ACS Medicinal Chemistry Letters

pubs.acs.org/acsmedchemlett

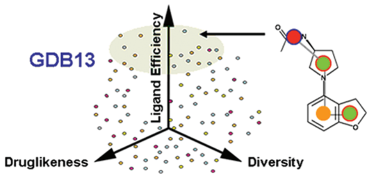
Design of a High Fragment Efficiency Library by Molecular Graph Theory

Jennifer Venhorst,[†] Sara Núñez,[*,†] and Chris G. Kruse

Abbott Healthcare Products, 1381 CP Weesp, The Netherlands

ABSTRACT Molecular graph theory was used to design a unique and diverse, high-efficiency fragment screening collection. A data set retrieved from the annotated database AurSCOPE GPS was used as the reference set, and the GDB-13 database, a virtual library of enumerated organic molecules, was used as a source for the fragment selection. The data graph collection of Discngine as implemented in PipelinePilot was applied to perform the graph pharmacophore similarity matching between the reference and the GDB-13 data sets, leading to the ultimate fragment screening library. The relevance of this unique fragment collection was demonstrated by means of a virtual screening exercise using human trypsin as a test case. Several novel entities with high similarity to known trypsin inhibitors were identified in the in silico exercise. The application of this unique, high fragment efficiency collection to other protein targets in the framework of fragment-based drug discovery is warranted.

KEYWORDS Fragment screening, ligand efficiency, graph pharmacophore, GDB-13, virtual screening, trypsin inhibition



High-throughput screening (HTS) has long been the dominant hit finding strategy in the pharmaceutical industry. Yet, throughout the past decade, significant shortcomings associated with these campaigns have become evident. For example, high false positive rates due to compound aggregation,[1,2] considerable false negative rates due to the suboptimal physicochemical properties of compound stocks, low hit rates due to the gratuitous complexity of screening libraries,[3] and negative campaign outcomes due to the poor intrinsic chemical diversity of compound stocks[4] have impaired the discovery of novel chemical matter for therapeutically attractive targets. The undermined productivity of HTS together with the rising success of structure-based drug design paved the way for the more provocative fragment-based drug discovery (FBDD) paradigm.[5] FBDD is nowadays an established technique for hit finding; in essence, it bases its strength on the competent binding of small chemical entities to their targets.[6] One apparent advantage of FBDD libraries is that their relatively small size (typically 500−5000 fragments) can thoroughly sample drug space in a proficient manner.[7] Yet, this remarkable advantage is accompanied by a potential liability: Oversimplifying the screening collection to a set of commercial entities regularly used by the entire FBDD community could give rise to comparable hit fragments. In such a case, the ensuing hit evolution toward patentable leads could prove challenging. With the purpose of avoiding this scenario, we advocate committing substantial computational, analytical, and synthetic resources toward the design and implementation of an exclusive FBDD screening armory. Specifically, we propose that the scaffold complexity of the rule-of-three compliant[8] screening library be significantly elevated so as to circumvent previously explored chemical space.

In this study, a unique and diverse, high-efficiency fragment collection was designed using the GDB-13 database[9] as a source of fragments. GDB-13 is a virtual collection of enumerated organic molecules of up to 13 atoms. This collection includes approximately 970 million C/N/O/S/Cl-containing compounds that conform to simple chemical stability and synthetic feasibility rules. To our knowledge, GDB-13 is the largest small molecule database to date that is publicly available. Recent hit finding efforts using the related GDB-11 database proved successful in discovering novel NMDA glycine site inhibitors.[10]

The GDB-13 database was preprocessed so as to retain only fragmentlike scaffolds suitable for FBDD screening. The applied chemical criteria are as follows: (1) rule-of-three compliance, (2) nontoxicity/reactivity compliance,[11−13] and (3) noncommercial availability[14] (section 1 in the Supporting Information). In parallel, AurSCOPE GPS,[15] an extensive knowledge database containing quantitative biological activity data for the majority of therapeutic drug targets, was mined to assemble the reference set. This set was composed of 2329 compounds obeying the rule-of-three and having a high affinity to their biological targets, which include ion channels, GPCRs, proteases, and kinases (section 2 in the Supporting Information). The FBDD profile of this reference

