdRp showing the active site tunnel (middle portion).

Similar to other positive strand RNA viruses, HCV NS5B functions both as a catalytic RdRp and as a cooperative RNA-binding protein that is required for viral RNA synthesis. Several lines of evidence suggest that like the kin of the Flaviviridae family, such as *Dengue virus* and *Japanese encephalitis virus*, the nonstructural proteins of HCV (NS3 to NS5B) assemble into membrane-associated ribonucleoprotein complexes (RNP), which are competent for authentic RNA replication. By virtue of its nature and mechanism of action, HCV RdRp is a potential target for antiviral drugs, since host cells probably have no analogous processes of RNA-dependent RNA synthesis and since the enzyme might recognize a specific template and primer of HCV RNA.

Accordingly, in the past decade, the development of inhibitors targeting the HCV polymerase activity has attracted the attention of investigators worldwide. In this respect, it should be noted that most of the nucleoside analogues that inhibit other viral polymerases are inactive against HCV polymerase; for instance, it is insensitive to most of the known inhibitors of DNA-dependent DNA polymerases or reverse transcriptases.[16−18]

Interestingly, the HCV polymerase activity can be inhibited by some transition metal ions such as Ni^{2+} and Zn^{2+} in the low micromolar concentration range.[18,19] Further, cerulenin and gliotoxin, originally identified as inhibitors for other viral RNA polymerases such as the poliovirus three-dimensional (3D) polymerase, display moderate inhibitory activity against HCV polymerase.[16,20]

Specific inhibitors for HCV polymerase were, however, not identified until recently. These include nucleoside analogues[21] and various non-nucleosides from different chemical classes.[22] Belonging to the latter group, over 200 compounds including alkyl-, phenyl-, pyrrole-, and thiophene-substituted diketo acids (ADKs) were evaluated in the HCV NS5B polymerase assay by Altamura et al.[23] Of these, several compounds demonstrated low micromolar IC_{50} values, and some of them also inhibit *hepatitis B virus* polymerase and HIV reverse transcriptase activities. As proposed by the authors, these molecules may interfere with the binding of phosphoryl groups of the nucleotide substrates at the active site of the viral polymerase and, therefore, inhibit the formation of phosphodiester bonds catalyzed by the enzyme. This inhibition could be explained by the ability of ADKs to interact with the catalytic metal ions found in the enzyme active site.

Recently, we were interested in the synthesis and biological evaluation of new inhibitors of the HIV-1 integrase (IN) enzyme.[24−29] These efforts culminated in the identification of 6-(1-arylmethylpyrrol-2-yl)-1,4-dioxo-5-hexenoic[27,28] and 3-quinolonyl-2,4-dioxobutanoic[29] acid derivatives belonging to the ADK class, which showed significant activities in the micromolar range against recombinant integrase.

On the basis of the above-mentioned reports concerning the anti-NS5B activity of the ADKs designed by Altamura et al.,[23] we designed a molecular modeling strategy using ADK derivatives as HCV polymerase inhibitors (Charts 1 and 2) with the aim of developing a 3D chemical-feature-based pharmacophore model. This model was then used to predict the anti-RdRp activity of 6-(1-arylmethylpyrrol-2-yl)-1,4-dioxo-5-hexenoic acid, 3-quinolonyl-2,4-dioxobutanoic acid, and other ADK derivatives previously designed as anti-HIV-1 agents targeted to IN and synthesized in our laboratories (Chart 3).[27−29]

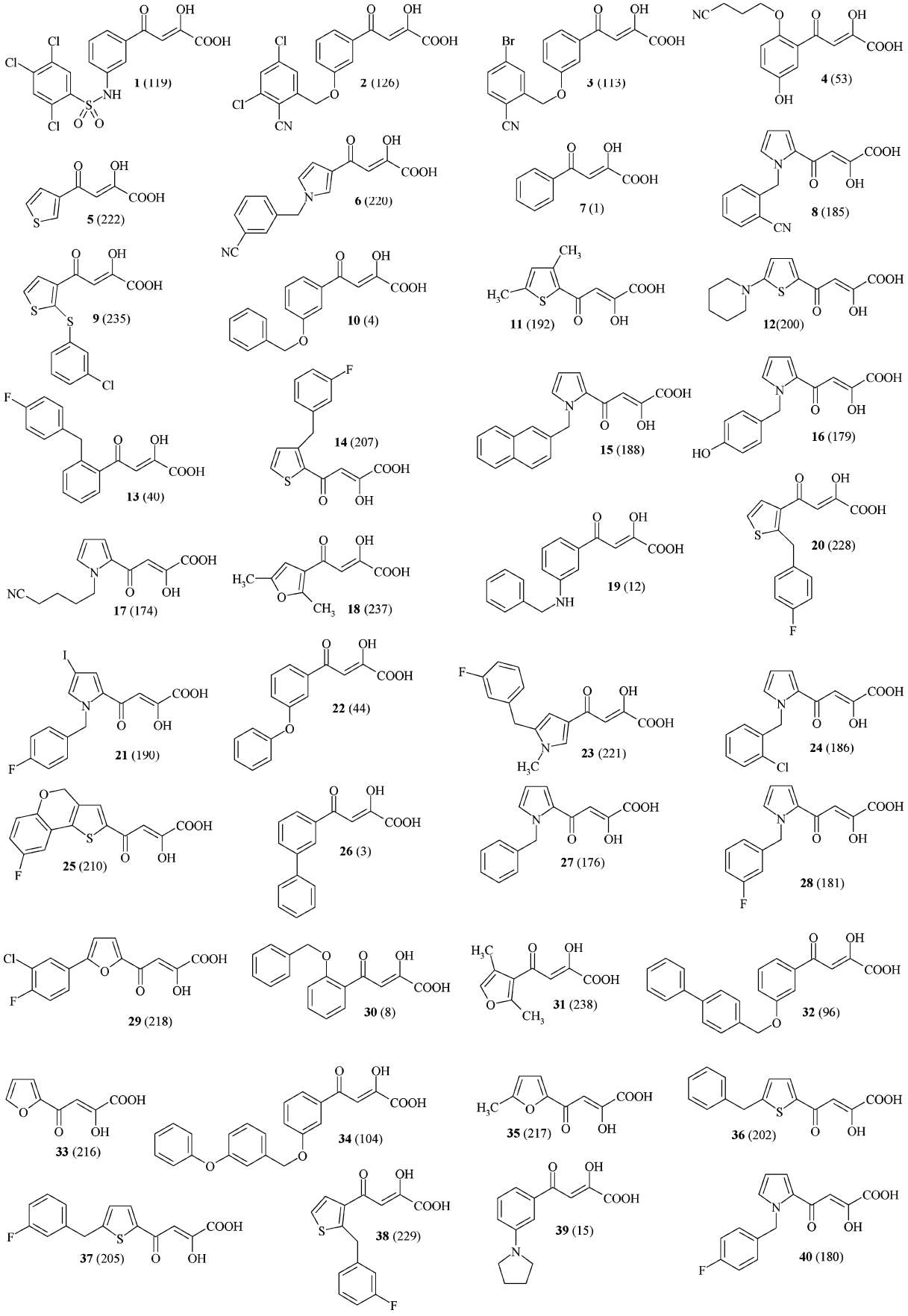
Furthermore, this theoretical study was validated by the comparison of predicted activities with some experimental data obtained by testing our derivatives against recombinant HCV polymerase in enzyme assays.

