Published in final edited form as:
Mol Pharm. 2011 October 3; 8(5): 1611–1618. doi:10.1021/mp200093z.

The Subcellular Distribution of Small Molecules: A Meta-Analysis

Nan Zheng[†], **Hobart Ng Tsai**[†], **Xinyuan Zhang**[†], **Kerby Shedden**[#], and **Gus R. Rosania**[*,†]
Department of Pharmaceutical Sciences, College of Pharmacy, and Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109

Abstract

To explore the extent to which current knowledge about the organelle-targeting features of small molecules may be applicable towards controlling the accumulation and distribution of exogenous chemical agents inside cells, molecules with known subcellular localization properties (as reported in the scientific literature) were compiled into a single data set. This data set was compared to a reference data set of approved drug molecules derived from the DrugBank database, and to a reference data set of random organic molecules derived from the PubChem database. Cheminformatic analysis revealed that molecules with reported subcellular localizations were comparably diverse. However, the calculated physicochemical properties of molecules reported to accumulate in different organelles were markedly overlapping. In relation to the reference sets of Drug Bank and Pubchem molecules, molecules with reported subcellular localizations were biased towards larger, more complex chemical structures possessing multiple ionizable functional groups and higher lipophilicity. Stratifying molecules based on molecular weight revealed that many physicochemical properties trends associated with specific organelles were reversed in smaller vs. larger molecules. Most likely, these reversed trends are due to the different transport mechanisms determining the subcellular localization of molecules of different sizes. Molecular weight can be dramatically altered by tagging molecules with fluorophores or by incorporating organelle targeting motifs. Generally, in order to better exploit structure-localization relationships, subcellular targeting strategies would benefit from analysis of the biodistribution effects resulting from variations in the size of the molecules.

Keywords

drug transport; pharmacokinetics; biodistribution; drug targeting; databases; mathematical modeling; drug delivery; cheminformatics

Introduction

To develop small molecule chemical agents that accumulate at specific sites within cells, one would need to address not only bioavailability and tissue distribution issues at a systemic level, but also focus on delivery and targeting strategies at the subcellular level. In this context, knowledge about the relationships between the physicochemical properties and subcellular distribution of exogenous chemical agents could lead to greater understanding of their biological effects and could serve as a basis for the rational design of chemical agents "supertargeted" to specific sites of action within cells [1]. As such, supertargeted collections

[*]To whom correspondence should be addressed. Address: College of Pharmacy, the University of Michigan. 428 Church Street, Ann Arbor, MI 48109-1065. Telephone: (734)763-1032. Fax: (734)615-6162. grosania@umich.edu.
[†]Department of Pharmaceutical Sciences, University of Michigan College of Pharmacy.
[#]Department of Statistics, University of Michigan.

of chemical agents could serve as a starting point for developing more potent and less toxic drug leads, focusing on molecules that concentrate at intended sites of action while avoiding unwanted interactions with unintended targets.

The scientific literature supports the notion that many small molecule chemical agents tend to accumulate in specific organelles. The localization is usually supported by evidence including physical interaction with organelle components, resulting changes in organelle structure and function, or it may be visualized microscopically when a molecule has a specific optical signature. At a microscopic level, tissue distribution profiles depend on drug molecules crossing cellular membranes. During this process, drug molecules may also accumulate in various subcellular organelles, or bind to components such as lipids, proteins, DNA, RNA that localize to different intracellular or extracellular compartments. Specific properties of small molecules (pK_a, log*P*, molecular size, formal charges, hydrogen bond forming capacity, etc.) have been associated with predictable differences in systemic bioavailability and tissue distribution [2–5]. Indeed, a comprehensive cheminformatic meta-analysis of the physicochemical and subcellular distribution properties of small molecules as reported in the scientific literature could lead to interesting insights and would be important to prioritize future research efforts in this area.

Here, to help assess the status of current knowledge about the distribution of small molecules inside cells and its application to subcellular drug targeting and delivery, we compiled a data set of small molecules with reported subcellular localization features. In turn, a meta-analysis was performed to reveal how chemical structure and physicochemical properties are associated with the subcellular transport and biodistribution properties of exogenous chemical agents inside cells.

Methods

Data collection

Manual, text-based searches were undertaken using PubMed, Web of Sciences, MEDLINE and a commercial catalogue of fluorescent probes (Molecular Probes Catalog, Invitrogen Inc) using standard MESH terms (i.e., lysosome, mitochondria, nucleus, cell membrane/plasma membrane/cytoplasmic membrane, endoplasmic reticulum, Golgi apparatus/Golgi complex, subcellular, intracellular, accumulation, distribution) to identify small molecules exhibiting organelle-specific intracellular localization patterns. This initial pool of references was expanded by searching for articles written by the same authors, articles citing or cited by these articles, as well as articles describing studies perforomed on related compounds as identified by searching chemical substance names (e.g. styryls, amines, etc.). For molecules that were found in review articles or catalogues, the chemical name (and synonyms) were used as key words to search PubMed and Google Scholar for original research articles describing experimental evidence documenting their subcellular localization.