Received Date: July 8, 2010
Accepted Date: September 1, 2010
Published on Web Date: September 03, 2010

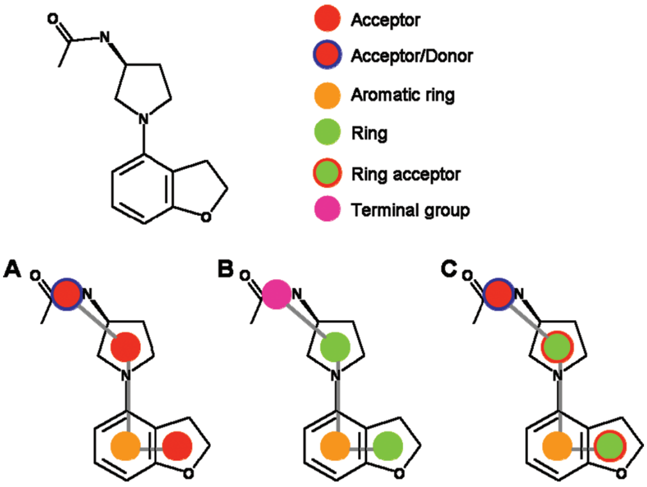


Figure 1. Three types of reduction schemes are possible within Discngine: (A) pharmacophore-based, (B) topology-based, and (C) pharmacophore and topology-based. The latter was used in this study. Nodes recognized for pharmacophore-based graphs are as follows: acceptor (hydrogen bond acceptor), donor (hydrogen bond donor), negative charge, positive charge, aromaticity, lipophilicity, and ionizable. Nodes for topology-based graphs include aromatic or planar ring, nonplanar ring, linker, and terminal group.

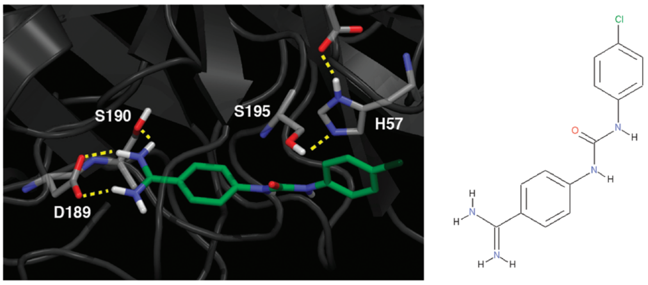


Figure 2. Active site of human trypsin with a crystallized inhibitor (left). The relevant S1 pocket formed by Asp189 (D189) and Ser190 (S190) and the secondary subsite formed by the catalytic Ser195 (S195), His57 (H57), and Asp102 triad are highlighted. The chemical structure of the small inhibitor is also shown (right).

set was ascertained by analysis of several physicochemical descriptors (Figure 1 in the Supporting Information); this collection thus constitutes a suitable reference data set with which to execute the molecular graph pharmacophore matching algorithm against GDB-13.

Subsequently, molecular graph theory[16] was used for the graph pharmacophore generation and subsequent matching of the GDB-13 database against the reference data set. A molecular graph pharmacophore is a simple way of representing an entity by a set of nodes and their spatial relationship and was here chosen for its successful performance in lead hopping practices.[17,18] The data graph collection of Discngine[19] as implemented in PipelinePilot[20] was used to encode the reference and GDB-13 data sets into graph pharmacophores (Figure 1). The pharmacophore graph matching algorithm was thereafter used to select those GDB-13 entities scoring 0.9 or higher toward the reference set, resulting in a total of 53600 nonredundant fragment hits.