Results and Discussion

During a HypoGen run, Catalyst distinguishes between alternatives of thousands of models by applying cost analysis and by searching for the simplest set of chemical functions that correlate best with the observed activity.[30]

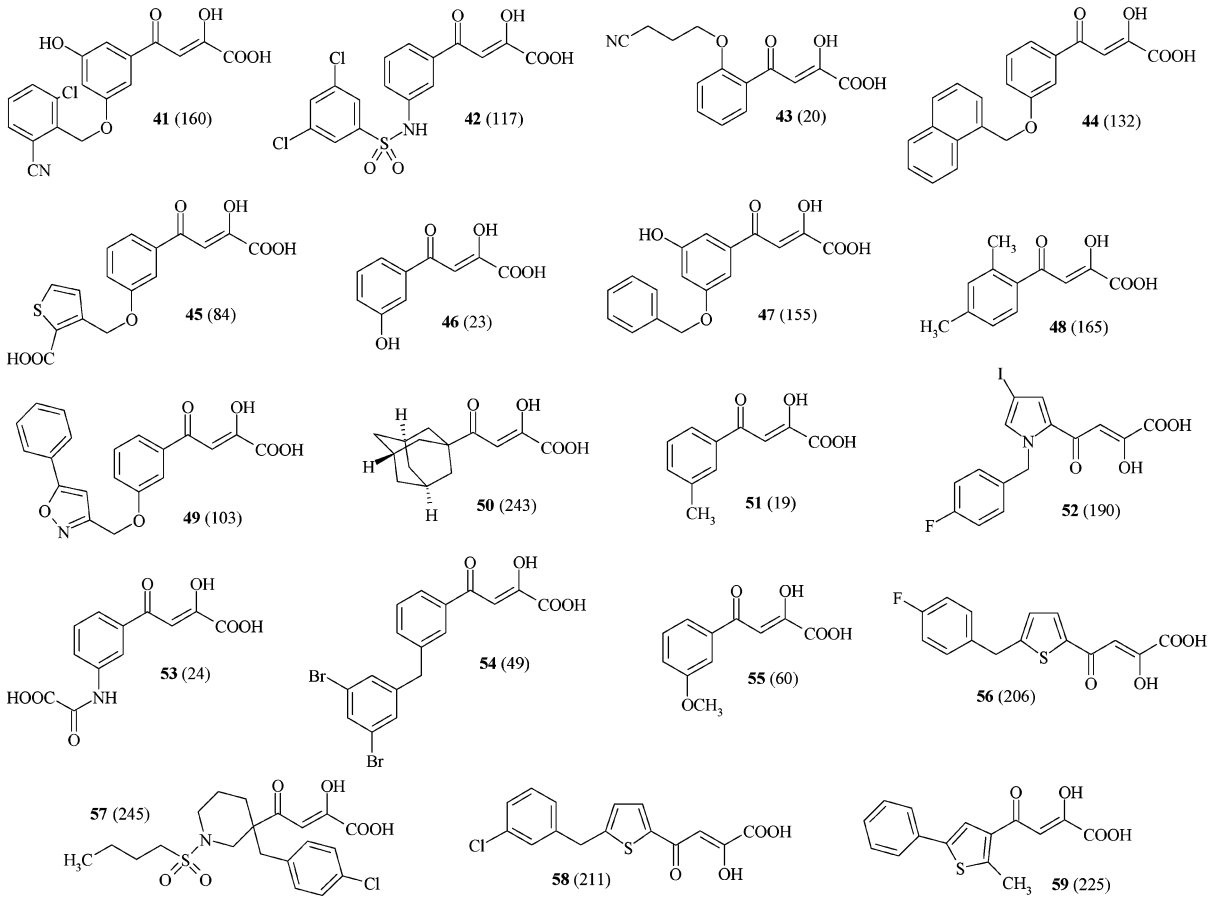
At the end of the run, Catalyst produced 10 hypotheses. The top-ranked one, Outhypo-38969.22 (Hypo1 for simplicity), is shown in Figure 2. It consists of two hydrogen bond acceptors, one negative ionizable group,

Chart 1. 2D Chemical Structures of the 40 Molecules Forming the Training Set Used To Obtain HypoGen Pharmacophore Hypotheses[a]



[a] The numbers in parentheses refer to the literature numbering.[23]

Chart 2. 2D Chemical Structures of the 29 Molecules Forming the Test Set Used to Validate HypoGen Pharmacophore Hypotheses[a]



[a] The numbers in parentheses refer to the literature numbering.[23]

and two hydrophobic aromatics. Hypo1 is the best pharmacophore hypothesis resulting from this study, as it is characterized by the highest cost difference, lower error cost, lowest root mean square (rms) divergence, and best correlation coefficient (Table 1). In detail, the null cost of the 10 top-scored hypotheses was equal to 194.0110, the fixed cost value was 118.7910, and the configuration cost was 11.4569. As the total cost of Hypo1 was equal to 128.041, the large difference between null and total hypothesis cost, Δcost = 65.970, coupled with a high correlation coefficient, $r = 0.965$, and a reasonable root mean square deviation, rms = 1.070, ensures that a true correlation will very likely be estimated by the model.

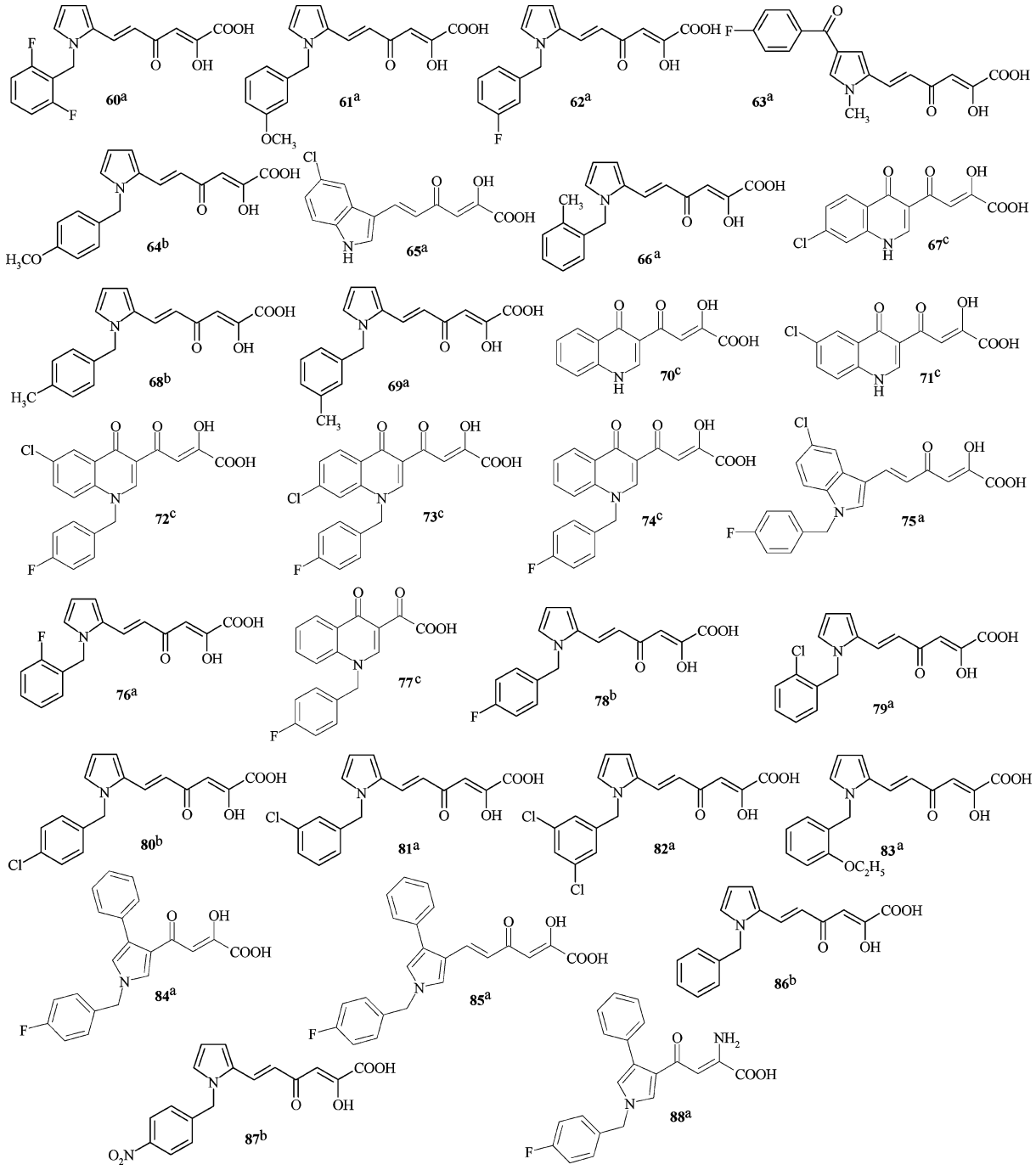
Besides this cost analysis, the most obvious method to validate Hypo1 is to check for its capacity to correctly predict the activity of the training set compounds. Table 2 lists the Hypo1-estimated activity values, together with the corresponding errors (i.e., the ratio between calculated and experimental activity). As we can see from this table, most of the IC_{50} values were predicted correctly and, in the worst cases, they were found to be within the same order of magnitude as the corresponding experimental evidences. On this hypothesis, all training set compounds mapped one hydrogen bond acceptor function, one hydrophobic aromatic function, and the negative ionizable function, while 90% of the compounds mapped the remaining hydrogen bond acceptor function and 70% mapped the second hydrophobic aromatic function.

