

50376
1996
282

THÈSE

présentée à

L'Université des Sciences et Technologies de Lille

pour obtenir le titre de

DOCTEUR EN CHIMIE

spécialité: Chimie organique et macromoléculaire

par

Dragos Horvath

**Les modèles de solvant continu. Applications à la
modélisation moléculaire de l'inhibition de
l'enzyme parasitaire Trypanothione Reductase.**

Date de soutenance: 26.11.1996

Président:	F. Wallart
Rapporteurs:	S.J. Wodak T. Simonson
Directeur du travail:	A. Tartar
Membres:	K. Sharp D. van Belle I. Silaghi-Dumitrescu H. Pauwels

SCD LILLE 1



D 030 213872 5

50376
1996
282

Jan 2000 7.02

Remerciements

Ce chapitre est complètement redondant et inutile. Non, il ne s'agit pas du fait qu'il n'ait personne à remercier, mais c'est tout simplement que dans l'excellente atmosphère de travail dans laquelle j'ai eu la chance de "respirer", les relations d'amitié et complicité établies au SCBM de Pasteur Lille aussi bien qu'à l'Université Libre de Bruxelles ont fait de la télépathie un instrument beaucoup plus approprié pour transmettre des remerciements qu'une simple feuille de papier.

J'ajouterais néanmoins un chapitre de remerciements, parce que la vie est faite de choses souvent redondantes et inutiles. Here we go:

- en premier mon "chef", Monsieur le Professeur André Tartar pour son soutien. Ces remerciements peuvent être interprétés comme une conséquence de la bonne éducation dont j'ai bénéficié pleinement pendant ma jeunesse derrière le "rideau de fer". Néanmoins, je le fais sincèrement (tout bon communiste aurait dit exactement ça, n'est-ce pas??)
- à Daniel van Belle, pour tout. However, the previous statement will not be detailed here in order to ensure a decent size of this chapter.
- for Shoshana: GO TO previous paragraph.
- à Eric Buisine, mon premier guide et aide pour la période où je ne parlais pas encore le "fy_biosym".
- à Guy Lippens, pour son sain scepticisme de physicien par rapport à mes interminables "fits". Il est l'auteur de la célèbre phrase "avec 70 paramètres, tu peux fitter un chameau".



- un grand merci pour toute l'équipe RMN de Lille (Céline, Christophe, Benoît O., Jean-Michel, Gérard). L'expression matérielle de nos "remerciements" réciproques peut être admirée sur l'armoire de la salle RMN.
- merci aux chimistes (combinatoires ou pas) du labo (Benoît D., Xavier, Patricia, Oleg, Béatrice, Iuliana, Christophe et tous les autres), pour m'avoir rappelé de temps en temps que les molécules sont des choses réelles. En particulier, merci à Benoît pour son intérêt constant pour tout ce qui est modélisation moléculaire - c'est réconfortant de ne pas se sentir le seul fou virtuel du labo.
- pour obtenir les noms de tous mes amis de l'Université Libre de Bruxelles auxquelles je remercie, il faut se connecter au <http://www.ulb.ac.be/>.
- en dehors de la réalité virtuelle, il y a néanmoins une réalité réelle habitée par des chimistes expérimentalistes. Un grand merci à toute l'équipe du Professeur Christian Sergheraert, pour sa très bonne collaboration entre le réel et le virtuel.
- merci à Elisabeth Davioud et. al. pour avoir cru à l'oracle des prédictions d'affinité des inhibiteurs et d'avoir pris le temps de les vérifier. Valérie, merci pour avoir mesuré les pourcentages d'inhibition.
- fuer Inge Jekeli, Author und Verbreiter der Theorie de "Inneren Freiheit & Distanz".
- pentru toti dascalii de la facultatea de chimie de la Cluj.
- pentru Mirmiriltza, dupa care ma topesc.
- je remercie le Conseil Régional du Nord et la société CEREP pour m'avoir soutenu financièrement pendant mes années de thèse. J'espère que cette thèse récompensera la confiance que vous m'avez témoignée.
- enfin, merci à tout le personnel administratif "mutant" qui m'a vraiment aidé ou qui au moins n'a pas compliqué inutilement ma vie. Ceci inclut Carole, pour

laquelle la différence entre le nombre de cas où elle m'a aidé et le nombre de fois où elle m'a engueulé reste supérieur à zero. On peut donc en déduire que le nombre de fois ou elle m'a aidé doit être véritablement astronomique. Merci beaucoup à l'équipe des Ressources Humaines de l'Institut qui m'a aidé dans des nombreuses démarches administratives.

- En ce qui concerne les bureaucrates de toute sorte (espèce qui, ma foi, n'est pas en voie de disparition en Europe), il leur revient de me remercier pour m'être contenté de ne pas le "remercier" d'une façon mettant clairement en évidence les richesses cachées d'une grande variété de langues du vieux continent.

Introduction générale

Parmi d'autres très récents domaines interdisciplinaires situés à la frontière entre la biologie, la chimie et l'informatique, la modélisation moléculaire constitue un outil important permettant l'étude approfondie des mécanismes d'action des molécules biologiquement actives. Dans ce contexte, la modélisation assistée par ordinateur des interactions entre les enzymes et leurs inhibiteurs ouvre la porte à des méthodes « rationnelles » de recherche de nouveaux médicaments¹ (« Rational Drug Design »). Ce nouveau domaine de recherche, fondamental mais aussi appliqué, se trouve aujourd'hui en plein essor. Bien que la puissance toujours croissante des ordinateurs modernes permette la modélisation de phénomènes de plus en plus complexes, la modélisation moléculaire nécessite encore un important travail de développement de stratégies et d'algorithmes plus « intelligents » de façon à minimiser l'effort calculatoire tout en améliorant la précision des résultats obtenus.

A ce jour, la modélisation moléculaire n'a pas atteint la précision ou la rapidité qui lui permettraient de se « substituer » aux mesures expérimentales². Cela n'a rien d'étonnant, étant donnée la complexité des phénomènes biologiques. Cependant, le nombre d'applications pour lesquelles les résultats des simulations concordent avec les valeurs expérimentales ne cesse d'augmenter. De plus, il ne faut pas perdre de vue que même des résultats semiquantitatifs ou des indications

¹ C. Hansch, "Comprehensive Medicinal Chemistry. Vol. 4: Quantitative Drug Design", Pergamon Press (1990)

² W.F. van Gunsteren, "Methods of calculating free energies and binding constants. Successes and problems" in "Computer Simulations of Biomolecular Systems", ESCOM, pp. 27-59 (1989)

qualitatives déduites d'une simulation moléculaire peuvent être très utiles au scientifique expérimentaliste.

D'autre part, le développement de nouveaux algorithmes de simulation, apportent de nouvelles perceptions du monde moléculaire. L'approximation des « champs de forces consistants » qui sera discutée en détail plus loin³ repose sur une description alternative, moins exacte, mais beaucoup plus intuitive des molécules par rapport au formalisme quantique. La question « Jusqu'où peut-on aller avec une description classique des molécules ? » n'a pas pour l'instant reçue aucune réponse claire. Même si "tout" peut être modélisé en choisissant un nombre approprié de paramètres ajustables dans les simulations, il n'est pas évident d'évaluer la généralité ou la solidité des fondements physiques d'un tel modèle. Il est donc essentiel de maintenir un équilibre entre d'une part, l'utilisation des paramètres ajustables permettant d'optimiser la concordance des résultats de la simulation par rapport à l'expérience, et d'autre part, le souci de ne pas trop s'écarter des bases physiques du modèle par l'introduction de telles « corrections empiriques ».

L'un des objectifs des travaux présentés dans la première partie de ce mémoire est l'étude théorique des phénomènes de solvatation⁴, un facteur très important contrôlant une grande variété de processus biologiques, notamment l'inhibition des enzymes par des ligands, ce qui fera l'objet de la deuxième partie de cette thèse.