Next, a pareto multiobjective optimization[21−23] was used to reduce the above-mentioned hit set to the smallest fragment collection that describes fragment diversity in drug space competently. The following objectives were taken into consideration: (1) high structural diversity (fragment level), (2) high structural diversity (scaffold level), (3) high scaffold complexity,[24] (4) high chemical tractability, and (5) low scaffold redundancy (section 3 in the Supporting Information). This resulted in the ultimate 1357 fragment set, here referred to as the high fragment efficiency (HFE) set (available upon request). Physicochemical descriptor distribution analysis of the HFE set confirmed the suitability of this diverse fragment collection for FBDD screening purposes (Figure 2 in the Supporting Information), which parallels the distribution of commercial fragment collections.[7] A plausible explanation for the modest variation in the descriptor distribution (e.g., molecular weight) stems from this unique library being assembled on basis of the 1−13 heavy atom premise.[9]

Next, the intrinsic chemical diversities of the HFE set, a commercial fragment collection (iNovacia AB[25]), and a random fragment selection (here used as a control) were examined. Such analysis established a higher diversity within the HFE set than that of iNovacia's or the random fragment selection (Figure 3 in the Supporting Information). Moreover, the chemical similarity between the HFE/iNovacia fragment collections was assessed and concluded to be low (Figure 4 in the Supporting Information). In addition, a substructure search in the Prous Science Integrity database[26] ascertained that the HFE library finds zero occurrences whereas 25% of iNovacia AB's fragment collection is often recovered in compounds in R&D stages.[26] In addition, a chemical tractability analysis of the HFE, iNovacia, and control fragment collections showed a higher number of chemical handles for fragments in the HFE data set, thus making this screening set a more tractable starting point for subsequent chemical elaboration (Figure 5 in the Supporting Information).

Next, the applicability of the HFE fragment collection was interrogated by means of a virtual screening exercise using the serine protease trypsin as a test case. For comparison purposes, the iNovacia and control fragment sets were studied in parallel. The binding of inhibitors to trypsin is characterized by a primary specificity pocket (S1) and additional secondary subsites (Figure 2). It is well-documented that the presence of Asp189, at the bottom of S1, is the major enthalpic determinant for the narrow affinity and specificity of trypsin for positively charged substrates/inhibitors.[27,28]

The crystal structure of human trypsin in complex with a small inhibitor obtained from the publicly available ZINC database[29] was used in this study. The HFE, iNovacia, and control fragment sets were docked into human trypsin so as to sample possible interaction poses using GOLD.[30] Molecular interaction fingerprints[31−33] were thereafter used as a postprocessing tool to discard nonrelevant docking poses (section 6 in the Supporting Information). Specifically, the following constraints were imposed for each screening fragment: (1) polar interaction with Asp189, (2) polar/hydrophobic/aromatic interaction with Ser195/His57, and (3) minimal intermolecular clashes.[34] After visual inspection,