Figure 3 depicts one of the most active conformations of compound **2** (126) in the training set mapped onto Hypo1. As we can see from this figure, compound **2** (126) fits all features of the pharmacophore model Hypo1 very well. Both phenyl rings of the molecule overlap with the hydrophobic aromatic (HpAr) features of Hypo1, whereas the bridge oxygen atom serves as a hydrogen bond acceptor (HBA). Another HBA point is located on the nitrogen atom of the cyano group in the ortho position of the dichlorobenzene ring. Finally, the negative ionizable (NI) feature is correctly disposed over the carboxylic group. Thus, the second most active compound, **2** (126), in the training set maps closely with the statistically most significant hypothesis, and the predicted activity of **2** (126) toward Hypo1 is reasonable (Table 2).

Further validation of this hypothesis is forthcoming from the fact that all compounds that constituted the most active group in the training set plus the next most active compounds in Table 2 mapped all five features of the hypothesis. Conversely, compounds that are estimated to have a lower activity did not map adequately to Hypo1, at least at one critical site, such as one HBA or one HpAr site.

Figure 4 shows, as an example, the mapping on Hypo1 of the training set compound **28** (181), for which the pair of HpAr features are placed over the pyrrole and phenyl rings, respectively, the NI feature is correctly located on the carboxylic group, and the carbonyl oxygen atom of the diketo acid moiety serves as the first HBA, whereas the second HBA mapping is missing.

Chart 3. 2D Chemical Structures of the Set of Diketo Acids for Which the Activity as HCV RdRp Inhibitors Was Predicted Using Hypo1



[a] Reference 28. [b] Reference 27. [c] Reference 29.

The validity and the predictive character of this five-feature hypothesis were further assessed by using the test set molecules listed in Chart 2. All members of the test set were predicted correctly, as expected, and the results are reported in Table 3. All test set compounds mapped one hydrogen bond acceptor function, one hydrophobic aromatic function, and the negative ionizable function, while 92% of the compounds mapped the remaining hydrogen bond acceptor function and 72% mapped the second hydrophobic aromatic function.

Figure 5 shows the mapping of the test set compounds **44** (132) and **54** (49) onto Hypo1. Both compounds map all the functional features of the best hypothesis with high scores. The two HpAr features overlap with the aromatic rings present in the molecules, the NI is always located on the COOH group, and the two HBA features are placed over the oxygen atoms of the diketo acid moiety. Interestingly, for compound **54** (49), the overlay of both HpAr features with the corresponding aromatic rings is not as neat as it is for the most active compound, **44** (132), in harmony with the lowest activity of this molecule.

In light of these results, Hypo1 was used to predict the activity of the set of recently synthesized diketo acid compounds listed in Chart 3.

Table 4 presents the comparison between predicted and experimentally determined activity values for this subset, while Table 5 lists the predicted activity values

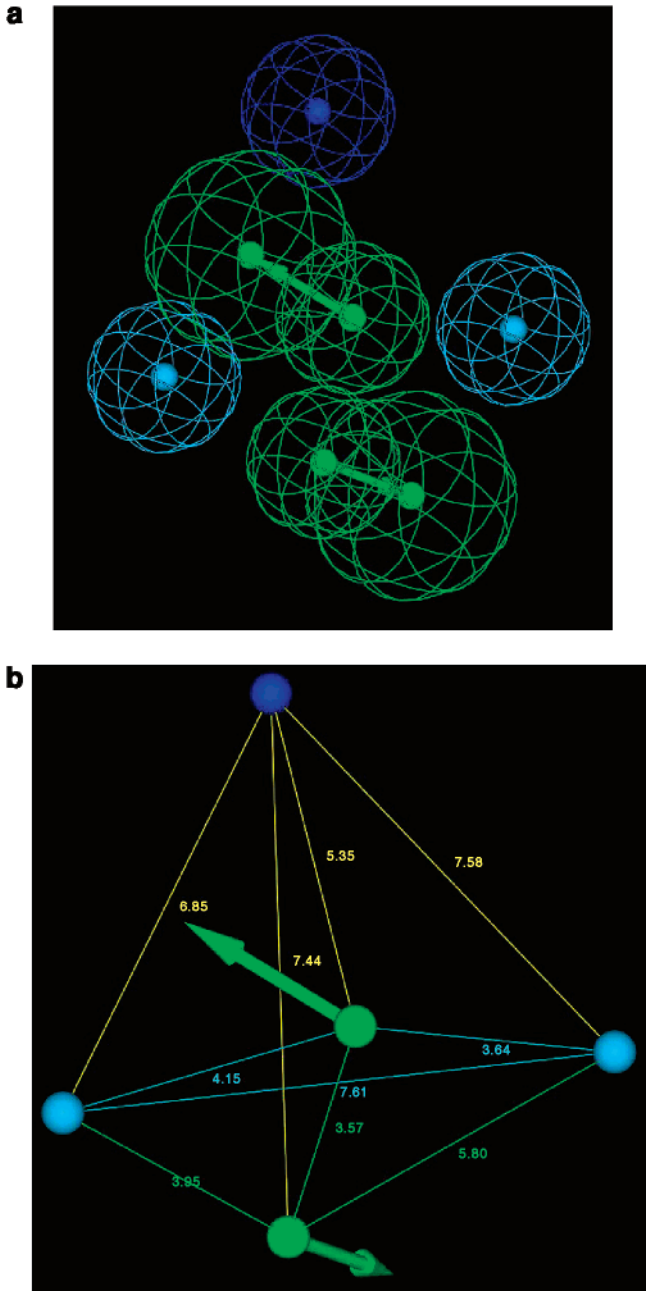


Figure 2. (a) Top-scoring HypoGen pharmacophore Hypo1. The hypothesis features are color-coded as follows: hydrophobic aromatic, light blue; negative ionizable group, dark blue; hydrogen bond acceptor, green. (b) Distances (Å) between all features in Hypo1.

for those compound for which the activity test is currently still in progress. From an inspection of Table 4, we readily see that the agreement between the experiment and simulation data is good, both qualitatively and quantitatively. There is a shift in the values, which is easily justified if we consider that the activity data for training and test sets and those for our diketo acids were obtained with different methodologies. In fact, for biochemical assays a GST tagged NS5B derivative was used.[38] This NS5B derivative has a C-terminal 27 amino acid truncation. Researchers typically use this NS5B derivative instead of the full-length NS5B, employed in the test performed by Altamura,[23] as it is more soluble and easier to isolate. Furthermore, this NS5B derivative has been shown to have the same characteristics as the full-length NS5B. However, we used derivative **8** (185), synthesized by Altamura et al.,[23] as a reference compound. In our test, **8** showed $IC_{50} = 61.5$ μM, while the activity reported in the literature was $IC_{50} = 6.3$ μM.[23] Therefore, we decided to calculate experimental IC_{50}'s as follows: (IC_{50} measured)/F, where $F =$

Table 1. Information of Statistical Significance and Predictive Power Presented in Cost Values Measured in Bits for the Top 10 Hypotheses as a Result of the Automated HypoGen Pharmacophore Generation Process[a]

hypothesis	total cost	Δcost	rms deviation	correlation (r)
1	128.041	65.970	1.070	0.965
2	129.667	64.344	1.333	0.955
3	130.504	63.507	1.383	0.901
4	132.627	61.384	1.386	0.903
5	132.966	61.045	1.397	0.929
6	133.105	60.906	1.412	0.900
7	133.205	60.806	1.417	0.887
8	133.234	60.777	1.400	0.922
9	134.790	59.221	1.423	0.878
10	135.284	58.727	1.433	0.859

[a] The null cost of the 10 top-scored hypotheses is 194.0110, the fixed cost value is 118.7910, and the configuration cost is equal to 11.4569.