³ Voir l'introduction dédié à ce sujet

⁴ T.E. Creighton, *Curr. Opin. Struct. Biol.* 1,5 (1991); D. Bashford, *Curr. Opin. Struct. Biol.* 1,175 (1991)

Les modèles de solvant continu⁵ traitent le solvant comme un milieu astructuré, homogène et possédant des propriétés diélectriques responsables des effets de polarisation subis par le solvant sous l'influence du champ électrique créé par la distribution de charges de la molécule de soluté. Notre étude des modèles de solvant continu s'est concrétisée par:

A: le développement d'algorithmes efficaces, résolvant l'équation différentielle de Poisson et permettant d'évaluer l'énergie de polarisation du solvant par le soluté. Un algorithme basé sur la méthode des éléments aux frontières (« boundary element method ») a fait l'objet d'une publication:

D. Horvath, D. van Belle, G. Lippens & S.J. Wodak « Development and Parametrization of Continuum Solvent Models. I. Models based on the Boundary Element Method », *J. Chem. Phys*, 104, 6679 (1996)

Le développement et la généralisation (en dehors du cadre de cette thèse) de cette méthode afin d'obtenir un potentiel de solvation utilisable dans les simulations de Dynamique Moléculaire a donné lieu à des communications à différentes conférences internationales:

-ECCC1- *Computational Chemistry*, Nancy (France), May 1994 (poster)

**-*Workshop on Protein-Ligand Interactions*, Heidelberg (BRD), April 1995
(présentation orale)**

⁵ Voir l'introduction dédié à ce sujet

-Gordon Research Conference on Computational Chemistry, New Hampton, New Hampshire (USA), July 1996 (poster)

B: l'étude des modèles continus empiriques, basés sur des équations simplifiées. Un cadre théorique unitaire pour l'analyse de tels modèles a été proposé et les performances de plusieurs techniques analogues ont été comparées. Ceci a fait l'objet d'une deuxième publication:

D. Horvath; D. van Belle; G. Lippens; « Développement and parametrization of continuum solvent models. II. An unified approach to the solvation problem. »
J.Chem.Phys. 105, 4197 (1996)

Les modèles de solvant continu mis au point ont ensuite été utilisée afin d'évaluer l'influence des effets de solvant sur l'affinité d'un ligand pour le site enzymatique de la Trypanothione Reductase⁶ (TR), une enzyme des parasites *Trypanosomatidae*, dont l'inhibition sélective sans interférer avec le métabolisme de l'espèce hôte (l'homme) peut s'avérer très efficace dans la lutte contre certaines maladies parasitaires comme la maladie du *Chagas*, qui affecte les pays du tiers monde⁷.

La simplicité des modèles de solvant empiriques constitue un atout essentiel pour leur application dans les simulations moléculaires. Une étude des relations entre la structure des inhibiteurs synthétisés et leur affinité et faisant appel au

⁶ E.A.O Etah, K. Smith, A.H. Fairlamb, "Trypanothione detoxication systems in trypanosomatids", Spring Meeting of the British Society for Parasitology, London (1993)

⁷ S.L. DeCastro, *Acta Trop.*, 53,83 (1993); M. Moutiez, *Thèse doctorale*, Université de Lille II (1995)

traitement « continu » du solvant montre clairement l'applicabilité de cette approche aux problèmes concrets de pharmacochimie (cette publication *ne fait pas partie de cette thèse*):

Baillet, S.; Buisine, E.; Horvath, D.; Maes, L.; Bonnet, B.; Sergheraert, C. « 2-Aminodiphenylsulphides as inhibitors of trypanothione reductase. » *Bioorg. Med. Chem.*, 6, 891 (1996)

D'autres études de modélisation de molécules inhibitrices de la Trypanothione Reductase ont fait l'objet de communications scientifiques:

- *Journées du Groupe de Graphisme Moléculaire, Nouan-Le-Fuzelier (France), 17-19 mai 1993 (poster)*

- *Secondes Conférences Européennes du Groupement des Pharmacochimistes de l'Arc Atlantique, Caen (France) 30 juin - 2 juillet 1993 (poster)*

- *BSP Trypanosomiasis and Leishmaniasis Seminar, Glasgow (UK), 3-6 september 1995 (poster).*

La dernière partie de cette thèse présente un modèle plus général permettant la prédiction de l'affinité d'un ligand pour la Trypanothione Reductase. Le modèle a été conçu dans le but d'estimer rapidement les affinités d'un grand nombre de composés pour cette enzyme. Cette approche repose essentiellement sur un algorithme destiné à déterminer les géométries des principaux conformères des

ligands et un algorithme de « docking »⁸, évaluant les interactions enzyme-ligand par rapport à l'orientation relative du ligand et de l'enzyme et estimant l'affinité d'un ligand pour l'enzyme en fonction de l'intensité de ces interactions.

Le développement d'un algorithme de construction des structures moléculaires et d'exploration conformationnelle automatique a été nécessaire afin d'obtenir de manière rapide des géométries correctes pour les molécules de ligands dont la structure tridimensionnelle n'est pas disponible ou n'a pas été déterminée expérimentalement.

L'algorithme de prédiction de l'affinité évalue les contributions des principales interactions (électrostatique, hydrophobe, van de Waals...) intervenant lors de la fixation du ligand à l'enzyme et impliquant également le solvant. L'un des plus simples modèles de solvant empiriques⁹ étudiés préalablement a été utilisé pour quantifier la contribution des effets de solvant. L'applicabilité de ce modèle a pu être toutefois validée par comparaison à des résultats obtenus à partir des modèles de solvant plus « précis », mais trop onéreux en temps calcul pour pouvoir être directement utilisés dans le cadre de l'algorithme de prédiction des affinités.

L'application du modèle de prédiction d'affinité au « criblage virtuel » d'une banque de données moléculaires a abouti à l'identification de nouveaux composés inhibiteurs de l'enzyme Trypanothione Reductase. Ceci est décrit dans la publication suivante:

⁸ Ajay, M. Murcko; *J. Med. Chem.*, 38,4953 (1995)

⁹ M.K. Gilson, B. Honig, *J. Comp.-Aided Mol. Des.*, 5,5 (1991)

D. Horvath, « A Virtual Screening Approach Applied to the Search of Trypanothione Reductase Inhibitors », *accepted for publication in J. Med. Chem (1996)*.

Cette thèse s'organise donc principalement autour des trois publications mentionnées antérieurement, auxquelles s'ajoutent deux introductions décrivant la problématique générale des modèles de solvant continus et la théorie générale des simulations moléculaires basées sur les champs de forces empiriques. Chaque chapitre est indépendant, avec sa propre numérotation de pages et sa propre liste bibliographique.

Introduction: Modélisation de la solvatation. Modèles "continus" du solvant . L'électrostatique des diélectriques

A. Modélisation de la solvatation. Modèles "continus" du solvant.

Ce chapitre est destiné à introduire et à clarifier certains aspects essentiels des modèles continus¹ de solvant. Ceux-ci sont basés sur l'hypothèse que le soluté est un *corps homogène, faiblement diélectrique* ($\epsilon_{\text{int}}=2\dots 4$), possédant une *distribution de charges* atomiques ponctuelles, et entouré par un *milieu extérieur homogène* (le solvant), caractérisé par la valeur ϵ_{ext} de sa *constante diélectrique*, à laquelle on peut ajouter des paramètres de "tension superficielle" effective pour décrire les effets non-électrostatiques de la solvatation. L'interface diélectrique soluté (ϵ_{int}) - solvant (ϵ_{ext}) dépend de la géométrie du soluté et des valeurs données aux rayons atomiques. Autrement dit, la structure moléculaire discontinue du solvant est ignorée dans l'approche "continue" de la solvatation. On suppose que la réorientation rapide des dipôles du solvant implique un "moyennage" des interactions soluté-solvant et que ce "potentiel de force moyenne" peut être expliqué en se référant uniquement aux propriétés diélectriques du solvant.

Il est hasardeux de parler des propriétés diélectriques à "l'intérieur" d'une molécule, étant donné que, par définition, la constante diélectrique est une grandeur macroscopique qui reflète la réponse d'un grand *ensemble de molécules* soumises à l'action d'un champ électrique extérieur. Cela comprend les restructurations des nuages électroniques (*polarisation électronique*; apparition des dipôles induits) aussi bien que les réorientations spatiales des dipôles permanents (*polarisation d'orientation*). La constante diélectrique ϵ_{int} à l'intérieur du soluté doit être perçue de façon formelle, afin de tenir compte des effets de polarisation *électronique* du soluté dans les modèles continus. Ce paramètre est différent du ϵ effectif du composé pur, qui inclut aussi la contribution de la polarisation d'orientation.

