DOI: 10.1002/anie.201105156

Drug Design

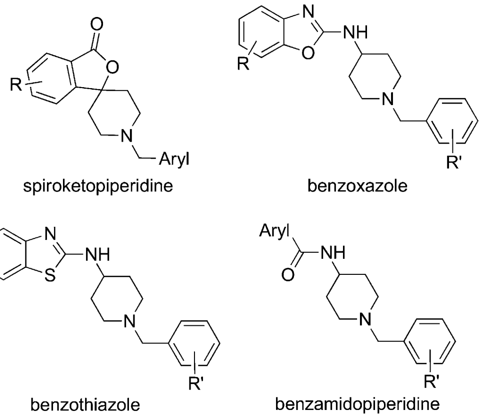
# Neighborhood-Preserving Visualization of Adaptive Structure–Activity Landscapes: Application to Drug Discovery**

*Michael Reutlinger, Wolfgang Guba, Rainer E. Martin, Alexander I. Alanine, Torsten Hoffmann, Alexander Klenner, Jan A. Hiss, Petra Schneider, and Gisbert Schneider**

Compound optimization from primary hits to pharmaceutical lead structures by organic synthesis is largely guided by the chemical feasibility and tractability of the candidate compounds, and the specific knowledge and intuition of the medicinal chemists involved. Herein we present a modeling approach that assists synthetic chemists in decision making and molecular design by visualizing and rationalizing structure–activity relationships (SAR) and structure–property relationships as SAR landscapes.

Visualization and analysis of SAR or "fitness" landscapes have been research topics in computational medicinal chemistry for approximately two decades.[1] Principal component analysis (PCA)[2] and projection to latent structures (PLS)[3] yield linear, statistically interpretable SAR models, and data projections from typically high-dimensional property spaces. As a result of the underlying mathematical models, the solutions provided by nonlinear projection are often more accurate, but also evade immediate interpretation. Despite this apparent drawback, nonlinear projection techniques like the self-organizing map (Kohonen network, SOM),[4] Sammon mapping,[5] multidimensional scaling (MDS),[6] and stochastic proximity embedding (SPE)[7]—to name just the most prominent approaches—have demonstrated their particular usefulness for SAR modeling. Their appeal lies in the ability to appropriately mirror the typically nonlinear dependencies between a structural (constitution-, topology-, conformation-based) molecular representation and some measured bioactivity or property.

We present an advanced modeling approach to SAR landscape visualization that results in easily interpretable biological response surfaces in chemical space [ligand-induced structure–activity relationship display (LiSARD)]. The LiSARD algorithm generates interactive graphics that can be used as intuitive roadmaps for molecular design and optimization. As a first practical application, we analyzed the antagonists for the human somatostatin receptor subtype 5 (hSST5R). This class A G-protein-coupled receptor is involved in several physiological processes, for example, *N*-methyl-D-aspartate (NMDA) receptor activation and control of hormonal secretion.[8] In a chemogenomics study aimed at finding nonpeptidic hSST5R antagonists, approximately 3000 compounds of which the majority belonged to four structural classes, were synthesized and tested at Roche (Scheme 1).[9]



**Scheme 1.** Scaffold classes of compounds synthesized and tested for hSST5R activity (or antagonization).

Although this application of LiSARD to a real dataset from Roche is of a retrospective nature, it serves as a proof-of-concept study to test the applicability of innovative approaches for interactive SAR visualization in medicinal chemistry.

For SAR landscape analysis, compounds need to be represented by meaningful structural attributes (descriptors) that correlate with the measured activities. In this study, we employed a topological pharmacophore representation (CATS descriptor),[10] which leads to a 150-dimensional space containing information of both molecular structure and potential ligand–receptor interaction points. Molecules that are neighbors in such a chemical space are more likely to have similar properties and activity than compounds with a large pairwise distance. This means that for dimensionality reduction and data visualization the preservation of the local neighborhood (context) might be more important than preservation of distances. We evaluated different dimensionality reduction methods (PCA, MDS, SPE, SNE)[2,6,7,11] for their ability to preserve the high-dimensional neighborhood in the low-dimensional projection for different chemical libraries (see Table S1 in the Supporting Information). Within