Table 2. Experimental Biological Data[23] and Estimated IC_{50} Values of the Training Set Molecules Based on the Pharmacophore Model Hypo1

compd	no. in the literature[23]	experimental IC_{50} [μM]	estimated IC_{50} [μM]	error[a]
1	119	0.049	0.022	−2.2
2	126	0.056	0.096	+1.7
3	113	0.10	0.48	+4.8
4	53	0.11	0.53	+4.8
5	222	4.0	2.6	−1.5
6	220	4.6	8.1	+1.8
7	1	5.6	2.4	−2.3
8	185	6.3	2.1	−3.0
9	235	7.5	10.1	+1.3
10	4	8.0	4.2	−1.9
11	192	8.2	5.1	−1.6
12	200	8.2	11	+1.3
13	40	8.3	6.2	−1.3
14	207	9.9	13	+1.3
15	188	13	18	+1.4
16	179	13	18	+1.4
17	174	13	16	+1.2
18	237	14	11	−1.3
19	12	17	21	+1.2
20	228	18	13	−1.4
21	190	18	21	+1.2
22	44	19	25	+1.3
23	221	21	19	−1.1
24	186	24	31	+1.3
25	210	27	21	−1.3
26	3	28	24	−1.2
27	176	29	24	−1.2
28	181	30	25	−1.2
29	218	41	45	+1.1
30	8	44	47	+1.1
31	238	48	55	+1.1
32	96	50	45	−1.1
33	216	50	41	−1.2
34	104	50	52	+1.0
35	217	58	55	−1.1
36	202	68	61	−1.1
37	205	74	77	+1.0
38	229	80	90	+1.1
39	15	84	73	−1.2
40	180	93	100	+1.1

[a] Values in the error column represent the ratio of the estimated activity to the measured activity or its negative inverse if the ratio is less than one.

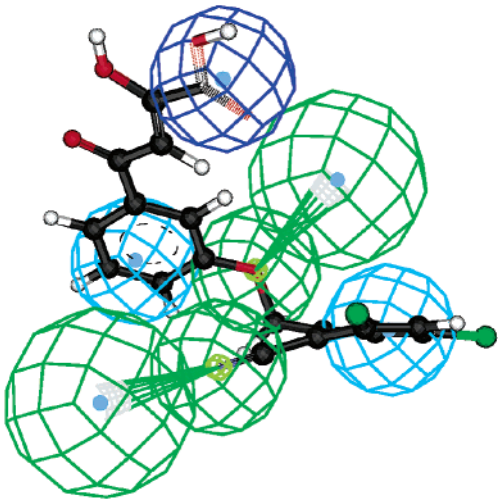


Figure 3. Mapping of the training set compound **2** (126) onto Hypo. Hypothesis features are color-coded as follows: hydrophobic aromatic, light blue; negative ionizable group, dark blue; hydrogen bond acceptor, green.

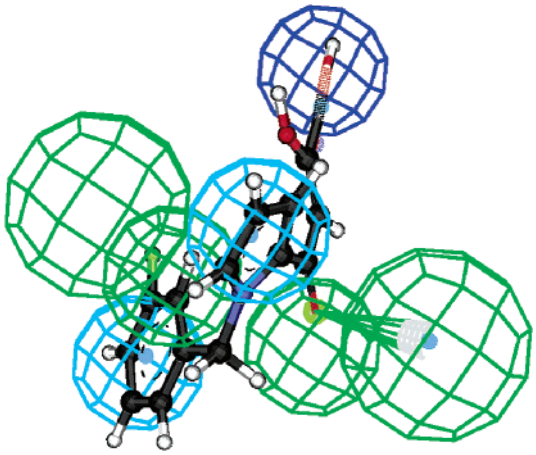


Figure 4. Training set compound **28** (181) mapped to Hypo1. Hypothesis features are color-coded as follows: hydrophobic aromatic, light blue; negative ionizable group, dark blue; hydrogen bond acceptor, green.

(IC_{50} measured for **8**)/(IC_{50} reported in the literature) (Table 4).

Parts a, b, and c of Figure 6 illustrate the mapping of compounds **60**, **69**, and **67**, respectively. The predicted most active compounds **60** and **69** map nicely over the five features of the hypothesis, overlapping the HpAr features over the aromatic rings and the NI feature onto the carboxylic group and with two oxygens of the diketo acid moiety serving as HBAs. In the case of compound **67**, the presence of the condensed ring system featuring only one aromatic group accounts for the missing second aromatic feature mapping and, hence, the lower activity predicted for this compound.

Conclusions

In the rational drug design process, it is common for the biological activity data of a set of compounds acting upon a particular protein to be known, while information on the three-dimensional structure of the protein active site is absent. A three-dimensional pharmacophore hypothesis that is consistent with known data should be useful and predictive in evaluating new compounds and directing further synthesis. A pharmacophore model postulates that there is an essential three-dimensional arrangement of functional groups that a molecule must possess to be recognized by the active site. It collects common features distributed in 3D space that are intended to represent groups in a molecule that participate in important interactions between drugs and their active sites. Hence, a pharmacophore model provides crucial information about how

Table 3. Experimental Biological Data[23] and Estimated IC_{50} Values of the Test Set Molecules Based on the Pharmacophore Model Hypo1

compd	no. in the literature	experimental IC_{50} [μM]	estimated IC_{50} [μM]	error[a]
41	160	0.1	0.22	+2.2
42	117	0.14	0.38	+2.7
43	20	0.38	0.15	−0.4
44	132	0.67	1.4	+2.1
45	84	0.95	1.2	+1.3
46	23	1.0	1.4	+1.4
47	155	1.4	0.67	−0.6
48	165	2	3.1	+1.6
49	103	6	4.1	−0.7
50	243	6.7	15.3	+2.0
51	19	12	16	+1.3
52	190	18	10	−0.6
53	24	19	25	+1.3
54	49	32	26	−0.8
55	60	40	55	+1.4
56	206	65	61	−0.9
57	245	81.4	77	−0.9
58	211	82	90	+1.1
59	225	167	144	−0.9

[a] Values in the error column represent the ratio of the estimated activity to the measured activity or its negative inverse if the ratio is less than one.

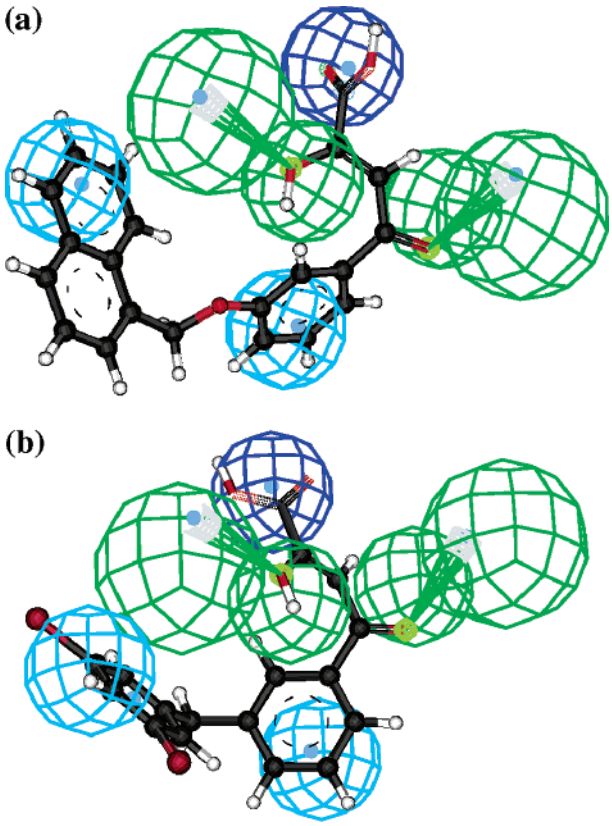


Figure 5. Test set compounds **44** (132) (a) and **54** (49) (b) mapped to Hypo1. Hypothesis features are color-coded as follows: hydrophobic aromatic, light blue; negative ionizable group, dark blue; hydrogen bond acceptor, green.