En ce qui concerne le solvant, les deux contributions (électronique et d'orientation) sont prises en compte dans le modèle continu. Le solvant se

¹ C.J. Cramer & D.G. Truhlar, "Continuum solvent models: Classical and Quantum Mechanical implementations", *Rev. Comp. Chem.* 6,1 (1995)

comporte comme un diélectrique par rapport au champ électrique généré par les charges des atomes du soluté, qui va simultanément induire des déplacements des électrons dans les orbitales des molécules de solvant et va essayer d'orienter les dipôles moléculaires permanents en s'opposant à l'agitation thermique. Ce sont les mêmes phénomènes qui se déroulent lors d'une mesure macroscopique de la constante diélectrique du solvant, sauf que dans ce cas, le champ électrique perturbant est introduit par l'expérimentateur et non par les charges d'une molécule solvatée. En conséquence, on peut supposer en première approximation que la valeur ϵ_{ext} caractérisant le solvant dans des modèles continus doit être égale à la valeur macroscopique de la constante diélectrique du solvant pur. En réalité, il faudrait prendre en compte le fait que le champ électrique produit localement par une molécule de soluté affectera uniquement les molécules de solvant proches de celui-ci et non pas l'ensemble du solvant. Ces molécules voisines peuvent avoir un comportement différent de celles dans le solvant pur. L'influence du soluté ne se limite pas uniquement à des effets de nature purement électrostatique, mais peut influencer d'autres façons la mobilité des dipôles du solvant. Sous l'influence d'un champ électrique important, des effets de saturation et de compression diélectriques affecteront la valeur de la constante diélectrique apparente. De tels phénomènes ont été empiriquement pris en compte lors de la calibration des paramètres intervenant dans les modèles présentés dans les deux articles suivants.

Ceci montre qu'une bonne compréhension de la nature des effets diélectriques est nécessaire pour connaître les limites de l'applicabilité de l'approche "continue" de la solvation. Celle-ci se base sur une modélisation macroscopique inspirée par l'électrostatique des diélectriques, mais les fondements moléculaires qui permettent d'avancer l'hypothèse d'un solvant "nonstructuré" et homogène ne peuvent être perdus de vue. Dans le chapitre suivant, l'aspect mixte -microscopique et macroscopique- des phénomènes diélectriques seront discutés brièvement, en accordant toujours une priorité au traitement intuitif et physique par rapport à la rigueur mathématique.

Les effets de solvation de nature non-électrostatique² comprennent l'effet de "cavitation" (énergie nécessaire pour la création d'une cavité capable d'accueillir le soluté dans le solvant), l'interaction de van der

² B.K. Lee, *Biopolymers*, 31,993 (1993)

Waals entre le soluté et le solvant et l'influence du soluté sur la structuration du solvant interfacial. L'ensemble de ces phénomènes définit l'effet "hydrophobe", c'est-à-dire la tendance des groupements hydrocarbonés ou aromatiques de s'associer dans l'eau en minimisant ainsi la surface de contact avec le milieu aqueux. En fait, ce phénomène n'est pas connu en détail, mais est traité empiriquement dans les modèles continus, en supposant que l'énergie libre correspondante varie proportionnellement à la surface moléculaire accessible au solvant. Evidemment, les contributions de cavitation ou van der Waals sont peu dépendantes des propriétés électrostatiques locales de la surface moléculaire, et peuvent donc être facilement décrites par des termes proportionnels à la surface moléculaire totale. Cependant, il n'est pas clairement démontré que les effets entropiques causés par la perturbation de la structure locale du solvant par le soluté sont plus importants autour des surfaces apolaires. Le modèle d' "iceberg"³, c'est-à-dire, une perte d'entropie du solvant lors du transfert de chaînes hydrocarbonées dans l'eau a suscité des nombreux débats. Nos travaux sont basés sur une description utilisant un terme proportionnel à la *surface moléculaire totale* plus un terme ne tenant compte que de la *surface des groupements apolaires*, sans toutefois approfondir la modélisation des processus physiques impliqués. Nos résultats nous ont néanmoins suggéré une définition "généralisée" de la polarité locale de la surface moléculaire en fonction de la densité de charge de polarisation en chaque point de l'interface diélectrique. Par la suite, nous avons proposé un modèle alternatif et obtenu des résultats encourageants en "*modulant*" en fonction de cette mesure de polarité, la contribution de chaque élément de surface dans les termes décrivant les effets non-électrostatiques de la solvation.

La facilité principale offerte par des modèles continus de solvant vient de l'applicabilité directe du formalisme bien établi de l'électrostatique classique⁴ pour décrire les interactions de nature électrostatique entre un soluté et son solvant. Pour cette raison, ce type d'approche a été une des premières appliquées pour comprendre les phénomènes de solvation⁵. Cependant, malgré les importantes simplifications à la base de ces méthodes, le traitement électrostatique reste assez élaboré même pour des molécules de petite taille (sauf pour des cas très simples comme les ions sphériques) et demande une puissance de calcul qui n'était pas accessible avant le

³ R.M. Jackson & M.E.J. Sternberg, *Protein Engineering*, 7,371 (1994)

⁴ J.D. Jackson, *Classical Electrodynamics*, J. Wiley & Sons, NY., (1975)

⁵ M. Born, *Z. Physik* 1,45 (1920)

développement de l'informatique moderne. Ces approches ont donc connu une période de "disgrâce". Les échecs des essais effectués afin de caractériser quantitativement les effets de solvant en appliquant la méthode "continue" ont été exagérés et attribués à l'inexactitude de l'hypothèse considérant que la solvation peut être réduite à un problème purement électrostatique⁶.

Après la mise au point des techniques d'intégration numérique des équations différentielles, les travaux ont montré que cette approche est en effet capable de fournir une description quantitative de la solvation et que les échecs précédents sont à imputer principalement au traitement mathématique insuffisamment élaboré et non pas aux approximations physiques sur lesquelles la méthode se base. En effet, afin d'obtenir une solution analytique, il était nécessaire de considérer une interface sphérique ou ellipsoïdale ou une représentation simplifiée de la distribution des charges. Ceci est insuffisant pour décrire les molécules complexes, dont la surface et la distribution des charges sont hautement irrégulières.

Il est intéressant de montrer que le calcul de l'énergie d'interaction entre un soluté et son environnement diélectrique revient au calcul de l'énergie libre du système soluté+molécules de solvant explicites, en effectuant l'intégration de la fonction de partition⁷ correspondante par rapport aux degrés de liberté du solvant. Le *potentiel de force moyenne (PFM)* ainsi obtenu décrit l'énergie libre d'une distribution de molécules de solvant entourant une géométrie bien définie du soluté, en fonction des coordonnées atomiques de celui-ci. Ce potentiel n'est pas la moyenne de l'énergie des interactions entre le soluté et le solvant, mais inclut aussi l'entropie du solvant, en réponse aux influences ressenties par ses molécules en présence du soluté. Cette énergie libre peut être exprimée comme une fonction -un *potentiel* - des degrés de liberté du soluté, au même titre que le potentiel intramoléculaire de Coulomb ou de van der Waals et peut être utilisé pour calculer le niveau d'énergie de chaque géométrie. En prenant la moyenne de ce PFM de solvation par rapport aux degrés de liberté du soluté, on obtient l'énergie libre totale du système soluté-solvant. Tout ceci met l'accent sur l'avantage majeur des modèles continus de solvation: la possibilité d'exprimer ce potentiel de force moyenne tout en évitant l'étape d'intégration explicite de la fonction de partition par rapport aux degrés de liberté du solvant, ce qui même dans les

⁶ J. Tomasi & M. Persico, *Chem. Rev.* 94, 2027 (1994)

⁷ D.A. McQuarrie, "*Statistical Mechanics*", Harper Collins, NY. (1976).

cas les plus favorables nécessite de très longues simulations dans le but d'obtenir un échantillonnage significatif de tous les états de la couche de solvation. Dans la plupart des situations impliquant des biomolécules, cela est impossible. C'est surtout l'évaluation des termes *entropiques* qui demande une exploration exhaustive de l'espace de phase. Une simulation effectuée avec des molécules explicites de solvant doit en principe aboutir à la même valeur d'énergie libre de solvation calculée à partir d'un modèle continu. Dans le cas le plus simple - le calcul de l'énergie de solvation des ions sphériques - on a démontré que de telles simulations⁸ convergent en effet (mais très difficilement) vers les valeurs données par l'expression de Born. Cette dernière offre une estimation raisonnable et rapide de l'énergie libre de solvation, à la condition que le "rayon ionique", paramètre intervenant dans cette expression, ait été convenablement défini. Par contre, aucune information sur la structure microscopique de la couche de solvation autour d'un ion ne peut être obtenue.