[*] M. Reutlinger, A. Klenner, Dr. J. A. Hiss, Dr. P. Schneider, Prof. Dr. G. Schneider
Department of Chemistry and Applied Biosciences
Wolfgang-Pauli-Strasse 10, 8093 Zurich (Switzerland)
E-mail: gisbert.schneider@pharma.ethz.ch
Dr. W. Guba, Dr. R. E. Martin, Dr. A. I. Alanine, Dr. T. Hoffmann
F. Hoffmann-La Roche Ltd., Discovery Chemistry
4070 Basel (Switzerland)

[**] This research was supported in parts by the Deutsche Forschungsgemeinschaft (DFG, FOR1406, TP4), the Swiss National Science Foundation (grant no. 205321_134783), and the OPO-Foundation Zurich.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201105156.

the given framework, SNE scored best for all calculated measures on all datasets. SNE aims at finding a projection from the original data space to the lower-dimensional embedding space such that the pairwise neighborhood distribution of points in the data and embedding space are approximately the same.

For visualization of the hSST5R compound distribution, the 150-dimensional-descriptor space was projected onto a three-dimensional (3D) space by SNE. We observed a projection trustworthiness of 96 % and continuity of 99 % for the hSST5R data.

For ease of interpretation, two-dimensional (2D) projections are preferably used for analyzing the structure of the SAR landscape, with the third dimension typically being an experimental measure of activity (e.g., $pIC_{50}$). To allow multiple views on such a landscape and reduce the risk of artifacts and potential misinterpretation, LiSARD enables manual rotation of the 3D compound cloud to select a suitable view on a 2D plane (Figure 1 a). Local activity values are computed on the fly from the available data points by Gaussian kernel regression with adaptive bandwidths and represented as a colored surface (Figure 1 b). As a result, a continuous surface is spanned over the data points. Its height corresponds to some observed variable, for example, local compound potency, and the transparency of the landscape indicates the local confidence of the model. This continuous SAR landscape representation resulted from a cooperation between the computer scientists and medicinal chemists. It offers the advantage of identifying local SARs by rotating the compound cloud and reducing the risk of over-interpretation of a static graphical model.