Table 4. Experimental Biological Data and Estimated IC_{50} Values of Our New DKAs Based on the Pharmacophore Model Hypo1[a]

compd	experimental IC_{50} [μM]	estimated IC_{50} [μM]	error[b]
60	7.4	0.98	−0.1
61	7.6	1.4	+0.2
62	18	4.0	+0.2
63	16	4.8	+0.3
64	16	5.9	−0.4
65	18	7.6	+0.4
66	18	8.5	−0.5
67	26	15	−0.6
68	28	16	+0.6

[a] Experimental IC_{50}'s calculated as follows: (IC_{50} measured)/F, where F = (IC_{50} measured for **8**)/(IC_{50} reported in the literature).[23] [b] Values in the error column represent the ratio of the estimated activity to the measured activity or its negative inverse if the ratio is less than one.

Table 5. Estimated IC_{50} Values of the Other DKAs Based on the Pharmacophore Model Hypo1

compd	estimated IC_{50} [μM]
69	2.8
70	3.5
71	3.8
72	5
73	5.3
74	5.3
75	6.3
76	6.7
77	7.5
78	8.3
79	9.8
80	11
81	11
82	11
83	11
84	12
85	13
86	13.3
87	18.2
88	19

well the common features of a subject molecule overlap with the hypothesis model. It also informs about the ability of molecules to adjust their conformations in order to fit an active site with energetically reasonable conformations. Such characterized 3D models convey important information in an intuitive manner.

Our present work illustrates how a set of various compounds may be treated statistically to uncover the molecular characteristics which are essential for high activity. These characteristics are expressed as common features disposed in 3D space and are collectively termed a hypothesis. Hypotheses obtained were applied to map the active or less active compounds. Important features, such as two hydrogen bond acceptors, a negative ionizable moiety, and two hydrophobic aromatics, of the surface-accessible models were found, which are the minimum components of a hypothesis for effective compounds.

In this work, we built three-dimensional pharmacophore hypotheses from a training set of 40 proprietary diketo acid compounds active as inhibitors of the HCV RNA-dependent RNA polymerase. We verified that the more active compounds map well onto all features of the best hypothesis, Hypo1. For less active compounds, their lack of binding affinity to the target protein can be primarily ascribed to their inability to achieve an energetically favorable conformation shared by the active compounds in order to fit the 3D common feature pharmacophore keys. Taken together, two hydrogen bond acceptors, a negative ionizable group, and two hydrophobic aromatics located on the molecules seem to be essential for high activity in inhibition of HCV RNA-dependent RNA polymerase.

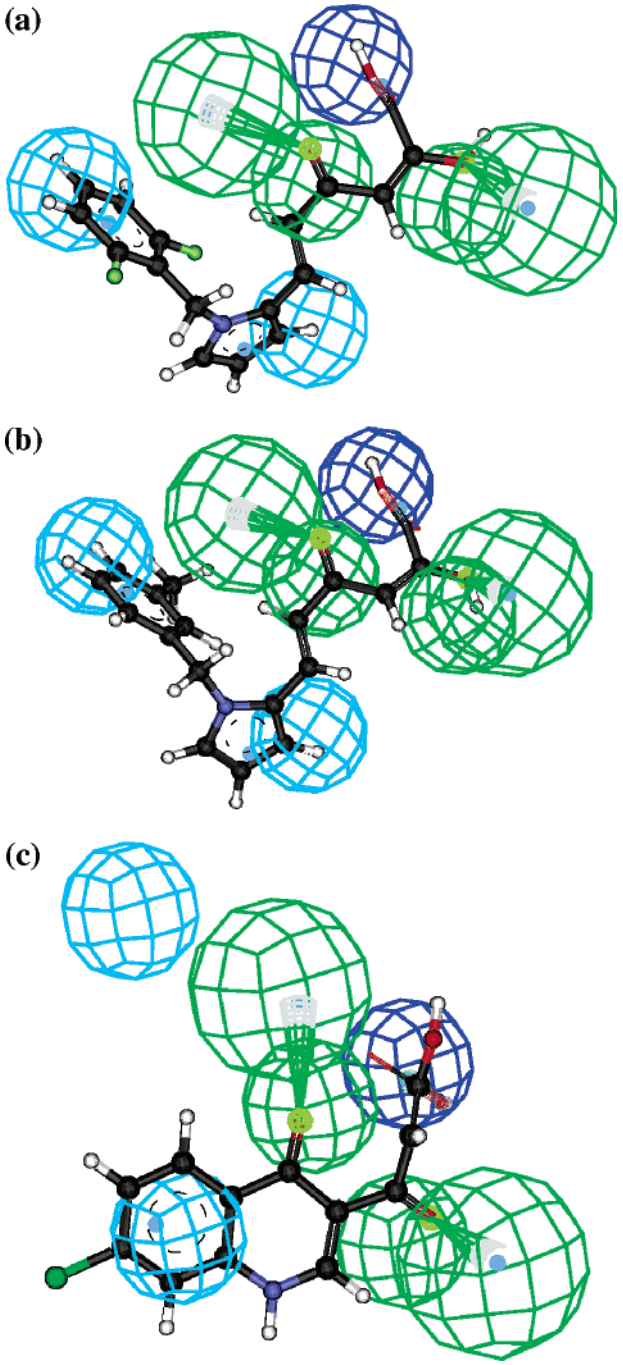


Figure 6. Mapping of three new diketo acid derivatives **60** (a), **69** (b), and **67** (c) to Hypo 1. Hypothesis features are color-coded as follows: hydrophobic aromatic, light blue; negative ionizable group, dark blue; hydrogen bond acceptor, green.

The same hypothesis has been applied to the prediction of the activity of recently synthesized diketo acid derivatives. For those compounds for which the activity was obtained concomitantly, the results obtained were both qualitatively and quantitatively in good agreement. Accordingly, the hypothesis derived in this study can be considered to be a useful tool in designing new leads

for hopefully more active compounds, and work is in progress in this respect.

Experimental Section

Computational Methodology. The study was performed using the Catalyst software package (version 4.9, Accelrys Inc., San Diego, CA) on a Silicon Graphics Octane computer, running the IRIX 6.5 operating system.

Catalyst is a relatively new approach that focuses on modeling the drug−receptor interaction from the point of view of the receptor, using information derived only from the drug.[30]

Molecular structures are treated as templates consisting of strategically positioned chemical functions that will bind effectively with complementary functions on receptors. Conformational flexibility is modeled by creating multiple conformers, which are treated simultaneously during the runs.

Chemical-feature-based pharmacophore hypotheses can be generated automatically using the HypoGen algorithm within Catalyst, provided that structure−activity relationship data of a well-balanced set of compounds is available. Two assumptions must be made about the data: (1) all compounds used in the training set have to bind to the same receptor in roughly the same fashion and (2) compounds having more binding interactions with the receptor are more active than those with fewer. A problem that must be addressed is that, since adverse steric interactions are not recognized during the automatic pharmacophore model generation, molecules that exhibit low activity due to a "forbidden" volume should not be present in the data set. After model generation, however, this information can be included in the model by adding exclusion spheres manually.