Database construction

Each molecule was incorporated into a database of 967 unique compounds with subcellular localization information about their chemical structure and distribution profile (Supplemental Table 1–10). Claims for a specific subcellular distribution pattern were established based on the authors' interpretation of the data. For example: "*compound X* targets *organelle Y*"; "*compound X* (strongly/mainly/predominantly/selectively) localized in *organelle Y*" [6–10]; "*compound X* exhibited a *organelle Y* localization" [8, 11]; "*compound X* mostly concentrating in *organelle Y*" [12, 13]; "*method Z* showed significant enrichment of *compound X* in *organelle Y*" [14]; "strong *organelle Y* accumulation was observed for

compound X" [15]; "*compound X* (preferentially) accumulated in *organelle Y*" [16–18]; "*Z* percentage of *compound X* was associated with *organelle Y*" [19]; "subcellular distribution of *compound X1* was almost identical with the distribution of *compound X2*" [20]; "*organelle Y* accounted for approximately *Z* percentage of the total distribution" [21–23]. Each entry was linked to the main reference source about the compound's subcellular localization. Compound chemical structures were sketched in MOE (Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada) using the Molecule Builder, then reduced to single connected components (i.e., without counter-ions) with MOE Wash algorithm, and then converted to Simplified Molecular Input Line Entry Specification (SMILES) strings.

Localization categories

For integrative analysis, we manually grouped the chemical agents into one of seven major categories, based on their reported site of accumulation (Supplemental Table 1–10). Functional considerations led us to consider lysosomes and endosomes as a single endolysosomal compartment because the molecular components of endosomes and lysosomes generally overlap in different cell lines, possessing an acidic lumen pH and readily exchanging contents. Molecules accumulating in the endoplasmic membrane (ER) and Golgi apparatus were also grouped together since these two organelles share similar protein markers and exchange content (localization to the Golgi and ER is generally reported together, because these two organelles are also difficult to distinguish using fluorescence microscopy).

Database comparisons

For comparison purposes, a random sample of 1000 compounds was downloaded from DrugBank [24, 25] which represents a collection of drugs that have been approved by the FDA (Supplemental Table 11). Similarly, a random sample of 982 compounds was downloaded from PubChem which represents an arbitrary sample of small organic compounds (Supplemental Table 12). The two reference datasets did not have overlapping molecules. ChemAxon and MOE were used to calculate molecular descriptors of the major micro-species at pH 7.4 for each compound in the subcellular localization dataset, or the PubChem and DrugBank reference sets. Z-score was computed to compare the mean descriptor value of molecules in the database to PubChem or DrugBank samples, according to the equation:

$$Z-\text{score}=(X_1-X_0)\div\sqrt{\frac{s_1^2}{n_1}+\frac{s_0^2}{n_0}}$$

, where X_1 and X_0 are the mean descriptor value of two subgroups (i.e. the subcellular localization dataset vs PubChem or DrugBank datasets; the lower vs. higher molecular weight compounds; or, the targeting vs non-targeting compounds); s_1^2 and s_0^2 are the sample variances of the corresponding populations; and, n_1 and n_0 are the number of molecules in the corresponding populations. A positive or negative Z-score with an absolute value greater than 3.1 indicates X_1 is significantly greater or less then X_0 (p-value < 0.001). The histograms of molecular descriptors of the compounds in both datasets were plotted and overlaid for visual comparison. Statistical analyses were performed with Python 2.5 (www.python.org).

Discriminant analysis

Linear discriminant analysis (LDA) was used to elucidate how the seven molecular descriptors that showed greatest association with individual subcellular localization categories (molecular weight, a_don, b_rotR, dipole, glob, logP_ow, and formal charge at pH7.4) were related to the reported localization of compounds at the four major sites (lysosomes, mitochondria, nuclei and plasma membrane). The LDA was restricted to those compounds with complete property data and that localized exclusively to one of these four

sites. Five of the seven properties (weight, a_don, b_rotR, dipole, glob) were non-negative and skewed, so they were logarithmically transformed using the function $\log_2(x+1)$. LDA was applied separately to compounds < 500 Daltons (n = 437) and > 500 Daltons (n = 332). Scatter plots of the points according to the first two discriminant directions were constructed and the points were labeled according to each subcellular localization category.

Chemical diversity analysis

The chemical structures were input into MOE to generate a Simplified Molecular Input Line Entry Specification (SMILES) strings. Next, the MACCS Structure Keys (Molecular ACCess System, a library of 166 generic chemical substructure features) was used to generate a binary fingerprint of each molecule, based on which MACCS substructure feature is present or absent in each molecule (as captured by the SMILES strings). To calculate the Tanimoto coefficient for each pair of compounds, the total number of features shared by each pair of molecules and the number of common, overlapping features present in both molecules are used, according to the equation: $Tc=\frac{c}{N1+N2-C}$, where *N1 + N2* represent the total number of unique features (bits) in the pair of molecules and *C* represents the number of unique features (bits) shared in common by the fingerprints of both molecules. Two molecules were considered as structurally similar if the *Tc* value was greater than 0.85 (Supplementary Figure 1). Average *Tc* of each sub-group of the subcellular localization dataset was calculated as the average of *Tc* values between all possible pairs of molecules present in each localization category.

Results

The physicochemical properties of compounds with reported subcellular localization features were compared with the corresponding properties of reference compounds obtained from two public repositories of small molecule information (PubChem and DrugBank databases [24, 25]; Figure 1 and 2). Relative to a random PubChem data set (Figure 1, line), compounds with reported localization properties (Figure 1, grey) were larger (e.g. higher molecular weight), possessed a broader charge distribution at physiological pH 7.4, and were more lipophilic (higher *logP_ow*). Compounds with reported localization properties also contained more hydrogen bond donors (*a_don*), more rotatable bonds (*b_rotR*, fraction of rotatable bonds) and were flatter (*glob*, or globularity, with a value of 1 indicating a perfect sphere and 0 indicating a one- or two- dimensional object) (Figure 1). Values for atom count, bond count, shape, volume and surface area-related descriptors of all localization categories were also greater than those of the reference PubChem compounds (histograms not shown).