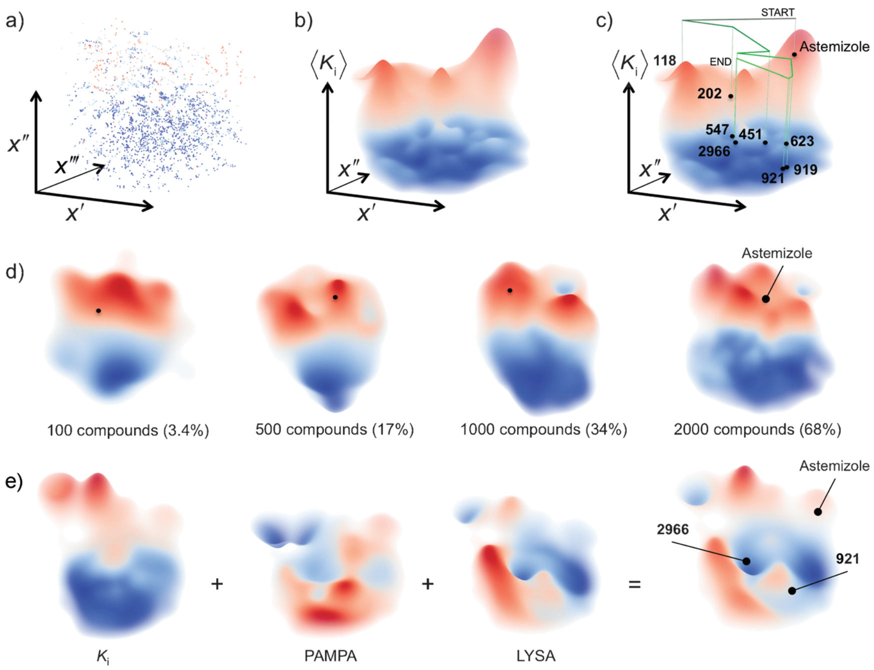


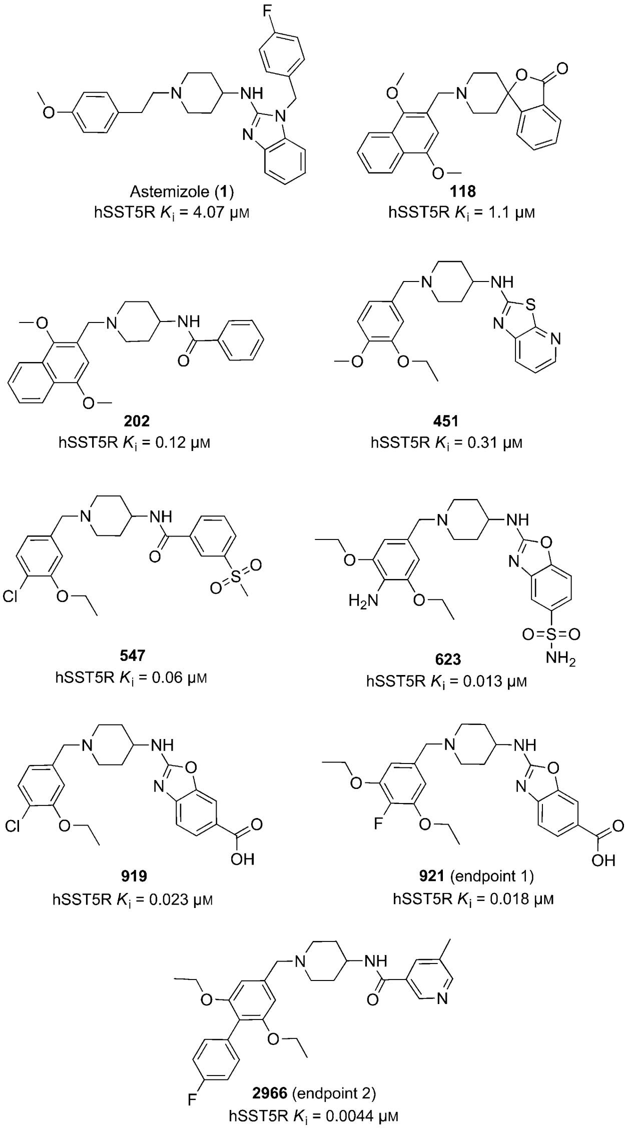
***Figure 1.*** Landscape generation and visualization by LiSARD. a) Three-dimensional compound distribution obtained by compressing a 150-dimensional compound representation (CATS pharmacophore feature vector) using stochastic neighbor embedding (SNE).[11] Dots represent compounds; blue: low $k_i$; red: high $k_i$. b) View on the hSST5R SAR landscape. Note that only two ($x'$, $x''$) of the three dimensions shown in (a) are used. Surface regions with low transparency correspond to areas of high model confidence. Coloring of the landscape is according to average local potency $\langle k_i \rangle$. c) Progress of the hSST5R project over time. The sample trajectory starts with the reference Astemizole and ends at compound **2966**. Compound numbers correspond to strategically preferred hits selected by the project team. d) Adaptive evolution of the structure–activity landscape for hSST5R agonists over the duration of the project. The snapshots contain increasing levels of detail that can be captured depending on the available number of compounds synthesized and tested. Note that active and inactive compounds contribute equally to the model (blue: low $k_i$, red: high $k_i$). e) Superimposition of different landscapes for various relevant drug properties results in a multidimensional fitness landscape for drug design. The locations of the reference Astemizole and the two endpoints of medicinal chemical optimization (**921**, **2966**) are shown. (blue: preferred regions, red: taboo regions).

In this study, experimentally determined biological activity values were added to the projected data points as the third dimension, which required interpolation between the data points. For continuous SAR estimation Bajorath and coworkers recently employed a geostatistic method called kriging.[12] As this technique turns out to be computationally demanding and a limiting factor for dynamic data visualization, we decided to implement a Kernel regression technique, which is suitable for large datasets and interactive calculation of fitness landscapes.

Visualization of the chemical space visited during a drug design project can help avoid taboo areas containing unwanted chemotypes and properties. Figure 1 c presents the trajectory of the hSST5R project progressing over time. Strategic decision points are marked, and the corresponding lead structures are presented in Scheme 2. Using a chemogenomics strategy, Astemizole (**1**) and the spiropiperidine class were identified as chemical entry points. Until then no small-molecule hSST5R antagonists were known. Both starting points evolved into the benzothiazole and the bioisosteric benzamidopiperidine series, which were optimized in parallel.