En ce qui concerne les fondements physiques des approches "explicite" et "continue" de la solvation, le traitement explicite du solvant semble être plus rigoureux. Le remplacement de l'intégration de la fonction de partition d'un système soluté-solvant par des résultats obtenus en supposant que le solvant se comporte comme un simple système de dipôles soumis au champ électrique généré par le soluté, est une hypothèse attaquable du point de vue physique. Il faut néanmoins tenir compte du fait que, même si le calcul des fonctions de partition en utilisant un modèle *explicite* des molécules de solvant est réalisable, l'utilisation d'un *champ de force empirique* pour décrire les interactions interatomiques au cours d'une telle simulation nous interdit d'affirmer que la représentation physique du système est meilleure. L'utilisation de modèles explicites de solvant n'offre pas une plus grande rigueur, à moins qu'un Hamiltonien quantique ne soit utilisé. Peut-être à l'exception des petites molécules, un tel effort de calcul est impensable aujourd'hui, et d'une façon ou d'une autre, dans l'état actuel de la science, les solutions pouvant être apportées à des problèmes d'une telle complexité ne peuvent être qu'empiriques.

En effet, le développement de la Mécanique Moléculaire⁹ se basant sur une description *empirique, mais efficace* des interactions interatomiques permet de tirer une parallèle avec le problème des modèles continus de solvant.

⁸ B. Roux, H.A. Yu & M. Karplus, *J. Chem. Phys.*, 94,4683 (1989)

⁹ voir références dans les listes bibliographiques des articles annexés

Dans les deux cas, on se base sur des descriptions fortement simplifiées de la réalité.

Le succès prédictif des *champs de forces empiriques* de la Mécanique Moléculaire est le résultat d'un travail de *calibration* de paramètres intervenant dans la description des interactions interatomiques. En utilisant un ensemble de molécules dont certaines propriétés physico-chimiques ont été mesurées, on recherche un jeu de paramètres pour le champ de forces qui *maximise* l'accord entre les valeurs mesurées et les valeurs calculées. Malgré l'empirisme de la méthode, on peut obtenir une description raisonnable du comportement moléculaire moyennant des calculs modérés.

Les modèles continus de solvant font intervenir beaucoup moins de paramètres que les champs de forces empiriques. Une étape de *calibration* n'est pas obligatoire pour obtenir une description conforme des énergies de solvation d'une grande variété de molécules. Les paramètres d'un modèle continu (les valeurs des constantes diélectriques "effectives", les rayons atomiques définissant l'interface molécule-solvant et les charges atomiques) peuvent être assignés sur une base physique, en utilisant $\epsilon_{\text{int}} \approx 2$ (la valeur correspondant à la polarisabilité électronique "pure"), ϵ_{ext} = la constante diélectrique macroscopique du milieu entourant la molécule, les rayons atomiques mesurés ou importés d'un champ de force moléculaire déjà établi et les charges atomiques calculées par n'importe quel procédé quantique ou semi-empirique. Cependant, il ne faut pas perdre de vue que les fondements de l'approche ne sont pas physiquement exacts - le solvant n'est pas un milieu astructuré, il n'y a pas d'interface nette séparant l'intérieur et l'extérieur de la molécule. Les modèles continus nous *forcent* d'accepter l'existence d'une telle interface afin de remplir les conditions d'applicabilité et il n'est pas évident que la "surface" van der Waals de la molécule (qui d'ailleurs n'est qu'une autre représentation idéalisée se rapportant à une autre catégorie d'interactions atomiques) soit le meilleur choix possible. Typiquement, les auteurs qui s'inspirent des valeurs expérimentales pour paramétriser leur modèle de solvant continu, sont toujours amenés à leur appliquer une correction empirique pour que les résultats calculés à partir de ce modèles soient en accord avec les effets de solvant mesurés¹⁰.

¹⁰ A.A. Rashin, K. Namboodiri, *J. Phys. Chem.*, 91, 6003 (1987)

Pour ces raisons, on peut se demander si la précision des méthodes "continues" peut augmenter en étant appliquées à la "philosophie" des champs de forces empiriques, c'est-à-dire en *optimisant* tous les paramètres qui interviennent dans la description de la solvation. Ceci constitue le thème central des deux publications suivantes "Development and Parametrization of Continuum Solvent Models". De plus, avec une stratégie efficace d'optimisation, on peut envisager une *simplification plus importante* des équations des modèles continus, en espérant que l'optimisation des paramètres puisse compenser une diminution du degré de réalisme physique.

Ceci est sans doute une démarche nécessaire, dans la mesure où, malgré la rapidité des ordinateurs actuels, les solutions numériques élaborées de l'équation de Poisson-Boltzmann (PB)¹¹ appliquées aux macro-molécules sont encore trop pénalisantes pour pouvoir être utilisées couramment dans les simulations nécessitant l'exploration exhaustive des conformations moléculaires¹² (Dynamique Moléculaire MD ou simulations Monte Carlo MC) des macromolécules. Cependant, certaines approches très simplifiées^{13,14}, dont la complexité est tout à fait analogue aux fonctions typiques d'un champ de force empirique, ont été proposées, sans être pourtant étudiées à fond. Ce sont des expressions empiriques, dont leur interprétation physique n'est pas évidente parce que ils ne sont pas rigoureusement déductibles de l'équation de Poisson-Boltzmann ou ils comportent des hypothèses difficilement acceptables d'un point de vue physique. Nous avons systématisé certaines de ces approches empiriques (voir article II - A Unified Approach to the Solvation Problem) et étudié leur capacité d'évaluation des énergies de transfert entre le vide et l'eau pour un lot représentatif des molécules incluant toutes les fonctions usuelles de la chimie organique. L'unification de ces modèles obtenue après ce travail de systématisation nous a permis de proposer d'autres expressions du même style. Nous avons également pu augmenter la qualité de ces modèles en appliquant une optimisation des paramètres impliqués.

¹¹ M.K. Gilson, M.E. Davis, B.A. Luty & J.A. McCammon, *J. Phys. Chem* 97,3591,(1993)

¹² A. Warshel, *Computer Modelling of Chemical Reactions in Enzymes and Solutions* NY, J. Wiley & Sons, Inc. (1991)

¹³ W.C. Still, A. Tempczyk, R.C. Hawley & T.E. Hendirckson, *J. Am. Chem. Soc.*, 112, 6127 (1990)

¹⁴ M.K. Gilson & B. Honig, *J. Comp.-Aided. Mol. Design*, 5,5,(1991)