Chemical agents with reported subcellular localization were also larger, more hydrophobic and contained more positive charges at physiological pH as compared to small molecule drugs currently on the market, represented by the DrugBank data set (Figure 2). When compared with DrugBank compounds [24, 25], compounds with reported subcellular localization possessed a more positive charge distribution at pH 7.4, higher *logP* values and higher molecular weight (Figure 2), although hydrogen donor count (*a_don*), rotatable bond fraction (*b_rotR*) and globularity (*glob*) were similar. Interestingly, while 84.7% and 71.6% DrugBank compounds conformed to Lipinski's Rule of Five or Oprea's Rule of Lead-likelihood, only 52.8% and 41.4% of compounds with known subcellular localization features conformed to the Rule of 5 [2] and the Rule of Lead-likelihood [26] (Table 1). Most of the violations of drug-likeliness or lead-likeliness tests were due to higher molecular weight and higher *logP_ow* of compounds with reported localization (data not shown). The majority of violations observed for compounds reported to localize at the plasma membrane, ER/Golgi, and cytosol.

Many compounds with reported subcellular localization were conjugated to a specific targeting motif or fluorophore, to enhance organelle-specific accumulation or to facilitate the detection of the compounds inside cells [9, 18]. Such conjugation is accompanied by an increase in molecular weight, which could impact the mechanisms of transport and accumulation inside cells. Therefore, to assess the effect of molecular weight on localization, compounds with subcellular localization information were stratified into lower and higher molecular weight groups using a molecular weight of 500 Dalton as a threshold. Compounds < 500 Daltons are more “drug-like” or “lead-like” based on Lipinski’s Rule of 5 or Oprea’s Rule of Lead-likeness, and generally lack extraneous fluorophore tags or delivery vectors. Molecules with lower molecular weight (Figure 3, grey filled line) contained less hydrogen bond donors, smaller dipole moments, lower fractions of rotatable bonds, and were less lipophilic and less globular than molecules with higher molecular weight (Figure 3, solid line).

Exploring the pH-dependent ionization states of molecules with reported subcellular localization, the overall formal charge increased from negative to positive as pH decreased, as expected from the protonation of the ionizing centers within each molecule. This trend was apparent in both low (Figure 4, grey filled line) and high (Figure 4, solid line) molecular weight compounds. Nevertheless, in most cases and especially under extreme pH conditions, higher molecular weight molecules showed a much broader distribution of formal charges than lower molecular weight compounds, reflecting the prevalence of multiple ionization centers in higher molecular weight compounds.

Other molecular properties of low and high molecular weight compounds were different, depending on the reported subcellular localizations (Table 2). Compared to larger compounds >500 Daltons, smaller compounds with reported endo-lysosomal localization were more positively charged at physiological pH, were smaller (lower *molecular weight*) and more spherical (higher *glob*). The smaller compounds with reported mitochondrial localization contained lower dipole moment (*dipole*). The smaller compounds with reported nuclear localization contained a lower fraction of rotatable bonds (*b_rotR*) and were flatter (lower *glob*). The smaller compounds with reported plasma membrane localizations were larger than non-localizing compounds but contained fewer hydrogen bond donors and were less spherical in shape.

Remarkably, for larger compounds within a given localization class, many trends observed between physicochemical properties and subcellular localizations appear reversed, when compared to the trends observed for smaller compounds (Table 2). This was especially striking in the case of molecular weight: lower molecular weight was associated with lysosomal localization for compounds <500 Daltons, while larger molecular weight was associated with lysosomal localization for compounds >500 Daltons. In addition, higher molecular weight was associated with mitochondrial, nuclear and plasma membrane localization for compounds <500 Daltons, while lower molecular weight was associated with mitochondrial and plasma membrane localization for compounds >500 Daltons. Similar molecular weight-dependent trend reversals were observed for other physicochemical properties in every localization category (Table 2).

Linear discriminant analysis was applied to find linear combinations of features which separate compounds with different reported subcellular localization sites in the endo-lysosomes, mitochondria, nucleus and plasma membrane, amongst the lower and higher molecular weight subsets (Figure 5). For compounds <500 Daltons (Figure 5, left plot), only a small portion of molecules with reported endo-lysosomal localization could be distinguished from the rest by the first and second combination of molecular properties (LDA 1 and LDA 2). These endo-lysosomal compounds possessed lower molecular weight

and lower lipophilicity (data not shown). However, these compounds were all derived from a single experimental report focusing on the pharmacological effects of closely related alkylamines [27]. For compounds >500 Daltons (Figure 5, right plot), molecules reported to localize to different subcellular compartments exhibited highly overlapping physicochemical properties.

Lastly, we confirmed that based on their chemical structure, molecules with reported subcellular localization features were reasonably diverse, irrespective of their organelle-targeting properties. The average Tanimoto coefficient (*Tc)* value is 0.350 for molecules with reported localization, which was close to the average *Tc* values of random PubChem (0.282) and DrugBank (0.292) datasets. The group of molecules with reported lysosomal localization had the lowest average *Tc* of 0.325 while the group of reported ER/Golgi localization had the highest average *Tc* of 0.438. No molecule in the database was similar to more than 24 (2.5%) molecules in the entire dataset for $Tc > 0.85$. Within each category, there were variations in terms of the similarity of the molecules to each other (Figure 6), with molecules localizing to mitochondria and lysosomes being most diverse, and molecules localizing to the ER/Golgi and plasma membrane being least diverse. This trend could reflect an intrinsic tendency for molecules possessing specific structural features to accumulate in the ER/Golgi and plasma membrane compartments, although it was also possible that this reported localization may also be biased by systematic chemical synthesis efforts of molecules incorporating specific organelle-targeting motifs.