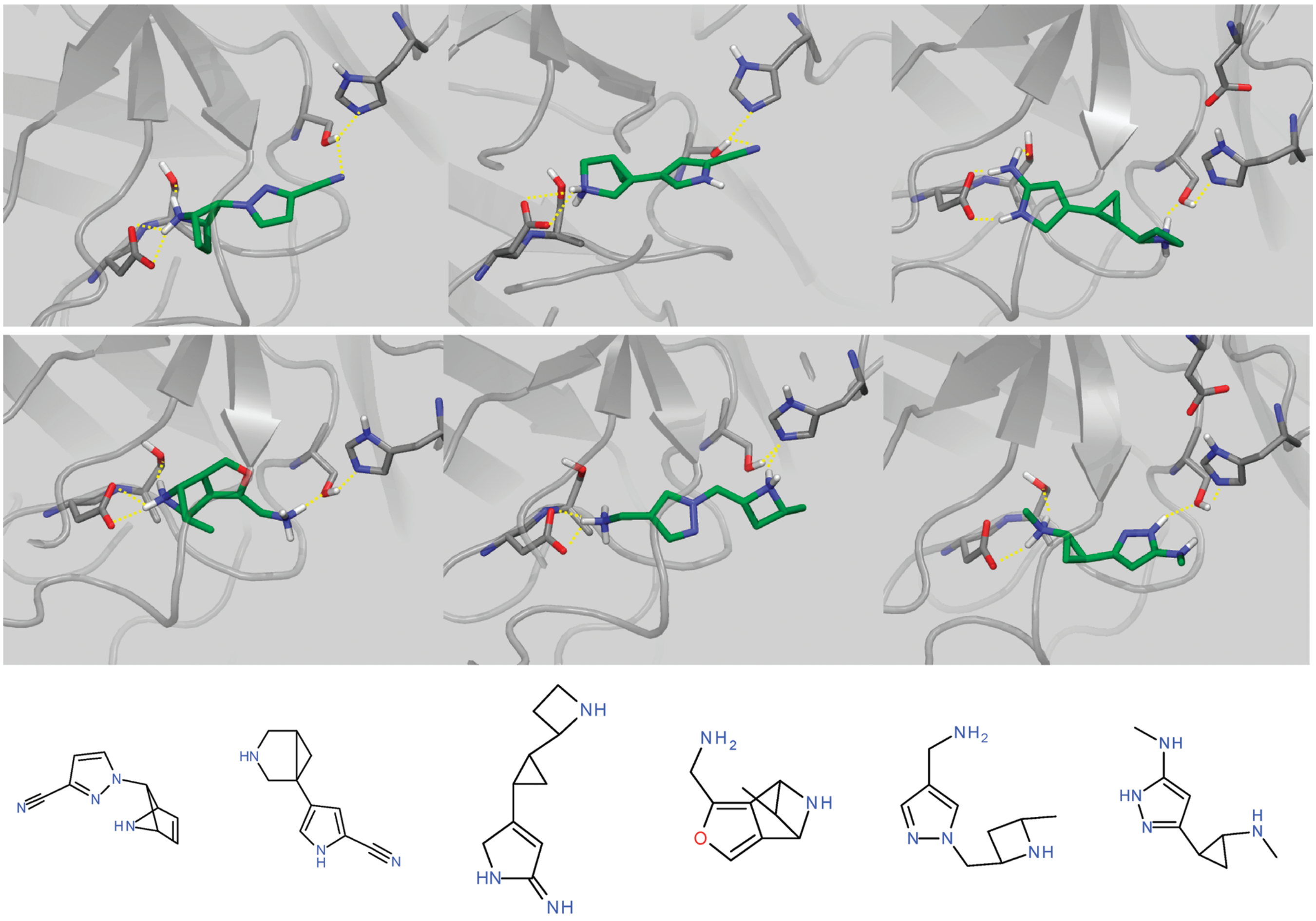


Figure 3. Three-dimensional representation of a small selection of the HFE hits. Intermolecular interactions between each fragment and the S1 pocket/catalytic triad of trypsin and the respective chemical structures are shown.

an average of 80 nonredundant hit fragments was accepted from the HFE, iNovacia, and control fragment collections.

To quantify the relevance of the aforementioned fragment hits, a chemical similarity analysis[35] was performed to determine the correspondence between several trypsin inhibitors and HFE/iNovacia/control hit sets. The trypsin inhibitor set was obtained from the ZINC database and further enriched with those contained in the Prous Science Integrity database. For chemical comparison simplicity, each inhibitor was thereafter reduced to the simplest scaffold that interacts with the binding site residues majorly involved in recognition (Figure 2).

The chemical similarity between the HFE/iNovacia/control hit sets and trypsin's inhibitor set was analyzed (Figure 6 in the Supporting Information). It can be readily observed that the similarity of the HFE hit set to the trypsin reference collection exceeds those of iNovacia/control hit sets. Moreover, it appears that the intrinsic diversity of HFE was somewhat higher than that of the iNovacia/control sets. Analysis of the chemical tractability of the hit sets showed that the HFE hit set presents a higher number of chemical handles, which would make fragment elaboration by manual/parallel chemistry toward leads more amenable. All in all, these data suggest that the HFE set represents a more diverse, higher efficiency FBDD collection with high chemical tractability potential.