In the early project phase, compound potency and selectivity towards the related H1 receptor (Astemizole was marketed as a H1R antagonist and withdrawn later) were the driving criteria to obtain potent and selective compounds. Compound **921** (endpoint 1) represents such an intermediate candidate. This phase was followed by a multidimensional optimization strategy aiming at the best compromise between potency and physicochemical as well as pharmacokinetic parameters. Optimization finally resulted in compound **2966** (endpoint 2), which was tested in vivo.[13]

Visualization of SAR landscapes at all project stages provides an additional criterion that is based not only on the actives found so far, but equally on the inactive compounds. Figure 1 d demonstrates the adaptive nature of the SAR



***Scheme 2.*** Selected compounds and their potency on hSST5R. Higher compound index numbers indicate later stages of the project.

landscape models. Depending on the number of compounds synthesized and tested, and on the project status, increasingly fine-grained models are computed. By using two thirds of the data, the final shape of the SAR landscape (see Figure 1 b) is clearly visible. It is of note that even the first approximate landscape model computed from only 100 compounds correctly structures chemical space into desired (blue) and taboo (red) regions (Figure 1 d). Having access to such knowledge at an early stage of the project provides valuable information for hit prioritization, and helps focus on relevant areas in chemical space earlier so that optimized lead structures may be identified faster.

Monitoring the SAR landscape over the duration of the project certainly is a desirable feature for medicinal chemists to explore innovative structural variations of a chemotype and to avoid walking in circles, as well as areas with potential off-target liabilities. Multiple activities and properties can be displayed simultaneously in LiSARD, thereby enabling multidimensional optimization with the aim to avoid compounds that have an undesired pharmacological activity and property profile. Figure 1 e presents the superimposition of the landscapes for experimentally determined potency (hSST5R antagonism), membrane permeability (PAMPA), and aqueous solubility (LYSA).[14,15] Such a multidimensional fitness landscape can be readily obtained by adding up the individual landscape functions and subsequently rescaling the $z$ axis to obtain pseudoprobabilities.

In this proof-of-concept study, we have demonstrated that a dynamic view on adaptive SAR landscapes can support molecular design by providing project-specific visual aids for compound prioritization. Potential compound liabilities can be avoided, multiple properties can be considered at a time, and the information contained in both active and inactive compounds is optimally exploited for early hit prioritization and progress monitoring.

## Experimental Section

Compound structures were standardized using the "wash" function in MOE v2010.10 (The Chemical Computing Group Inc., Montreal, Canada). Properties were computed with MOE. CATS descriptors were computed using the speedcats software (0–9 bonds, type-sensitive scaling), as described.[10]

Stochastic neighbor embedding (SNE) defines two conditional probabilities: 1) $p_{ij}$, the probability that a data point $\xi_i$ has $\xi_j$ as its neighbor [Eq. (1)], and 2) $q_{ij}$, the induced probability that point $i$ picks point $j$ as its neighbor as a function of the low-dimensional images $x_i$ of all data points $\xi_i$ [Eq. (2)]. The cost function minimized in the embedding is a sum of the Kullback-Leibler ($KL$) divergences between the original ($p_{ij}$) and induced ($q_{ij}$) distributions [Eq. (3)].

$$p_{ij} = \exp\left(-\frac{\|\xi_i-\xi_j\|^2}{2\sigma_i^2}\right) \bigg/ \sum_{k\neq i} \exp\left(-\frac{\|\xi_i-\xi_k\|^2}{2\sigma_i^2}\right), \text{ where} \tag{1}$$

$$q_{ij} = \exp\left(-\|x_i-x_j\|^2\right) \bigg/ \sum_{k\neq i} \exp\left(-\|x_i-x_k\|^2\right) \tag{2}$$

$$C = \sum_i \sum_j p_{ij} \log\frac{p_{ij}}{q_{ij}} = \sum_i KL(P_i\|Q_i) \tag{3}$$

$\sigma_i$ was chosen by a binary search, such that the entropy of the distribution over neighbors was equal to $\log k$, where $k$ is the number of local neighbors or "perplexity".