Parallèlement, nous avons essayé de trouver un modèle de solvation suffisamment simple pour être applicable aux simulations moléculaires, sans pourtant s'éloigner d'une description physiquement fondée (dans le cadre des hypothèses du solvant "continu"). L'article I - "Models Based on the Boundary Element Method" - propose une approche simplifiée pour la résolution de l'équation de PB aboutissant à un modèle efficace, mais plus élaboré, et ne comprenant pas de termes empiriques dont l'interprétation physique soit obscure. Cette approche est beaucoup plus rapide que les approches classiques faisant appel à la technique des éléments de surface (Boundary Element Method - BEM) pour résoudre l'équation de PB. On a ainsi prouvé qu'en regroupant convenablement les termes, on peut éviter l'étape la plus coûteuse en temps de calcul, c'est à dire l'inversion d'une matrice dont la dimension peut atteindre des milliers pour des molécules de la taille d'une petite protéine. Autrement dit, des solutions approximatives et très rapides de l'équation PB ont été trouvées, d'une complexité compatible avec les champs de force empiriques, sans toutefois sacrifier la qualité et le sens physique des solutions exactes. Au niveau de la rapidité, notre approche est comparable aux derniers progrès effectués dans le domaine d'autres techniques numériques d'intégration de l'équation de PB¹⁵. L'optimisation des paramètres intervenants améliore la qualité du modèle ainsi obtenu. Les valeurs calculées des énergies de transfert entre le vide et l'eau sont en bon accord avec les données expérimentales. Notamment, la qualité du modèle est équivalente à celle d'autres approches plus élaborées, faisant appel à une description quantique des molécules¹⁶. Certains phénomènes, ne pouvant être traité explicitement par notre approche, sont pris en compte *implicitement* en permettant aux rayons atomiques de s'ajuster librement afin de maximiser la concordance entre le calcul et l'expérience. Par exemple, les énergies de solvation de certains dérivés halogénés ne peuvent pas être correctement reproduites en se basant sur une distribution de charges fixes, les liaisons C-X étant en effet polarisables en milieu aqueux. Pour cette raison, les charges effectives des atomes C et X sont plus importantes en solution que dans le vide. Lorsque le modèle quantique recalcule les charges en fonction de l'Hamiltonien comprenant le terme de solvation, notre approche considère une distribution fixe des charges. L'étape d'optimisation des paramètres de notre modèle est introduite dans le but de compenser indirectement les défauts

¹⁵ M.J. Holst & J. Saied, *J. Comp. Chem.* 16,347 (1994)

¹⁶ C.J. Cramer & D.G. Truhlar, *J. Comp-Aided Mol. Des.* 6,629 (1992)

due aux inexactitudes inhérentes, comme par exemple les erreurs dans les charges atomiques utilisées. Pour contrebalancer l'erreur due à la charge trop petite de l'halogène, l'optimisation des rayons atomiques a diminué¹⁷ les rayons des halogènes, de façon à ce que la contribution de ces atomes à la solvation soit correctement évaluée. Ceci est un bon exemple qui montre d'un côté les limites des modèles de solvant non-quantiques et de l'autre, la façon dont la "philosophie du champ de force empirique" peut surmonter ces difficultés en introduisant des paramètres effectifs ajustés pour reproduire correctement des effets qui ne sont pas modélisés explicitement.

L'application directe de ce modèle dans les simulations moléculaires est en cours, l'étape la plus délicate étant le calcul du *gradient* du potentiel de solvation, c'est à dire la force exercée par le solvant sur chaque atome du soluté. De prochaines études seront axées essentiellement sur le problème très peu exploré des propriétés dynamiques d'un tel "solvant continu effectif". En effet, l'hypothèse centrale du solvant continu postule que le mouvement des molécules de solvant est très rapide et que cette couche peut être regardé comme étant à tout instant *en équilibre* avec le soluté. Par contre, l'échelle de temps des vibrations atomiques est bien inférieure et le temps des transitions conformationnelles est à peu près du même ordre de grandeur que le temps caractéristique de réorientation des molécules d'eau ("tumbling time" τ_t). Le modèle d'une géométrie fixe du soluté en équilibre avec un solvant "continu" est une description correcte pour le calcul de l'énergie de solvation, mais permet-il aussi de suivre les *fluctuations* de cette énergie par rapport aux mouvements rapides et de faibles amplitudes des atomes du soluté? Ce problème ne devrait pas remettre en question l'utilité du modèle dans les simulations de dynamique moléculaire couvrant des échelles de temps importantes par rapport au τ_t , où l'effet global de solvant sera correctement reproduit même si les fluctuations locales ne le sont pas. Le modèle continu peut sûrement être appliqué aux problèmes d'exploration de l'espace conformationnel d'une molécule, pouvant distinguer les différences d'énergie de solvation entre deux conformations bien distinctes. Un domaine où l'application de ce modèle ne

¹⁷ selon la loi de Born, l'énergie de solvation varie proportionnellement au carré de la charge atomique et inversement proportionnellement au rayon. Ceci s'applique uniquement aux ions monoatomiques, mais reste qualitativement vrai pour les contributions des atomes liés dans des molécules.

semble pas possible est la simulation des spectres de vibration, où l'approximation continue de la couche de solvation n'est pas adéquate.

De plus, le modèle continu ne tient pas compte de la *viscosité* du solvant réel et cela laisse entrevoir la possibilité de quantifier l'effet spécifique de ce facteur sur la dynamique des molécules en comparant des simulations dans un solvant explicite avec la dynamique dans le solvant continu. Si cela s'avère nécessaire, la viscosité pourra être simulée dans le modèle continu par des "chocs stochastiques" appliqués aux atomes du soluté. Ceci suggère qu'en dehors des applications directes en modélisation des systèmes biologiquement importants, l'étude des modèles de solvant continus va contribuer à l'élargissement de nos connaissances théoriques sur les interactions intermoléculaires et la chimie physique de la solvation.

B. L'électrostatique . Rappel des équations fondamentales

Ceci est une présentation succincte et intuitive des équations de l'électrostatique classique et un bref rappel des théories décrivant les diélectriques. Ensuite, le problème de l'énergie du champ électrique en présence de diélectriques sera discuté à l'aide de quelques exemples simples, dans le but d'éclaircir les concepts qui seront appliqués et généralisés dans les publications ci-jointes.

La loi fondamentale de l'électrostatique est la loi de Coulomb, décrivant la force exercée entre deux particules chargées Q_1 et Q_2 dans le vide, comme proportionnelle au produit des charges et inversement proportionnelle au carré de la distance et dont le vecteur définissant leur position relative est \vec{r} :

$$\vec{F} = \frac{Q_1 Q_2}{4\pi\epsilon_0 r^3} \vec{r} \quad (1)$$

Dans les unités du Système International (SI), les unités des mesure correspondantes sont $[F]=N$; $[Q] = C$; $[r] = m$; $[\epsilon_0] = CV^{-1}m^{-1}$. Le système Gaussien considère ϵ_0 égal à l'unité et en conséquence il n'apparaît plus explicitement dans les équations. De plus, pour les applications de modélisation moléculaire, il est plus facile de considérer les longueurs en Å, les charges en unités électroniques (eu) et les énergies en kcal/mol, ce qui donne $1/4\pi\epsilon_0 = 332.0 \text{ kcal.Å.mol}^{-1}.\text{eu}^{-2}$.

La permittivité absolue ϵ_0 est une constante universelle caractérisant l'intensité intrinsèque de l'interaction entre deux quanta de charge (1 unité électronique de charge (eu) = 1.6×10^{-19} C). Les forces électrostatiques sont ainsi plus faibles que les forces nucléaires, mais plus fortes que les forces gravitationnelles.

L'intensité du champ électrique \vec{E} représente la force agissant sur une charge unitaire:

$$\vec{E} = \frac{\vec{F}}{q} \quad (2)$$

l'intensité du champ autour d'une charge ponctuelle isolée devient donc:

$$\vec{E} = \frac{Q}{4\pi\epsilon_0 r^3} \vec{r} \quad (3)$$

La définition du vecteur \vec{D} de déplacement électrique dans le vide s'écrit:

$$\vec{D} = \epsilon_0 \vec{E} = \frac{Q}{4\pi r^3} \vec{r} \quad (4)$$

Une conséquence directe de la dépendance en $1/r^2$ de l'intensité du champ électrique est le théorème de Gauss qui s'énonce sous sa forme intégrale:

$$\epsilon_0 \int_{\Sigma} \vec{E} \cdot d\vec{s} = \int_{V_{\Sigma}} \rho dv = Q_{\Sigma} \quad (5)$$

où $d\vec{s}$ est le vecteur relatif à l'élément de surface ds , dirigé vers l'extérieur de la surface fermée Σ et dans la direction normale, Q_{Σ} est la charge totale contenue dans le volume V_{Σ} délimité par la surface Σ et donc égale à l'intégrale volumique de la densité de charge ρ . L'équation (5) est équivalente à l'expression différentielle du théorème de Gauss:

$$\text{div } \vec{E} = \nabla \cdot \vec{E} = \frac{\rho}{\epsilon_0} \quad (6)$$