Representative HFE hit fragments and their docking poses within trypsin's binding pocket are shown in Figure 3. It can be observed that the fragments are well-anchored in the binding site, making strong and specific interactions with crucial residues, thus suggesting high enthalpic contribution to the binding event. Structural similarity can be observed between the distinctive benzamidine motif and several of its amine analogues.[27,36,37] Specifically, the calculated chemical similarity for these six fragments ranges between 0.64 and 0.77. Moreover, all fragments are readily derivatizable through one chemical handle that is not accountable for their binding to trypsin.

A substructure search in the Prous Science Integrity database with the iNovacia AB and HFE hit sets ascertained that our HFE hits have zero occurrences within databases containing entities in R&D stages, whereas iNovacia AB's hit set was often recovered in known compounds. This implies that the structural optimization of these hit fragments into an IP-free drug space may be more burdensome and emphasizes our thesis of investing resources in the design and implementation of a unique screening fragment library that capitalizes on more amenable hit-to-lead and lead optimization efforts.

In summary, unprecedented efforts to design a diverse, high-efficiency, unique fragment collection based on molecular graph theory are here described. This fragment collection navigates a chemical space not previously populated by typical commercial FBDD libraries. Chemical matter was carefully selected by applying classical Ro3 filters and chemical tractability, diversity, scaffold complexity, and noncommercial criteria. The relevance of this unique fragment collection was demonstrated by means of a virtual screening exercise using human trypsin as the test case. Some novel and highly tractable fragments with high chemical similarity to trypsin inhibition motifs were identified in the virtual screening exercise, opening a door to the discovery of novel trypsin inhibitors. The applicability of this diverse, high-efficacy fragment collection to other protein targets in the frame of FBDD is inferred.

SUPPORTING INFORMATION AVAILABLE Computational procedures as well as outcome of the analyses discussed in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author: *To whom correspondence should be addressed. Tel: 0031-294-479868. E-mail: sara.nunez@abbott.com.

Author Contributions: † Both authors have contributed equally to the studies presented.

ACKNOWLEDGMENT We acknowledge the help of Dr. Antoniou and Dr. McCormack for proofreading the manuscript, Dr. Bronger for assistance with the chemical tractability of the fragments, and Dr. Lange and Dr. Ducrot for valuable scientific discussions.

REFERENCES

(1) Coan, K. E.; Shoichet, B. K. Stoichiometry and physical chemistry of promiscuous aggregate-based inhibitors. *J. Am. Chem. Soc.* **2008,** *130*, 9606–9612.

(2) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **2007,** *50*, 2385–2390.

(3) Irwin, J. J. How good is your screening library? *Curr. Opin. Chem. Biol.* **2006,** *10*, 352–356.

(4) Snowden, M.; Green, D. V. The impact of diversity-based, high-throughput screening on drug discovery: "Chance favours the prepared mind. *Curr. Opin. Drug Discovery Dev.* **2008,** *11*, 553–558.

(5) Hajduk, P. J. SAR by NMR: putting the pieces together. *Mol. Interventions* **2006,** *6*, 266–272.

(6) Schulz, M. N.; Hubbard, R. E. Recent progress in fragment-based lead discovery. *Curr. Opin. Pharmacol.* **2009,** *9*, 615–621.

(7) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007,** *6*, 211–219.

(8) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A "rule of three" for fragment-based lead discovery? *Drug Discovery Today* **2003,** *8*, 876–877.

(9) Blum, L. C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009,** *131*, 8732–8733.

(10) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Discovery of NMDA glycine site inhibitors from the chemical universe database GDB. *Chem. Med. Chem.* **2008,** *3*, 1520–1524.

(11) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010,** *53*, 2719–2740.

(12) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000,** *14*, 251–264.