Discussion

Knowing the bioaccumulation and biodistribution patterns of exogenous chemical agents inside cells could be useful to develop subcellular drug targeting and delivery approaches for increasing drug efficacy and decreasing toxicity. In this study, we have evaluated the relationship between the chemical structure of small molecules and the subcellular distribution patterns, based on published reports compiled from the scientific literature. In an accompanying review article, we have reviewed the evolution of the methods that have been used for performing subcellular distribution studies. Our major conclusion is that understanding of small molecule distribution inside cells has been biased by the experimental strategies that have been used for studying subcellular distribution, which have largely ignored the effect of molecular weight on the observed structure-localization relationships.

Today, fluorescence imaging constitutes the most common method used to establish the subcellular distribution of organelle-targeted small molecules. For this purpose, molecules are tagged with fluorescent probes and are studied because of their specific, organelle-targeting properties. Perhaps for this reason, molecules with known subcellular localization properties appeared to be more complex, larger, possessed many ionizable centers, and were more lipophilic as compared with references sets of molecules representing drugs currently on the market, or random samples of PubChem compounds without subcellular localization information.

As presented in the accompanying review article, there are many more reports of molecules that localize to a single organelle, as compared to reports of molecules that localize to multiple organelles. Perhaps this is because it is easier to focus analysis on localization to single organelles, but it could also be because most molecules that have been studied in terms of their localization are analyzed because of their specific targeting property. To target a single organelle, complex chemical structures with multiple functional groups may allow for strong and specific interactions with resident organelle components. Our results indicate that multiple ionizing centers are associated with larger compounds reported to accumulate

in specific organelles. While multiple ionization centers may underlie highly specific, organelle-targeting properties, high lipophilicity would be a necessary prerequisite for these molecules to penetrate inside cells. Our results also confirm that higher lipophilicity is a characteristic of compounds that have been reported to accumulate in specific organelles.

Molecular weight is an important parameter affecting transport properties and drug-likelihood [2, 26, 28] because of its direct inverse effect on diffusivity and plasma membrane permeability [29]. Using 500 Daltons as a threshold, molecular properties associated with specific subcellular compartments were identified and different trends of molecular properties distribution were observed for molecules lesser or greater than 500 Daltons. The differences in the observed trends emphasize the importance of molecular weight as a key property determining the transport mechanisms and molecular interactions affecting subcellular distribution.

In retrospect, the effect of molecular weight on the other physicochemical properties affecting localization may have been expected based on what is known about the molecular and cellular mechanisms responsible for organelle targeting and retention. For example, in the case of endolysosomal localization, the smallest molecules enter the cells and accumulate in lysosomes by passive diffusion while being retained by pH-dependent ion trapping. However, large, charged molecules enter the cells and accumulate in endolysosomes by pinocytosis or endocytosis, while being retained there by virtue of being intrinsically membrane impermeant. Similarly, flat, rigid, hydrophobic, small molecules accumulate in the nucleus by directly traversing the membranes of the nuclear envelope while being retained there by intercalating between the bases of DNA. However, larger, more globular, less membrane-permeant molecules possessing multiple positive charges may preferentially accumulate in the nucleus by entering through the nuclear pores while being retained there by forming electrostatic ion complexes with the phosphate backbone of DNA. Only in the case of the plasma membrane were our results consistent with a single common mechanism affecting localization: lipophilic partitioning of hydrophobic molecules possessing lipid-like characteristics.

Based on this meta-analysis, the ability to derive chemical-structure localization relationships of small molecules could benefit from more focused, quantitative structure-localization relationship studies performed on molecules possessing closely-related chemical structures, taking into account how transport mechanisms are molecular size-dependent. In addition, experimental analysis of nonspecific subcellular distribution patterns of compounds lacking targeting motifs should be a priority. High throughput chemical analytical techniques including chemical imaging modalities that do not rely on a fluorescence signal, such as Raman confocal microscopy, could improve understanding of the subcellular transport and distribution properties without the need of fluorescent tags for detection. Today, physiologically-based models consider log*P*, pK_a and charge as key input parameters to formulate quantitative pharmacokinetic hypothesis. Our results argue for the importance of research aimed at elucidating the effect of molecular weight (and related molecular size-dependent properties) in predictive pharmacokinetic models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by NIH grant 1RO1GM078200 to G. Rosania. H.N. Tsai was supported with funds from the Center for Molecular Drug Targeting at the University of Michigan, College of Pharmacy.