Le potentiel électrostatique Φ est une grandeur scalaire caractérisant chaque point de l'espace et qui correspond au travail nécessaire pour le transport d'une charge unitaire de l'infini (où, par définition, ce potentiel est nul) à ce point. La force étant égale et de direction opposée au gradient de l'énergie potentielle, l'intensité du champ électrique devient:

$$\vec{E} = -\nabla \Phi \quad (7)$$

Comme \vec{E} est le gradient d'un potentiel scalaire, cela implique que le rotor de ce champ vectoriel est toujours nul:

$$\nabla \times \vec{E} = -(\nabla \times \nabla) \Phi = 0 \quad (8)$$

En appliquant le théorème de Stokes, il résulte que l'intégrale de ligne du vecteur \vec{E} est également nulle pour n'importe quel circuit fermé. Cette intégrale est constante pour n'importe quel parcours ouvert reliant les deux points A et B. Autrement dit, le champ électrostatique¹⁸ est *conservatif*:

$$\int_{(A \rightarrow B)} \vec{E} \cdot d\vec{r} = - \int_{(A \rightarrow B)} \nabla \Phi \cdot d\vec{r} = \Phi(A) - \Phi(B) \Rightarrow \int \vec{E} \cdot d\vec{r} = 0 \quad (9)$$

¹⁸ en absence de forces électromotrices; la discussion des équations complètes de Maxwell n'est pas le but de cette introduction.

A partir de (6) et (7) on établit l'équation de *Poisson*:

$$\nabla \cdot (-\nabla \Phi) = -\nabla^2 \Phi = \frac{\rho}{\epsilon_0} \quad (10)$$

Dans le cas particulier d'un espace libre des charges ($\rho=0$), cette équation est connue sous le nom d'équation de *Laplace*.

Pour une densité de charge ρ distribuée dans le volume V , le potentiel au point \vec{x} de l'espace s'écrit:

$$\Phi(\vec{x}) = \frac{1}{4\pi\epsilon_0} \int_V \frac{\rho(\vec{r})}{|\vec{x} - \vec{r}|} d^3r \quad (11)$$

L'énergie électrostatique d'une charge Q située en un point et de potentiel Φ est $U=Q\cdot\Phi$. La construction d'un groupement de charges ponctuelles Q_i définie par les positions \vec{r}_i de chaque charge implique un travail correspondant à la somme des énergies nécessaires pour transporter chaque charge de l'infini à sa place, dans le champ électrique créé par celles qui ont déjà été amenées:

$$U = \sum_i Q_i \Phi(Q_j; j < i) = \sum_i Q_i \cdot \sum_{j < i} \frac{Q_j}{4\pi\epsilon_0 |\vec{r}_i - \vec{r}_j|} = \frac{1}{2} \sum_{i \neq j} \frac{Q_i Q_j}{4\pi\epsilon_0 |\vec{r}_i - \vec{r}_j|} \quad (12)$$

Cette expression peut se généraliser dans le cas d'une distribution continue de charge:

$$U = \frac{1}{2} \int_V \rho(\vec{r}) \Phi(\vec{r}) d^3r \quad (13)$$

ou d^3r représente l'élément de volume situé au point \vec{r} . En éliminant ρ de l'équation (13) en accord avec (10) et utilisant la relation (7), on peut reformuler l'expression de cette énergie en termes de vecteurs du champ électrique:

$$U = \frac{1}{2} \int_V \epsilon_0 |\nabla \Phi|^2 d^3r = \frac{1}{2} \int_V \epsilon_0 E^2 d^3r = \frac{1}{2} \int_V \vec{E} \cdot \vec{D} d^3r \quad (14)$$

Cette dernière expression suggère l'interprétation intuitive que l'espace parcouru par les lignes du champ électrique est "chargé" d'énergie électrostatique, la densité volumique d'énergie étant $u=1/2\epsilon_0 E^2$. Cette énergie

est de plus strictement *positive*, comme intégrale d'une grandeur carrée, et indépendante du signe des charges impliquées. Ceci contredit apparemment l'attraction entre charges de signes opposés, en accord avec l'équation (12). Ce problème provient des états de référence différents appliqués aux calculs de l'énergie électrostatique. L'équation (12) considère l'énergie d'une charge isolée comme nulle, alors que les expressions (13) et (14) incluent explicitement l'interaction de la charge *avec son propre champ électrique*, c'est à dire son énergie "intrinsèque". Ceci représente donc la différence entre les énergies obtenues par les deux voies, qui peut être assimilée à une "constante d'intégration" qui bien-sûr ne dépend pas de la géométrie de la distribution de charges. Dans le cas d'une charge ponctuelle, ce terme dont le sens physique est obscur est de plus *infini*, donc seule l'équation (12) reste applicable.

Deux charges (+q,-q) séparées par une distance l considérée comme très petite par rapport à l'échelle des phénomènes étudiées forment un *dipôle électrique* de moment dipolaire μ :

$$\vec{\mu} = ql \vec{1} \quad (15)$$

le sens du vecteur est conventionnellement défini comme sortant du pôle négatif et rentrant dans le pôle positif. Les dipôles électriques ne sont pas des entités physiques fondamentales (comme les dipôles magnétiques), mais un simple système composé de deux charges de signe opposé. Les phénomènes électrostatiques peuvent être complètement décrits sans l'usage du concept du dipôle, en traitant explicitement les deux charges le composant. Cependant, ce concept est utile parce qu'il simplifie le formalisme mathématique décrivant les interactions de tels objets. L'expression du potentiel électrostatique, produit par un dipôle placé à l'origine du système de référence et faisant un angle θ avec le vecteur de position \vec{r} s'écrit (à condition que $r \gg l$):

$$\Phi = \frac{\vec{\mu} \cdot \vec{r}}{4\pi\epsilon_0 r^3} = \frac{\mu \cos \theta}{4\pi\epsilon_0 r^2} \quad (16)$$

L'énergie d'un dipôle situé dans un champ électrique \vec{E} dépend aussi de l'angle défini par les deux vecteurs:

$$U = -\vec{\mu} \cdot \vec{E} = -\mu E \cos \theta \quad (17)$$

En conséquence, l'orientation optimale d'un dipôle est "parallèle " avec le champ électrique local, autrement dit avec la "tête" positive du dipôle dirigé vers le pôle négatif de la source du champ électrique.

C. L'électrostatique des diélectriques .

i. *Définition des diélectriques. Discussion du condensateur plan-parallèle rempli par un diélectrique.* La dispersion diélectrique est l'expression macroscopique de la *polarisation* de la matière par un champ électrique, c'est à dire la génération et l'orientation des dipôles microscopiques sous l'influence d'un champ électrique externe. La nature de la réponse diélectrique a été élucidée pour la première fois par Debye¹⁹. La mesure expérimentale des propriétés diélectriques peut se faire facilement à l'aide d'un condensateur plan-parallèle, dont la capacité dans le vide C_0 est définie comme le rapport entre la quantité de charge Q sur ses plaques et la différence de potentiel appliquée $\Delta\Phi$:

$$C_0 = \frac{Q}{\Delta\Phi} \quad (18)$$

L'intensité \vec{E}_0 du champ entre les plaques peut être facilement calculée à partir du théorème de Gauss, en tenant compte des propriétés de symétrie du système:

$$E_0 = \frac{Q}{\epsilon_0 S} = \frac{\sigma_0}{\epsilon_0} \quad (19)$$

où S représente la surface des plaques et $\sigma_0 = Q/S$ la densité superficielle de charge correspondante. Ce champ peut être considéré comme homogène et orthogonal aux plaques, au moins pour des condensateurs parfaits ayant des plaques de dimensions "infinies" par rapport à leur distance de séparation.

Tout matériel introduit entre les plaques d'un condensateur (voir Fig. 1) va ainsi subir l'effet du champ \vec{E}_0 , qui tend à *séparer* les charges positives et négatives distribuées dans ce matériel (c'est-à-dire le *polariser*).