We used the implementation of SNE from the Matlab Toolbox for Dimensionality Reduction v0.7.2[16] and Matlab 7.10.0 (The MathWorks Inc., Natick, USA).

The Nadaraya–Watson estimator was applied to fitting a surface to the projected data points.[17] The value for an unobserved location is estimated as a locally weighted average of the given data, using a kernel as weighting function. For a set of $n$ observations ($\mathbf{x}_i$,$y_i$) with $\mathbf{x} \in \mathbb{R}^2$ and $y \in \mathbb{R}$ the Nadaraya–Watson estimator is defined as given in Equation (4).

$$\begin{aligned} \hat{m}(\mathbf{x},\mathbf{h}) &= \sum_{i=1}^{n} \kappa_{\mathbf{h}}(\mathbf{x}-\mathbf{x}_i)y_i \bigg/ \sum_{i=1}^{n} \kappa_{\mathbf{h}}(\mathbf{x}-\mathbf{x}_i), \text{ where} \\ \kappa_{\mathbf{h}}(\mathbf{x}) &= \frac{1}{h_1 h_2}\kappa\left(\frac{x_1}{h_1},\frac{x_2}{h_2}\right) \end{aligned} \tag{4}$$

Here, $\mathbf{h} = (h_1,h_2)$ is a vector of bandwidths, and $\kappa(x)$ a multivariate kernel function, for which we used the multivariate Gaussian kernel

[Eq. (5)].

$$\kappa_{\mathbf{H}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{H}|^{-\frac{1}{2}}} e^{-\frac{1}{2}x^{T}H^{-1}x} \quad (5)$$

The bandwidth matrix **H** is defined as $\mathbf{H} = diag\left(h_1^2, h_2^2\right)$.[18] The optimal bandwidth $h_1$,$h_2$ was estimated from the data according to the normal reference rule [Eq. (6)].[19] As data points are in general unevenly distributed, a fixed bandwidth only represents a compromise for both densely and sparsely populated areas. We thus combined the local density adaptive bandwidth estimator with the Nadaraya–Watson approach to obtain the local bandwidth h(**x**) [Eq. (6)].[20,21]

$$\mathrm{h}(\mathbf{x}) = k\left(\sqrt{\sum_{i=1}^{n} \kappa_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}\right)^{-1} \quad (6)$$

The factor $k$ defines the degree of smoothing. We used $k = 1$. The bandwidth varies with the estimation position, inversely proportional to local data density. All surface interpolations were calculated in our visualization software LiSARD, which was implemented in Java SE 6 (Oracle Corporation, Redwood Shores, USA). For license requests contact the authors.

Received: July 22, 2011
Published online: October 7, 2011

**Keywords:** chemical space · computational chemistry · drug discovery · medicinal chemistry · structure–activity relationships

[1] a) G. Schneider, J. Schuchhardt, P. Wrede, *Biol. Cybern.* **1995**, *73*, 245–254; b) H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205–1213; c) B. Bienfait, J. Gasteiger, *J. Mol. Graphics Modell.* **1997**, *15*, 203–215; d) T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157–166; e) G. Schneider, *Curr. Med. Chem.* **2002**, *9*, 2095–2101; f) R. Guha, J. H. Van Drie, *J. Chem. Inf. Model.* **2008**, *48*, 646–658; g) G. Schneider, M. Hartenfeller, M. Reutlinger, Y. Tanrikulu, E. Proschak, P. Schneider, *Trends Biotechnol.* **2009**, *27*, 18–26; h) M. T. Sisay, L. Peltason, J. Bajorath, *J. Chem. Inf. Model.* **2009**, *49*, 2179–2189; i) P. Iyer, Y. Hu, J. Bajorath, *J. Chem. Inf. Model.* **2011**, *51*, 532–540; j) R. Guha, *Methods Mol. Biol.* **2011**, *672*, 101–117.

[2] K. Pearson, *Philos. Mag.* **1901**, *2*, 559–572.

[3] L. Ståhle, S. Wold, *J. Pharmacol. Methods* **1986**, *16*, 91–110.