References

1. Rosania GR. Supertargeted chemistry: identifying relationships between molecular structures and their sub-cellular distribution. Curr Top Med Chem. 2003; 3:659–85. [PubMed: 12570858]
2. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001; 46:3–26. [PubMed: 11259830]
3. Feng MR. Assessment of blood-brain barrier penetration: in silico, in vitro and in vivo. Curr Drug Metab. 2002; 3:647–57. [PubMed: 12369891]
4. Rodgers T, Leahy D, Rowland M. Physiologically based pharmacokinetic modeling 1: predicting the tissue distribution of moderate-to-strong bases. J Pharm Sci. 2005; 94:1259–76. [PubMed: 15858854]
5. Rodgers T, Rowland M. Physiologically based pharmacokinetic modelling 2: predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions. J Pharm Sci. 2006; 95:1238–57. [PubMed: 16639716]
6. Berry JP, Lespinats G, Escaig F, Boumati P, Tlouzeau S, Cavellier JF. Intracellular localization of drugs in cultured tumor cells by ion microscopy and image processing. Histochemistry. 1990; 93:397–400. [PubMed: 2139017]
7. Best TP, Edelson BS, Nickols NG, Dervan PB. Nuclear localization of pyrrole-imidazole polyamide-fluorescein conjugates in cell culture. Proc Natl Acad Sci U S A. 2003; 100:12063–8. [PubMed: 14519850]
8. Croce AC, Bottiroli G, Supino R, Favini E, Zuco V, Zunino F. Subcellular localization of the camptothecin analogues, topotecan and gimatecan. Biochem Pharmacol. 2004; 67:1035–45. [PubMed: 15006540]
9. Fernandez-Carneado J, Van Gool M, Martos V, Castel S, Prados P, de Mendoza J, Giralt E. Highly efficient, nonpeptidic oligoguanidinium vectors that selectively internalize into mitochondria. J Am Chem Soc. 2005; 127:869–74. [PubMed: 15656624]
10. Ross MF, Prime TA, Abakumova I, James AM, Porteous CM, Smith RA, Murphy MP. Rapid and extensive uptake and activation of hydrophobic triphenylphosphonium cations within cells. Biochem J. 2008; 411:633–45. [PubMed: 18294140]
11. Horton KL, Stewart KM, Fonseca SB, Guo Q, Kelley SO. Mitochondria-penetrating peptides. Chem Biol. 2008; 15:375–82. [PubMed: 18420144]
12. Villa P, Sassella D, Corada M, Bartosek I. Toxicity, uptake, and subcellular distribution in rat hepatocytes of roxithromycin, a new semisynthetic macrolide, and erythromycin base. Antimicrob Agents Chemother. 1988; 32:1541–6. [PubMed: 3190183]
13. Li W, Yuan XM, Ivanova S, Tracey KJ, Eaton JW, Brunk UT. 3-Aminopropanal, formed during cerebral ischaemia, is a potent lysosomotropic neurotoxin. Biochem J. 2003; 371:429–36. [PubMed: 12513695]
14. Glaumann H, Motakefi AM, Jansson H. Intracellular distribution and effect of the antimalarial drug mefloquine on lysosomes of rat liver. Liver. 1992; 12:183–90. [PubMed: 1406082]
15. Lichtner RB, Rotgeri A, Bunte T, Buchmann B, Hoffmann J, Schwede W, Skuballa W, Klar U. Subcellular distribution of epothilones in human tumor cells. Proc Natl Acad Sci U S A. 2001; 98:11743–8. [PubMed: 11562452]
16. Cramb G. Selective lysosomal uptake and accumulation of the beta-adrenergic antagonist propranolol in cultured and isolated cell systems. Biochem Pharmacol. 1986; 35:1365–72. [PubMed: 3008762]
17. Lansiaux A, Tanious F, Mishal Z, Dassonneville L, Kumar A, Stephens CE, Hu Q, Wilson WD, Boykin DW, Bailly C. Distribution of furamidine analogues in tumor cells: targeting of the nucleus or mitochondria depending on the amidine substitution. Cancer Res. 2002; 62:7219–29. [PubMed: 12499262]
18. Burns RJ, Smith RA, Murphy MP. Synthesis and characterization of thiobutyltriphenylphosphonium bromide, a novel thiol reagent targeted to the mitochondrial matrix. Arch Biochem Biophys. 1995; 322:60–8. [PubMed: 7574695]

19. Houghton PJ, Sosinski J, Thakar JH, Boder GB, Grindey GB. Characterization of the intracellular distribution and binding in human adenocarcinoma cells of N-(4-azidophenylsulfonyl)-N′-(4-chlorophenyl)urea (LY219703), a photoaffinity analogue of the antitumor diarylsulfonylurea sulofenur. Biochem Pharmacol. 1995; 49:661–8. [PubMed: 7887981]

20. Pettersen JE, Aas M. Subcellular localization of hexadecanedioic acid activation in human liver. J Lipid Res. 1974; 15:551–6. [PubMed: 4372285]

21. Yokogawa K, Nakashima E, Ishizaki J, Maeda H, Nagano T, Ichimura F. Relationships in the structure-tissue distribution of basic drugs in the rabbit. Pharm Res. 1990; 7:691–6. [PubMed: 2395795]

22. Yokogawa K, Nakashima E, Ishizaki J, Hasegawa M, Kido H, Ichimura F. Brain regional pharmacokinetics of biperiden in rats. Biopharm Drug Dispos. 1992; 13:131–40. [PubMed: 1550908]

23. Ishizaki J, Yokogawa K, Ichimura F, Ohkuma S. Uptake of imipramine in rat liver lysosomes in vitro and its inhibition by basic drugs. J Pharmacol Exp Ther. 2000; 294:1088–98. [PubMed: 10945864]

24. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006; 34:D668–72. [PubMed: 16381955]

25. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008; 36:D901–6. [PubMed: 18048412]

26. Oprea TI. Property distribution of drug-related chemical databases. J Comput Aided Mol Des. 2000; 14:251–64. [PubMed: 10756480]

27. Seglen PO, Gordon PB. Effects of lysosomotropic monoamines, diamines, amino alcohols, and other amino compounds on protein degradation and protein synthesis in isolated rat hepatocytes. Mol Pharm. 1980; 18:468–75.