Training Set and Test Set Selection and Conformational Models. The correct choice of the compounds of the training set is a key issue in an automated pharmacophore generation process. In fact, according to Sutter et al.,[31] the constructed pharmacophore model can be as good as the information data input. Nevertheless, to achieve such a quality, some must-obey rules should be respected in a three-dimensional quantitative structure−activity relationship (QSAR) generation using Catalyst. Thus, for instance, the set must be widely populated (at least 16 items) by structurally diverse representatives covering an activity range of at least 4 orders of magnitude. The most active compounds should inevitably be included in the training set, and all biologically relevant data should be obtained by homogeneous procedures. Taking into account these requirements, we defined a training set of 40 compounds, presented in Chart 1, taken from a single reference,[23] and characterized by IC_{50} values ranging from 49 nM to 160 μM.

Molecules were edited using the Catalyst 2D/3D visualizer. Conformational models of the training set compounds were generated using a Monte Carlo-like algorithm together with poling. Poling is a method for promoting conformational variation that forces similar conformers away from each other. Conclusively, poling improves the coverage of the conformational space.[32−34]

The number of conformations needed to produce a good representation of a compound's conformational space obviously depends on its molecular structure and flexibility. This last feature was taken into account by considering each compound as a collection of conformers representing a different area of conformational space accessible to a molecule within a given energy range. For this purpose, Catalyst provides two types of conformational analysis: Fast and Best. In our case, the Best option was used, specifying 250 as the maximum number of conformers. Our models emphasized a conformational diversity within the constraint of a 20 kcal/mol energy threshold above the estimated global minimum based on use of the CHARMM force field.[32−35] All other parameters used were kept at their default settings. This should ensure an exhaustive characterization of conformational space.[36] The molecules associated with their conformational models were then submitted to Catalyst hypothesis generation.

Generation of Pharmacophore Hypotheses with HypoGen. A pharmacophore is a representation of generalized molecular features including 3D (hydrophobic groups, charged/ionizable moieties, and hydrogen bond donors/acceptors), 2D (substructures), and 1D (physical and biological data) aspects that are considered to be responsible for a desired biological activity. Thus, the next step in the generation of a pharmacophore model with HypoGen, after the selection of the training set, is the definition of the correct, feasible features, defined by chemical functions, locations, orientations in space, tolerance in locations, and weights.

Taking into account the chemical nature of the compounds considered in this work, the following five features were selected to form the essential information in the hypothesis generation process: hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), hydrophobic aromatic (HpAr), ring aromatic (Ar), and a negative ionizable group (NI). The NI was used in order to broaden the search for deprotonated or protonated carboxylic groups at a physiological pH. The HypoGen algorithm was then forced to select only five features due to molecular flexibility and molecular complexity. For molecules larger than dipeptides, Catalyst will often find five-feature hypotheses automatically, but for smaller molecules, three- or four-feature hypotheses might constitute the majority. Since hypotheses with a larger number of features are more likely to be stereospecific and generally more restrictive models, the minimum of total features was set to 5 in order to force Catalyst to search for five-feature hypotheses.

Finally, prior to a HypoGen run, the value for the uncertainty in biological data must be set. Indeed, Catalyst generates a chemical-feature-based model on the basis of the most active compounds. These compounds are determined by performing a simple calculation based on the activity and uncertainty. As a matter of fact, the activity of the most active compound is multiplied by the uncertainty (which is set equal to 3.0 by default in the software) to establish a comparison number, "A". The activity of the next most active compound is divided by the uncertainty, and this results in "B", which is then compared to A. If B is smaller than A, the compound is included in the most active set; if not, the procedure stops. Accordingly, the uncertainty is a crucial parameter. In our case, as no experimental information was available, an uncertainty of 1.2 was preferred over the default factor of 3.0. The chosen value of 1.2 is more convenient for diketo acid compounds because their activities barely span the required 4 orders of magnitude; this choice has been further confirmed by preliminary calculations and by other literature evidence.[37]

Validation of HypoGen Hypotheses. The first way to validate the output of a HypoGen run is to apply a cost analysis. During a hypothesis generation run, Catalyst considers and discards many thousands of models. It attempts to minimize a cost function consisting of three terms: the weight cost, the error cost, and the configuration cost. The weight cost is a value that increases in a Gaussian form as the feature weight in a model deviates from an idealized value of 2.0. The error cost is a major value that increases as the rms difference between estimated and measured activities increases. Finally, the configuration cost is a fixed cost that depends on the complexity of the hypothesis, is equal to the entropy of the hypothesis space, and should always be less than 17. The overall assumption used in tuning the cost function is based on Occam's razor, which states that, between otherwise equivalent alternatives, the simplest model is the best. Simplicity is defined using the minimum description length principle from information theory.

Besides providing a numerical score for each generated hypothesis, Catalyst provides three numbers to help the chemist assess the validity of a hypothesis. One is the cost of an ideal hypothesis, which is the lower bound of the cost of the simplest hypothesis that still fits the data perfectly. Another is the cost of the null hypothesis, which presumes that there is no statistically significant structure in the data and that the experimental activities are normally distributed around their mean. Generally, the greater the difference

between the two costs, the higher the probability for finding useful models. In terms of hypothesis significance, what really matters is the magnitude of the difference between the cost of any returned hypothesis and the cost of the null hypothesis. In general, if this difference is greater than 60 bits, there is an excellent chance that the model represents a true correlation. Since most returned hypotheses will be higher in cost than the fixed cost model, a difference between the fixed cost and the null cost of 70 bits or more will be necessary to achieve the 60 bits difference. If a returned hypothesis has a cost that differs from the null hypothesis by 40−60 bits, there is a high probability (a 75−90% chance) of it representing a true correlation in the data. Under these conditions, it may be difficult to find a model that can be shown to be predictive.

Finally, correlation values are obtained by linear regression of the geometric fit index. Fit functions check mapping of chemical substructures into feature constraints, as well as distance deviations of chemical functions from the center of the feature. Therefore, the geometric fit points out exactly how the function is localized in the center of a feature sphere. The correlation coefficient is based on linear regression derived from the geometric fit index.

Another, maybe more intuitive, validation method is represented by the capacity of a given hypothesis to yield a correct activity prediction. For this purpose, Catalyst also calculates errors, defined as the ratio between the experimental and the estimated activity. Thus, for instance, an error of +1.2 implies that the experimental activity is 20% higher than the predicted value, whereas an error of −1.1 means that the calculated activity is overestimated by 10%, referring to the pharmacophore model.

Last, to fully check the predictive capacity of the proposed pharmacophore hypothesis, a test set made up of another 29 molecules[23] was considered (Chart 2). The compounds were built and their conformational analysis was performed following the same procedure applied to the training set compounds.

Biological Assay. The RNA-dependent RNA polymerase assay[38] was performed in a total volume of 40 μL containing 1 nM enzyme, 20 mM Tris-HCI (pH 7.5), 5 mM $MgCl_2$, 1 mM DTT, 1 mM EDTA, 20 units of RNase inhibitor, 50 μg/mL of actinomycin D, 5 μCi of [α-^{32}P]UTP, and 0.5 mM each of the remaining NTPs (i.e., ATP, CTP, and GTP) with 10 μg/mL of RNA template.[39] The concentration of the limiting nucleotide was adjusted to 10 μM. The reaction mixtures were incubated at 30 °C for 2 h. After incubation, the reaction was stopped by digestion with 50 μg of proteinase K (Boehringer Mannheim) in a proteinase K buffer (150 mM NaCI, 50 mM Tris-HCI (pH 7.5), and 0.5% SDS) for 30 min. The RNA products were extracted with phenol/chloroform (1:1) before ethanol precipitation. After heat denaturation (94 °C, 2 min), these products were analyzed by 8 M urea, 8% PAGE. The gels were dried and analyzed using a BAS 1000 BioImage analyzer system (Fuji).