¹⁹ P.Debye, "Polar Molecules" , Chemical Catalog Co., NY. (1929)

Le mécanisme de la *polarisation électronique* s'explique par les déplacements des centres de charge négative des nuages électroniques vers le pôle positif du champ électrique inducteur, tant que les noyaux atomiques sont attirés dans le sens opposé. Cette séparation des charges signifie l'apparition d'un *dipôle induit*. Cependant, la tendance du champ externe d'arracher les électrons périphériques "gravitant" autour des noyaux est bien sûr compensée par l'attraction électrostatique au sein même des atomes. La polarisabilité électronique des molécules diminue ainsi avec l'intensité croissante de l'attraction entre les nuages électroniques et les noyaux, facteur qui dépend du type d'orbitale électronique considéré et de la charge globale des noyaux:

-la polarisabilité des électrons d'une même orbitale atomique diminue avec le nombre atomique.

-les électrons dans les orbitales d'énergie plus basses sont attirés plus fortement par les noyaux, donc moins polarisables. Dans les molécules, l'ordre de polarisabilité croissante des paires électroniques est: orbitales liantes σ < orbitales liantes π localisées < (orbitales nonliantes, orbitales liantes π délocalisées).

-la géométrie de certaines orbitales permet un mouvement préférentiel des électrons dans des directions bien spécifiques (le plan du système π délocalisé), donc le moment dipolaire induit est plus important si le champ inducteur est parallèle à ces directions. Par contre, pour les molécules *isotropes*, la polarisation électronique ne dépend pas de l'orientation relative de la molécule par rapport au champ extérieur.

La polarisation *d'orientation* se rencontre pour des composés polaires, dont les centres de gravité des charges positives et négatives dans ces molécules ne coïncident pas et génèrent un moment dipolaire permanent. En accord avec l'équation (17), le champ électrique extérieur essaie d'aligner ces dipôles, en s'opposant à leur mouvement thermique. La projection moyenne dans la direction des lignes de champ de ces dipôles sera un compromis entre ces deux tendances opposées et dépendra aussi bien de la température que de la valeur du moment dipolaire moléculaire. De plus l'apparition ou la relaxation de la polarisation d'orientation est limitée par la mobilité des molécules, ce qui est à son tour une conséquence de l'état d'agrégation où, plus généralement, de la nature des forces intermoléculaires gouvernant le comportement de ce matériel. Ainsi, si on

effectue des mesures dans un champ électrique oscillant qui change de direction à une fréquence supérieure à celle de l'orientation des dipôles, la polarisation d'orientation ne va pas être détectée. Ceci permet d'estimer la mobilité des molécules dans un matériau donné et aussi de mesurer la polarisation électronique, pratiquement instantanée, pour des molécules possédant un dipôle permanent.

Indépendamment du mécanisme moléculaire qui est à l'origine, la polarisation se traduit par l'apparition de dipôles "moyens" induits dans le matériau, parallèles au champ extérieur inducteur \vec{E}_0 ²⁰. Cette distribution des dipôles induits est la source d'un champ électrique macroscopique \vec{E}_p généré en réponse aux effets polarisants du champ extérieur, et opposé à celui-ci. Le champ électrique *effectif* \vec{E} dans le matériau, résultant de la superposition de \vec{E}_0 et \vec{E}_p est donc inférieur en intensité par rapport au champ dans le vide. La *constante diélectrique relative* du matériau est définie comme:

$$\epsilon_r = \frac{\vec{E}_0}{\vec{E}} \quad (20)$$

Considérons une tranche d'épaisseur l et de surface S , parallèle aux plaques du condensateur, où l est de l'ordre des dimensions moléculaires. On peut donc considérer ceci comme une simple "couche" de dipôles. Si la densité du matériau est telle qu'il y a n dipôles moléculaires par unité de volume, le nombre de molécules dans cette tranche est nS et la densité superficielle des molécules est nl . Ces dipôles sont exposés à l'influence d'un champ électrique *local* \vec{E} , dont l'ordre de grandeur reste à déterminer. Admettons aussi que la projection moyenne des dipôles (permanents et induits) des molécules dans la direction du champ effectif \vec{E} est $\langle\mu\rangle$, ce qui implique qu'en moyenne, une surface de la tranche considérée (la plus proche du pôle négatif du condensateur) contient un excès de "têtes de dipôle" (charge positive), alors que l'autre présente le même excédent des "queues de dipôle". La charge ainsi séparée est égale à $nS \cdot \langle\mu\rangle$, donc $nS\langle\mu\rangle$, et la densité superficielle correspondante revient à $\sigma = n\langle\mu\rangle$. Cette

²⁰ Attention, le terme "champ inducteur" est utilisé pour mettre en évidence que E_0 est la cause de tous les phénomènes électrostatiques qui se déroulent dans le diélectrique, mais ceci ne veut pas dire que E_0 est le champ qui agit au niveau microscopique et "commande" l'orientation des dipôles. Ce champ électrique *microscopique* "vu" par les molécules sera introduit et discuté dans les paragraphes suivants; on peut montrer qu'il est colinéaire au champ effectif E , et non au champ extérieur E_0 sauf pour le cas particulier du condensateur plan-parallèle.

charge est dite "virtuelle" ou "liée" parce qu'elle est le résultat de la perturbation d'une distribution de charges globalement neutre et non de la présence d'un excès de particules chargées libres. En superposant de telles tranches pour reconstruire un bloc macroscopique de matériau diélectrique d'épaisseur d , les excès de charge portés par les surfaces des tranches macroscopique vont s'annuler mutuellement partout à l'intérieur du matériau, mais resteront inchangées sur les deux interfaces latérales entre le vide et le diélectrique.

Ce deux interfaces parallèles chargées forment un condensateur de polarité opposée, et le champ électrique généré par la densité de charge virtuelle $\sigma = n \langle \mu \rangle$ représente donc le terme \vec{E}_p s'opposant au champ extérieur \vec{E}_0 :

$$\vec{E}_p = - \vec{u} \frac{\sigma}{\epsilon_0} \quad (21)$$

\vec{u} étant le vecteur unitaire de la direction du champ \vec{E}_0 .

La densité volumique des dipôles induits définit le vecteur *depolarisation*:

$$\vec{P} = n \langle \vec{\mu} \rangle = -n \langle \mu \rangle \vec{u} = -\sigma \vec{u} \quad (22)$$

L'unité de la quantité P est celle d'une densité superficielle de charge (C/m^2).

Il est important de souligner que les distributions de charge "virtuelles" à l'interface entre deux milieux de constante diélectrique différente offrent une *description complète des effets de polarisation et du comportement diélectrique*. N'importe quel problème d'électrostatique impliquant des diélectriques peut être réduit à un problème d'électrostatique dans le vide en complétant la distribution des charges "libres" génératrices du champ électrique inducteur par la distribution des charges "virtuelles" décrivant la réponse du diélectrique. Les phénomènes de polarisation se déroulent cependant dans tout le volume du diélectrique, mais le seul résultat global de toutes ces participations microscopiques est l'accumulation de la charge virtuelle à l'interface de séparation. La difficulté majeure est donc de déterminer cette distribution de charges virtuelles, après quoi il est facile d'écrire le champ électrique effectif comme une superposition de champs provenant des charges *libres* et *liées*:

$$\vec{E} = \vec{E}_0 + \vec{E}_p = \vec{E}_0 - \frac{\vec{P}}{\epsilon_0} = \frac{\vec{E}_0}{\epsilon_r} \quad (23)$$

d'où:

$$\vec{P} = (\epsilon_r - 1)\epsilon_0 \vec{E} \quad (24)$$

Pour reprendre l'exemple du condensateur, nous pouvons exprimer l'intensité effective du champ électrique en termes de distributions de charge libre (σ_0) et virtuelle (σ):

$$\vec{E} = \frac{\sigma_0 - \sigma}{\epsilon_0} \cdot \vec{u} = \frac{\sigma_0}{\epsilon_0 \epsilon_r} \cdot \vec{u} \quad (25)$$

L'équation (20) exprime le champ électrique effectif en fonction de l'atténuation diélectrique du champ dans le vide en chaque point de l'espace occupé par de la matière polarisable. L'équation (25), par contre, décrit le même champ en faisant référence uniquement aux distributions de charge des plaques du condensateur et de l'interface diélectrique-vide, en montrant aussi comment la dernière densité découle des propriétés diélectriques du matériel. Ces deux points de vue sont complètement équivalents et on va utiliser l'un ou l'autre pour mieux mettre en évidence certains aspects des problèmes.