(13) Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today* **2003,** *8*, 86–96.

(14) Symyx Solutions, Inc. http://www.symyx.com.

(15) Aureus Pharma. http://www.aureus-pharma.com/Pages/Products/Aurscope_GPS.php.

(16) Amigó, J. M.; Gálvez, J.; Villar, V. M. A review on molecular topology: Applying graph theory to drug discovery and design. *Naturwissenschaften* **2009,** *96*, 749–761.

(17) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem.* **2004,** *47*, 6144–6159.

(18) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003,** *43*, 338–345.

(19) Discngine S.A.A. Romainville, France. http://www.discngine.com.

(20) *PipelinePilot*, version 7.0.1; Accelrys, Inc.: San Diego, CA, 2008.

(21) Andersson, J. A. A survey of multiobjective optimization in engineering design, Technical report: LiTH-IKP-R-1097. Dept. of Mechanical Engineering, Linköping University, 2000.

(22) Fonseca, C. M.; Fleming, P. J. An overview of evolutionary algorithms in multiobjective optimization. *Evol. Comput.* **1995,** *3*, 1–16.

(23) Geoffrion, A. M. Proper efficiency and the theory of vector optimization. *J. Math. Anal. Appl.* **1968,** *22*, 618–630.

(24) Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem.* **2002,** *45*, 5311–5320.

(25) iNovacia AB, Stockholm, Sweden. http://www.inovacia.se.

(26) Prous Science Integrity. http://www.prous.com/integrity.

(27) Leiros, H. K.; Brandsdal, B. O.; Andersen, O. A.; Os, V.; Leiros, I.; Helland, R.; Otlewski, J.; Willassen, N. P.; Smalås, A. O. Trypsin specificity as elucidated by LIE calculations, X-ray structures, and association constant measurements. *Protein Sci.* **2004,** *13*, 1056–1070.

(28) Katz, B. A.; Elrod, K.; Luong, C.; Rice, M. J.; Mackman, R. L.; Sprengeler, P. A.; Spencer, J.; Hataye, J.; Janc, J.; Link, J.; Litvak, J.; Rai, R.; Rice, K.; Sideris, S.; Verner, E.; Young, W. A novel serine protease inhibition motif involving a multicentered short hydrogen bonding network at the active site. *J. Mol. Biol.* **2001,** *307*, 1451–1486.

(29) Irwin, J. J.; Shoichet, B. K. ZINC: A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005,** *45*, 177–182.

(30) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997,** *267*, 727–748.

(31) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007,** *47*, 195–207.

(32) Venhorst, J.; Núñez, S.; Terpstra, J. W.; Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular

interaction fingerprints. *J. Med. Chem.* **2008,** *51*, 3222–3229.

(33) Núñez, S.; Venhorst, J.; Kruse, C. G. Assessment of a novel scoring method based on solvent accessible surface area descriptors. *J. Chem. Inf. Model.* **2010,** *50*, 480–486.

(34) Katona, G.; Berglund, G. I.; Hajdu, J.; Gráf, L.; Szilágyi, L. Crystal structure reveals basis for the inhibitor resistance of human brain trypsin. *J. Mol. Biol.* **2002,** *315*, 1209–1218.

(35) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivaschenko, A. A. New diversity calculations algorithms used for compound selection. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 249–258.

(36) Renatus, M.; Bode, W.; Huber, R.; Stürzebecher, J.; Stubbs, M. T. Structural and functional analyses of benzamidine-based inhibitors in complex with trypsin: implications for the inhibition of factor Xa, tPA, and urokinase. *J. Med. Chem.* **1998,** *41*, 5445–5456.

(37) Zhou, Y.; Johnson, M. E. Comparative molecular modeling analysis of-5-amidinoindole and benzamidine binding to thrombin and trypsin: Specific H-bond formation contributes to high 5-amidinoindole potency and selectivity for thrombin and factor Xa. *J. Mol. Recognit.* **1999,** *12*, 235–241.