Acknowledgment. Thanks are due to the MIUR, PRIN, and FIRB project on "Design, synthesis and evaluation of antiviral activity of inhibitors aimed at different targets in the replication cycle of Flaviviridae" (Grant Nos. RBNE01J3SK_006 and RBNE01J3SK_010). R.D.S. also thanks the Ministero della Sanità, Istituto Superiore di Sanità, Progetto AIDS (Grant No. 30F.19).

References

(1) Choo, Q. L.; Kuo, G.; Weiner, A. J.; Overby, L. R.; Bradley, D. W.; Houghton, M. Isolation of a cDNA Clone Derived from a Blood-Borne non-A non-B Viral Hepatitis Genome. *Science* **1989**, *244*, 359−362.

(2) Alter, M. J.; Kruszon-Moran, D.; Nainan, O. V.; McQuillan, G. M.; Gao, F.; Moyer, L. A.; Kaslow, R. A.; Margolis, H. S. The Prevalence of Hepatitis C Virus Infection in the United States, 1988 through 1994. *N. Engl. J. Med.* **1999**, *341*, 556−562.

(3) Leyssen, P.; De Clercq, E.; Neyts, J. Perspectives for the Treatment of Infections with *Flaviviridae*. *Clin. Microbiol. Rev.* **2000**, *13*, 67−82.

(4) Memon, M. I.; Memon, M. A. Hepatitis C: an epidemiological review. *J. Viral Hepatitis* **2000**, *9*, 84−100.

(5) Liang, T. J.; Rehermann, B.; Seeff, L. B.; Hoofnagle, J. H. Pathogenesis, Natural History, Treatment and Prevention of Hepatitis C. *Ann. Intern. Med.* **2000**, *132*, 296−305.

(6) Lauer, G. M.; Walker, B. D. Hepatitis C Virus Infection. *N. Engl. J. Med.* **2001**, *345*, 41−52.

(7) (a) Di Bisceglie, A. M.; Hoofnagle, J. H. Optimal Therapy of Hepatitis C. *Hepatology* **2002**, *36* (Suppl. 1), S121−127. (b) Chandler, G.; Sulkowski, M. S.; Jenckes, M. W.; Torbenson, M. S.; Herlong, H. F.; Bass, E. B.; Gebo, K. A. Treatment of Chronic Hepatitis C: A Systematic Review. *Hepatology* **2002**, *36*, S135−144.

(8) Scott, J. J.; Perry, C. M. Interferon-α-2b Plus Ribavirin. A Review of its Use in the Management of Chronic Hepatitis C. *Drugs* **2002**, *62*, 507−556.

(9) Manns, M. P.; McHutchison, J. G.; Gordon, S. C.; Rustgi, V. K.; Shiffman, M.; Goodman, Z. D.; Koury, K.; Ling, M.; Albrecht, J. K. Peginterferon alfa-2b Plus Ribavirin Compared with Interferon alfa-2b Plus Ribavirin for Initial Treatment of Chronic Hepatitis C: a Randomised Trial. *Lancet* **2001**, *358*, 958−965.

(10) Murphy, F. A.; Fauquet, C. M.; Bishop, D. H.; Ghabrial, S. A.; Jarvis, A. W.; Martelli, G. P.; Mayo, M. A.; Summers, M. D. *Classification and Nomenclature of Viruses: Sixth Report of the International Committee on Taxonomy of Viruses*; Springer-Verlag: Wien, Germany, 1995; pp 424−426.

(11) Lindenbach, B. D.; Rice, C. M. *Flaviviridae*: the Viruses and their Replication. In *Fields Virology*, 4th ed.; Knipe, D. M., Howley, P. M., Griffin, D. E., Lamb, R. A., Martin, M. A., Riozman, B., Eds.; Lippincott, Wilkins and Wilkins: Philadephia, PA, 2001; pp 991−1041.

(12) Behrens, S.-E.; Tomei, L.; De Francesco, R. Identification and Properties of the RNA-Dependent RNA Polymerase of Hepatitis C Virus. *EMBO J.* **1996**, *15*, 12−22.

(13) Ago, H.; Adachi, T.; Yoshida, A.; Yamamoto, M.; Habuka, N.; Yatsunami, K.; Miyano, M. Crystal Structure of the RNA-Dependent RNA Polymerase of Hepatitis C Virus. *Structure* **1999**, *7*, 1417−1426.

(14) Bressanelli, S.; Tomei, L.; Roussel, A.; Incitti, I.; Vitale, R. L.; Mathieu, M.; De Francesco, R.; Rey, F. A. Crystal Structure of the RNA-Dependent RNA Polymerase of Hepatitis C Virus. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13034−13039.

(15) Lesburg, C. A.; Cable, M. B.; Ferrari, E.; Hong, Z.; Mannarino, A. F.; Weber, P. C. Crystal Structure of the RNA-Dependent RNA Polymerase from Hepatitis C Virus Reveals a Fully Encircled Active Site. *Nat. Struct. Biol.* **1999**, *6*, 937−943.

(16) Lohmann, V.; Roos, A.; Korner, F.; Koch, J. O.; Bartenschlager, R. Biochemical and Kinetic Analyses of NS5B RNA-Dependent RNA Polymerase of Hepatitis C Virus. *Virology* **1998**, *249*, 108−118.

(17) Ishii, K.; Tanaka, Y.; Yap, C. C.; Aizaki, H.; Matsuura, Y.; Miyamura, T. Expression of Hepatitis C Virus NS5B Protein: Characterization of its RNA Polymerase Activity and RNA Binding. *Hepatology* **1999**, *29*, 1227−1235.

(18) Johnson, R. B.; Sun, X. L.; Hockman, M. A.; Villarreal, E. C.; Wakulchik, M.; Wang, Q. M. Specificity and Mechanism Analysis of Hepatitis C Virus RNA-Dependent RNA Polymerase. *Arch. Biochem. Biophys.* **2000**, *377*, 129−134.

(19) Ferrari, E.; Wrigtht-Minogue, J.; Fang, J. W.; Baroudy, B. M.; Lau, J. Y.; Hong, Z. Characterization of Soluble Hepatitic C Virus RNA-Dependent RNA Polymerase Expressed in *Escherichia coli*. *J. Virol.* **1999**, *73*, 1649−1654.

(20) Rodriguez, P. L.; Carrasco, L. Gliotoxin: Inhibitor of Poliovirus RNA Synthesis that Blocks the Viral RNA Polymerase 3Dpol. *J. Virol.* **1992**, *66*, 1971−1976.

(21) (a) Stuyver, L. J.; Whitaker, T.; McBrayer, T. R.; Hernandez-Santiago, B. I.; Lostia, S.; Tharnish, P. M.; Ramesh, M.; Chu, C. K.; Jordan, R.; Shi, J.; Rachakonda, S.; Watanabe, K. A.; Otto, M. J.; Schinazi, R. F. Ribonucleoside Analogue That Blocks Replicaton of Bovine Viral Diarrhea and Hepatitis C Viruses in Culture. *Antimicrob. Agents Chemother.* **2003**, *47*, 244−254. (b) Carroll, S. S.; Tomassini, J. E.; Bosserman, M.; Getty, K.; Stahlhut, M. W.; Eldrup, A. B.; Baht, B.; Hall, D.; Simcoe, A. L.; LaFemina, R.; Rutkowski, C. A.; Wolanski, B.; Yang, Z.; Migliaccio, G.; DeFrancesco, R.; Kuo, L. C.; MacCoss, M.; Olsen, D. B. Inhibition of Hepatitis C Virus RNA Replication by 2′-Modified Nucloside Analogues. *J. Biol. Chem.* **2003**, *278*, 11979−11984. (c) Shim, J.; Larson, G.; Lai, G.; Naim, V.; Wu, S.; Canonical, J. Z. 3′-Deoxyribonucleotides as Chain Terminator for HCV NS5B RNA Polymerase. *Antiviral Res.* **2003**, 243−251.