28. Ursu O, Oprea TI. Model-free drug-likeness from fragments. J Chem Inf Model. 2010; 50:1387–94. [PubMed: 20726597]

29. Balaz S. Modeling kinetics of subcellular disposition of chemicals. Chem Rev. 2009; 109:1793–899. [PubMed: 19265398]

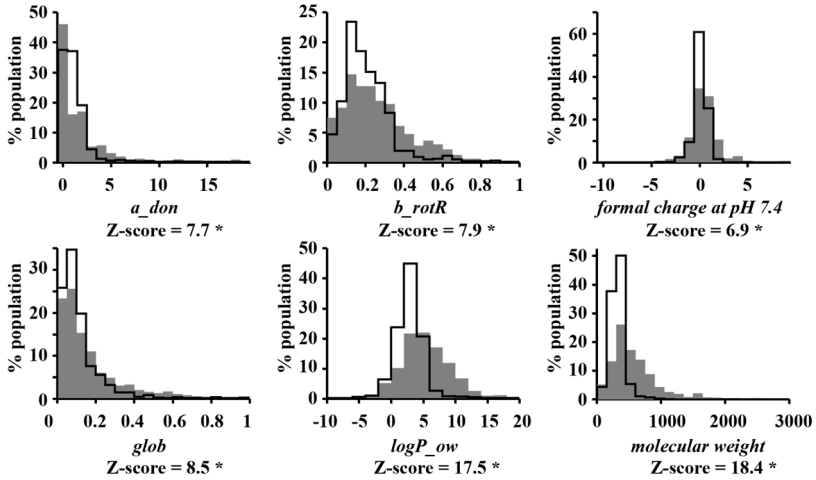

Figure 1.

Descriptor distributions of molecules with reported subcellular localization (filled gray area) and a random PubChem sample (solid line). Z-scores with an asterisk indicate a significant difference between the mean values of a descriptor in the group of compounds with reported localization and the reference PubChem dataset (p-value < 0.001). *a_don*: Hydrogen bond donor count. *b_rotR*: The fraction of rotatable bonds. *glob*: Globularity, a value of 1 indicates a perfect sphere while a value of 0 indicates a two- or one-dimensional object. *logP_ow*: Log of the octanol/water partition coefficient. *weight*: Molecular weight (including implicit hydrogens) in atomic mass units.

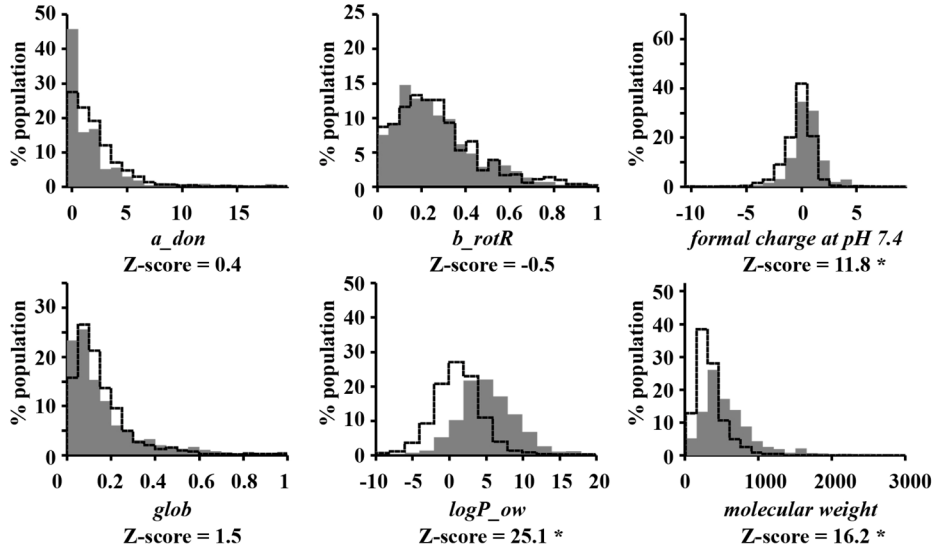

Figure 2.

Descriptor distributions of molecules with reported subcellular localization (filled gray area) and random DrugBank dataset (solid line). Z-scores with an asterisk indicate a significant difference between the mean values of a descriptor in the group of compounds with reported localization and the reference DrugBank sample (p-value < 0.001). *a_don*: Hydrogen bond donor count. *b_rotR*: The fraction of rotatable bonds. *glob*: Globularity, with a value of 1 indicating a perfect sphere and a value of 0 indicating a two- or one-dimensional object. *logP_ow*: Log of the octanol/water partition coefficient. *weight*: Molecular weight (including implicit hydrogens) in atomic mass units.