[4] a) T. Kohonen, *Biol. Cybern.* **1982**, *43*, 59–69; b) J. Gasteiger, J. Zupan, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinheim, **1999**; c) P. Schneider, Y. Tanrikulu, G. Schneider, *Curr. Med. Chem.* **2009**, *16*, 258–266.

[5] a) W. Sammon, *IEEE Trans. Comput.* **1969**, *C-18*, 401–409; b) G. Schneider, S.-S. So, *Adaptive Systems in Drug Design*, Landes Bioscience, Georgetown, **2001**; c) S. Ekins, K. V. Balakin, N. Savchuk, Y. Ivanenkov, *J. Med. Chem.* **2006**, *49*, 5059–5071.

[6] W. S. Torgerson, *Psychometrika* **1952**, *17*, 401–419.

[7] a) D. K. Agrafiotis, H. Xu, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 475–484; b) D. K. Agrafiotis, H. Xu, F. Zhu, D. Bandyopadhyay, P. Liu, *Mol. Inf.* **2010**, *29*, 758–770.

[8] a) T. Reisine, G. I. Bell, *Endocr. Rev.* **1995**, *16*, 427–442; b) G. Ösapay, K. Ösapay, *Expert Opin. Ther. Pat.* **1998**, *8*, 855–870.

[9] a) R. E. Martin, L. G. Green, W. Guba, N. Kratochwil, A. Christ, *J. Med. Chem.* **2007**, *50*, 6291–6294; b) W. Guba, L. G. Green, R. E. Martin, O. Roche, N. Kratochwil, H. Mauser, C. Bissantz, A. Christ, M. Stahl, *J. Med. Chem.* **2007**, *50*, 6295–6298; c) R. E. Martin, P. Mohr, H. P. Maerki, W. Guba, C. Kuratli, O. Gavelle, A. Binggeli, S. Bendels, R. Alvarez-Sánchez, A. Alker, L. Polonchuk, A. D. Christ, *Bioorg. Med. Chem. Lett.* **2009**, *19*, 6106–6113; d) A. Alker, A. Binggeli, A. D. Christ, L. Green, H. P. Maerki, R. E. Martin, P. Mohr, *Bioorg. Med. Chem. Lett.* **2010**, *20*, 4521–4525.

[10] a) G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem.* **1999**, *111*, 3068–3070; *Angew. Chem. Int. Ed.* **1999**, *38*, 2894–2896; b) U. Fechner, G. Schneider, *QSAR Comb. Sci.* **2004**, *23*, 19–22.

[11] G. Hinton, S. Roweis in *Advances in Neural Information Processing Systems* (NIPS) **2002**, *15*, 857–864.

[12] L. Peltason, P. Iyer, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50*, 1021–1033.

[13] U. Sprecher, P. Mohr, R. E. Martin, H. P. Maerki, R. Alvarez Sánchez, A. Binggeli, B. Künnecke, A. D. Christ, *Regul. Pept.* **2010**, *159*, 19–27.

[14] M. Kansy, F. Senner, K. Gubernator, *J. Med. Chem.* **1998**, *41*, 1007–1010.

[15] a) J. Alsenz, M. Kansy, *Adv. Drug Delivery Rev.* **2007**, *59*, 546–567; b) J. U. Peters, P. Schnider, P. Mattei, M. Kansy, *ChemMedChem* **2009**, *4*, 680–686.

[16] L. J. P. van der Maaten, E. O. Postma, H. J. van den Herik, *Dimensionality Reduction: A Comparative Review*, Tilburg University Technical Report, TiCC-TR 2009–005, **2009**.

[17] a) A. E. Nadaraya, *Theory Probab. Its Appl. (Engl. Transl.)* **1964**, *9*, 141–142; b) G. S. Watson, *Sankhya Ser. A* **1964**, *26*, 359–372.

[18] M. P. Wand, M. C. Jones, *J. Am. Stat. Assoc.* **1993**, *88*, 520–528.

[19] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York, **1992**.

[20] S. R. Sain, *Adaptive Kernel Density Estimation*, Ph.D. Dissertation **1994**, Department of Statistics, Rice University.

[21] D. O. Loftsgaarden, C. P. Quesenberry, *Ann. Math. Stat.* **1965**, *36*, 1049–1051.