(22) (a) Reddy, T. J.; Chan, L.; Turcotte, N.; Proulx, M.; Pereira, O. Z.; Das, S. K.; Siddiqui, A.; Wang, W.; Poisson, C.; Yannopoulos, C. G.; Bilimoria, D.; L'Heureux, L.; Alaoui, H. M. A.; Nguyen-Ba, N. Further, SAR Studies on Novel Small, Molecule Inhibitors of the Hepatitis C (HCV) NS5B Polymerase. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3341−3344. (b) Wang, M.; Ng, K. K. S.; Cherney, M. M.; Chan, L.; Yannopoulos, C. G.; Bedard, J.; Morin, N.; Nguyen-Ba, N.; Alaoui-Ismaili, M. H.; Bethell, R. C.; James, M.

N. G. Nonnucleoside Analogue Inhibitors Bind to Allosteric Site on HCV NS5B Polymerase. *J. Biol. Chem.* **2003**, *278*, 9489−9495. (c) Dhanak, D.; Duffy, K. J.; Johnston, V. K.; Lin-Goerke, J.; Darcy, M.; Shaw, A. N.; Gu, B.; Silverman, C.; Gates, A. T.; Nonnemacher, M. R.; Earashaw, D. L.; Casper, D. J.; Kaura, A.; Baker, A.; Greenwood, C.; Gutshall, L. L.; Maley, D.; DelVecchio, A.; Macarron, R.; Hofmann, G. A.; Alnoah, Z.; Cheng, H.-Y.; Chan, G.; Khandekar S.; Keenan, R. M.; Sarisky, R. T. Identification and Biological Characterization of Heterocyclic Inhibitors of the Hepatitis C Virus RNA-Dependent RNA Polymerase. *J. Biol. Chem.* **2002**, *277*, 38322−38327. (d) Chan, L.; Reddy, T. J.; Proulx, M.; Das, S. K; Pereira, O.; Wang, W.; Siddiqui, A.; Yannopoulos, C. G.; Poisson, C.; Turcotte, N.; Drouin, A.; Alaoui-Ismaili, M. H.; Bethell, R.; Hamel, M.; L′Heureux, L.; Bilimoria, D.; Nguyen-Ba, N. Identification of *N*,*N*-Disubstituted Phenylalanines as a Novel Class of Inhibitors of Hepatitis C NS5B Polymerase. *J. Med. Chem.* **2003**, *46*, 1283−1285. (e) De Francesco, R.; Tomei, L.; Altamura, S.; Summa, V.; Migliaccio, G. Approaching a New Era for Hepatitis C Virus Therapy: Inhibitors of the NS3−4A Serine Protease and the NS5B RNA-Dependent RNA Polymerase. *Antiviral Res.* **2003**, *58*, 1−16. (f) Summa, V.; Petrocchi, A.; Pace, P.; Matassa, V. G.; DeFrancesco, R.; Altamura, S.; Tomei, L.; Koch, U.; Neuner, P. Discovery of α,γ-Diketo Acids as Potent, Selective and Reversible Inhibitors of Hepatitis C Virus NS5B RNA Dependent RNA Polymerase. *J. Med. Chem.* **2004**, *47*, 14−17. (g) Gu, B.; Johnston, V. K.; Gutshall, L. L.; Nguyen, T. T.; Gontarek, R. R.; Darcy, M. G.; Tedesco, R.; Dhanak, D.; Duffy, K. J.; Kao, C. C.; Sarisky, R. T. Arresting Initiation of Hepatitis C Virus RNA Synthesis Using Heterocyclic Derivatives. *J. Biol. Chem.* **2003**, *278*, 16602−16607. (h) Love, R. A.; Parge, H. E.; Yu, X.; Hickey, M. J.; Diehl, W.; Gao, J.; Wriggers, H.; Ekker, A.; Wang, L.; Thomson, J. A.; Dragovich, P. S.; Furhman, S. A. *J. Virol.* **2003**, *77*, 7575−7581.

(23) Altamura, S.; Tomei, L.; Koch, U.; Neuner, P. J.; Summa, V. Diketoacid-Derivatives as Inhibitors of Polymerases. PCT Int. Appl. WO 200006529, 2000; CAN 132: 132323, 2000.

(24) Artico, M.; Di Santo, R.; Costi, R.; Novellino, E.; Greco, G.; Massa, S.; Tramontano, E.; Marongiu, M. E.; De Montis, A.; La Colla, P. Geometrically and Conformationally Restrained Cinnamoyl-Compounds as Inhibitors of HIV-1 Integrase: Synthesis, Biological Evaluation and Molecular Modeling. *J. Med. Chem.* **1988**, *31*, 3948−3960.

(25) Di Santo, R.; Costi, R.; Artico, M.; Tramontano, E.; La Colla, P.; Pani A. HIV-1 Integrase Inhibitors that Block HIV-1 Replication in Infected Cells. Planning Synthetic Derivatives from Natural Products. *Pure Appl. Chem.* **2003**, *75*, 195−206.

(26) Costi, R.; Di Santo, R.; Artico, M.; Massa, S.; Ragno, R.; Loddo, R.; La Colla, M.; Tramomtano, E.; La Colla, P.; Pani. A. 2,6-Bis-(3,4,5-trihydroxybenzylidene) derivatives of cyclohexanone: novel potent HIV-1 integrase inhibitors that prevent HIV-1 multiplication in cell-based assays. *Bioorg. Med. Chem.* **2004**, *12*, 199−215.

(27) Costi, R.; Di Santo, R.; Artico, M.; Roux, A.; Ragno, R.; Massa, S.; Tramomtano, E.; La Colla, M.; Loddo, R.; Marongiu, M. E.; Pani, A.; La Colla, P. 6-Aryl-2,4-dioxo-5-hexenoic Acids, Novel Integrase Inhibitors Active Against HIV-1 Multiplication in Cell-Based Assays. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1745−1749.

(28) Di Santo, R.; Costi, R.; Artico, M.; Ragno, R.; Greco, G.; Novellino, E.; Marchand, C.; Pommier, Y. Design, synthesis and biological evaluation of heteroaryl diketohexenoic and diketobutanoic acids as HIV-1 integrase inhibitors endowed with antiretroviral activity. *Farmaco* **2005**, *60*, 409−417.

(29) Di Santo, R.; Costi, R.; Artico, M.; Marchand, C.; Pommier, Y. Quinoline inhibitors of retroviral integrase. U.S. Patent prov. num. US60/522,423, date of deposit March 10, 2004.

(30) Sprague, P. Automated Chemical Hypothesis Generation and Database Searching with Catalyst. In *Perspectives in Drug Discovery and Design*; Anderson, P. S., Kenyon, G. L., Marshall, G. R., Muller, K., Eds.; ESCOM: Leiden, 1995; Vol. 3, pp 1−20.

(31) Sutter, J.; Güner, O. F.; Hoffmann, R.; Li, H.; Waldman, M. In *Pharmacophore, Perception Development and Use in Drug Design*; Güner, O. F., Ed.; International University Line: La Jolla, CA, 2000; pp 504−506.

(32) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1995**, *16*, 171−187.

(33) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285−294.

(34) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 2. Application of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295−304.

(35) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(36) *Catalyst*, version 4.9; Accelrys Inc.: San Diego, CA, 2003.

(37) Briens, F.; Bureau, R.; Rault, S. Applicability of Catalyst in Ecotoxicology, a New Promising Tool for 3D-QSAR: Study of Chlorophenols. *Ecotoxicol. Environ. Saf.* **1999**, *43*, 241−251.

(38) Qin, W.; Luo, H.; Nomura, T.; Hayashi, N.; Yamashita, T.; Murakami, S. Oligomeric Interaction of Hepatitis C Virus NS5B Is Critical for Catalytic Activity of RNA-Dependent RNA Polymerase. *J. Biol. Chem.* **2002**, *277*, 2132−2137.

(39) Yamashita, T.; Kaneko, S.; Shirota, Y.; Qin, W.; Nomura, T.; Kobayashi, K.; Murakami, S. RNA-Dependent RNA Polymerase Activity of the Soluble Recombinant Hepatitis C Virus NS5B Protein Truncated at the C−Terminal Region. *J. Biol. Chem.* **1998**, *273*, 15479−15486.

JM0504454