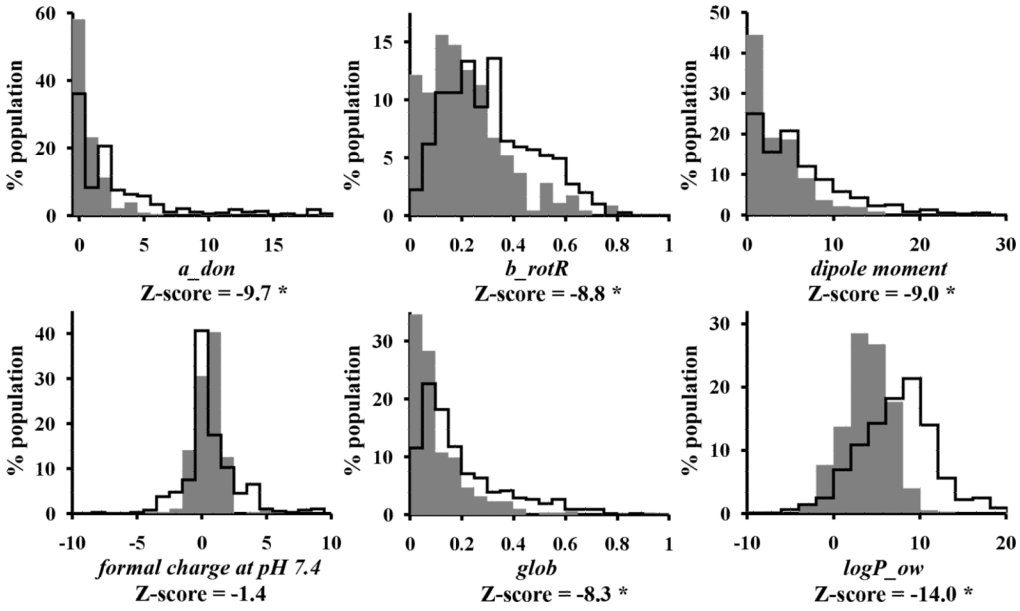

Figure 3.

Descriptor distributions of lower molecular weight (filled gray area; <500 Daltons) and higher molecular weight (solid line; > 500 Daltons) molecules with reported subcellular localization. Z-scores with an asterisk indicate a significant difference between the mean values of the descriptor in the lower and higher molecular weight groups (p-value < 0.001). *a_don*: Hydrogen bond donor count. *b_rotR*: The fraction of rotatable bonds. *dipole moment*: Dipole moment calculated from the partial charges of the molecule. *glob*: Globularity, with value of 1 indicating a perfect sphere and a value of 0 indicating a two- or one-dimensional object. *logP_ow*: Log of the octanol/water partition coefficient.

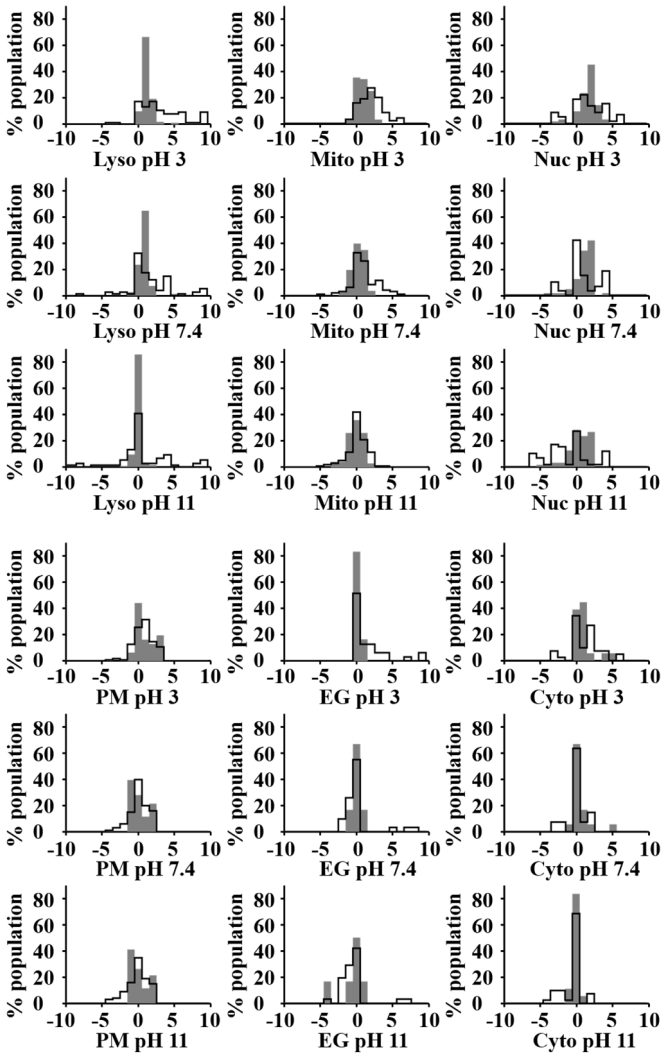

Figure 4.

Calculated, formal charge distributions of lower molecular weight (filled gray area; <500 Daltons) and higher molecular weight (solid line; >500 Daltons) compounds with reported subcellular localization, at three different pH values.

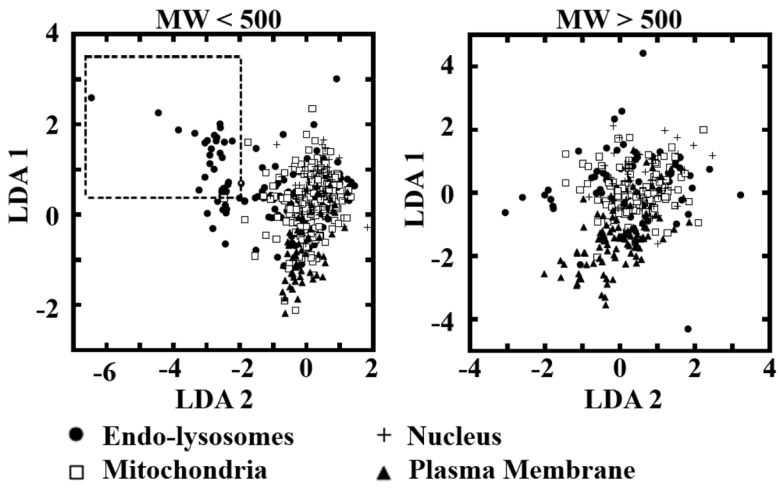

Figure 5.

Linear discriminant analysis of low (<500 Daltons) and high (>500 Daltons) molecular weight compounds with reported subcellular localizations. The axes of the plot represent linear combinations of seven molecular properties, identified using linear discriminant analysis to maximize the separation amongst the localization classes. LDA1 and LDA2 corresponded to the two, dominant linear combinations, with the “between class” variance accounting for 37% and 11% of the total variance, respectively. Additional discriminant factors (not shown) explained less than 3% of the total variance. The units on the two axes are relative and arbitrary.

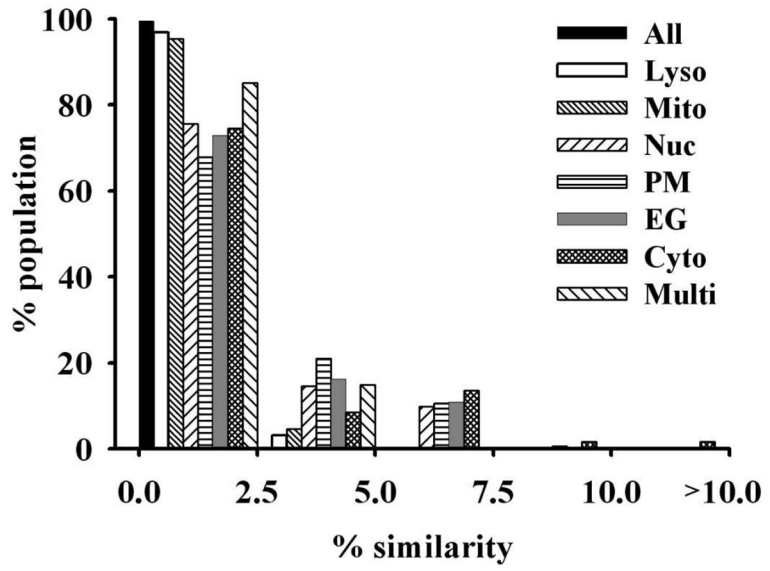

Figure 6. The major subcellular localization categories are represented by diverse subsets of molecules

In the plot, the x-axis indicates the percentage similarity threshold and the y-axis indicates the percentage of population in the group that falls between the similarity thresholds. With a *Tc* threshold of 0.85, above 65% percent of the compounds in each localization category were similar to no more than 2.5% of the subset, indicating highly diverse compounds representing each category. The relatively high percentage (21%) of PM molecules that are similar to 2.5% to 5% other molecules in PM group indicates that PM molecules are least diverse (Key: Lyso = lysosomes; Mito = Mitochondria; Nuc = nuclei; PM = plasma membrane; EG = Endoplasmic reticulum/Golgi body; Cyto = cytosolic; Multi = multiple localizations).

Table 1

Drug-likeness based on Lipinski's Rule of Five and lead-likeness based on Oprea's Rules of compounds with reported subcellular localizations. The number of drug-likely or lead-likely compounds in each location was calculated with MOE and divided by the total number of molecules in each location to calculate the percentage pass rate. The reference set of DrugBank compounds was used for comparison.

	Drug-likely		Lead-likely	
	Count	**Percentage**	**Count**	**Percentage**
Endo-lysosomes	149	65.93	134	59.29
Mitochondria	170	66.15	134	51.74
Nuclei	68	55.28	47	38.52
Plasma membrane	40	24.69	31	19.14
ER/Gogli	9	24.32	3	8.11
Cytosol	19	32.20	16	27.12
Multiple	49	48.51	35	34.65
Total	528	52.80	400	41.37
DrugBank	847	84.70	716	71.60

Table 2

Physicochemical property trends of small molecules stratified into lower (<500 Daltons) and higher (>500 Daltons) molecular weight categories, and associated with various subcellular localizations. For each localization class, Z-scores were used to compare differences in molecular properties of localizing vs. non-localizing molecules. Z-scores with an absolute value greater than 3.1 were highlighted in bold, indicating a significant trend associated with a specific localization. Z-scores with absolutes value greater than 3.1 and with a different sign from the Z-score of molecular weight were underscored. Z-scores for molecular descriptors that exhibited consistent (same sign) and significant differences between localizing vs. non-localizing compounds in both lower and higher MW groups were shaded in grey. Chemicals with reported ER/Golgi and Cyto localization were excluded from this analysis due to the small number of chemicals.

Lower MW molecules	**Number (%)**	**difference of targeting from non-targeting small molecules (Z-scores)**						
		weight	***a_don***	***b_rotR***	***charge***	***dipole***	***glob***	***logP_ow***
Endo-lysosomes	148 (65.5)	**−8.8**	−2.1	2.2	5.3	−1.4	5.3	**−8.4**
Mitochondria	164 (63.3)	**3.6**	−0.8	−2.6	**−3.1**	−3.7	−0.2	**3.1**
Nuclei	64 (52.0)	**4.6**	**4.7**	−3.9	**3.8**	−0.9	−4.7	−1.6
Plasma membrane	61 (37.7)	**6.2**	−6.9	**4.6**	−2.7	**7.0**	−8.8	**10.4**

Higher MW molecules	**Number (%)**	**difference of targeting from non-targeting large molecules (Z-scores)**						
		weight	***a_don***	***b_rotR***	***charge***	***dipole***	***glob***	***logP_ow***
Endo-lysosomes	78 (34.5)	**4.6**	**3.9**	−3.5	2.6	**3.4**	2.9	0.6
Mitochondria	95 (36.7)	**−3.8**	−0.9	−1.2	1.3	−4.5	−0.9	−1.6
Nuclei	59 (48)	0.4	1.9	−4.0	0.4	−0.7	−0.4	−5.7
Plasma membrane	101 (62.3)	**−7.0**	−9.6	8.4	**−3.9**	1.6	−3.2	8.0