Néanmoins, la généralisation de la discussion précédente sur le condensateur plan-parallèle nous oblige à revenir à certaines relations qui ont pu être écrites uniquement sur la base des propriétés particulières de symétrie du système considéré. Ainsi, il est très important de souligner que la relation (20), qui est généralement utilisée comme définition de ce "facteur atténuant" de l'intensité du champ électrique, c'est à dire la constante diélectrique relative, et qui s'énonce habituellement comme $\vec{E} = 1/\epsilon_r \cdot \vec{E}_0$, n'est pas une relation générale. En effet, \vec{E} et \vec{E}_0 ne sont pas généralement colinéaires, sauf dans les milieux diélectriques homogènes (très loin des interfaces diélectriques) où dans des systèmes présentant symétrie cylindrique ou sphérique. Au contraire, le moment dipolaire moyen ($\vec{\mu}$) est toujours colinéaire au champ effectif \vec{E} . On peut montrer que l'équation (24) est toujours valable²¹, et peut servir à relier la grandeur des dipôles qui apparaissent dans un matériel à sa constante diélectrique

²¹ pour les diélectriques *linéaires*, la seule catégorie traitée ici. Il existe des matériaux où la réponse diélectrique n'est pas proportionnelle au champ électrique appliqué.

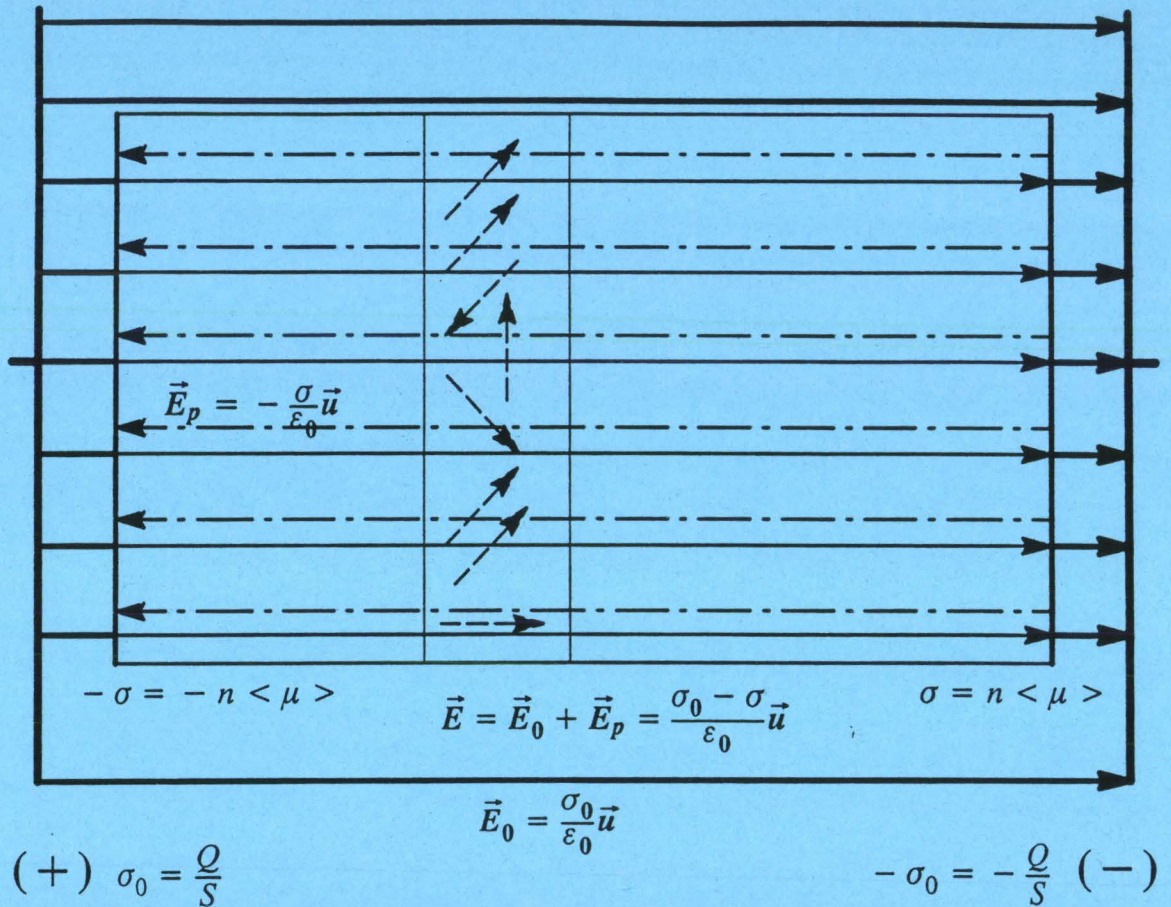


Fig. 1 Les différents champs électriques apparaissant dans un diélectrique sous l'influence d'un champ extérieur.

La polarisation P est le champ macroscopique émergeant des champs individuels des dipôles induits et moléculaires réorientés. Il peut être également vu comme un champ produit par des charges non-compensées à l'interface, celles-ci étant la conséquence du réarrangement global des dipôles. La polarisation P est opposée au champ extérieur E_0 , le champ résultant à l'intérieur du diélectrique est $E = E_0 - P = E_0 / \epsilon$

relative. Cette discussion sera reprise dans le chapitre sur les *théories microscopiques des phénomènes diélectriques*.

ii. *Les équations générales du champ électrostatique décrivant les milieux diélectriques hétérogènes.* Dans le cas général impliquant une distribution quelconque de charges libres et une géométrie quelconque de l'interface séparant des matériaux diélectriques différents, on est amené à considérer pour chaque élément de volume autour d'un point \vec{r} un champ effectif *localement* homogène $\vec{E}(\vec{r})$ et un vecteur de polarisation locale $\vec{P}(\vec{r})$, colinéaire à $\vec{E}(\vec{r})$ si le milieu est *isotrope*.

Evidemment, le vecteur de polarisation représente le champ engendré par les charges virtuelles. Il faut donc ajouter les contributions des dipôles dans le diélectrique, dont la densité volumique est P , à l'expression du potentiel électrostatique du à la distribution des charges libres ρ .

$$\Phi(\vec{x}) = \frac{1}{4\pi\epsilon_0} \int_v \left[\frac{\rho(\vec{r})}{|\vec{x} - \vec{r}|} + \frac{P(\vec{r}) \cdot (\vec{x} - \vec{r})}{|\vec{x} - \vec{r}|^3} \right] d^3r \quad (26)$$

ce qui peut s'écrire plus simplement, après intégration partielle du deuxième terme²²:

$$\Phi(\vec{x}) = \frac{1}{4\pi\epsilon_0} \int_v \frac{\rho(\vec{r}) - \nabla_r P(\vec{r})}{|\vec{x} - \vec{r}|} d^3r \quad (27)$$

Cette expression est tout à fait analogue à l'équation du potentiel dans le vide (11), sauf que la distribution de charges libres est maintenant remplacée par une distribution *effective*:

$$\rho_{eff}(\vec{r}) = \rho(\vec{r}) + \rho_{virt}(\vec{r}) = \rho(\vec{r}) - \nabla_r P(\vec{r}) \quad (28)$$

Cette distribution effective inclut les charges libres et les charges virtuelles. Le symbole ρ_{virt} représentant la densité volumique de charges virtuelles a été utilisé uniquement dans le souci de rester consistant avec le formalisme

²² N.E. Hill, W.E. Vaughan, M. Davies, "Dielectric Properties and Molecular Behaviour", van Nostrand Reinhold Company, London (1969).

décrivant les phénomènes électrostatiques; il s'agit en effet d'une densité superficielle σ_{virt} située à l'interface de séparation entre deux diélectriques.

Le champ électrique \vec{E} s'écrit toujours comme le gradient du potentiel, dans le vide aussi bien que en présence des diélectriques, et on peut montrer que:

$$\epsilon_0 \nabla \vec{E} = \rho_{eff} = \rho + \rho_{virt} \quad (29)$$

$$\nabla \vec{P} = -\rho_{virt} \quad (30)$$

$$\nabla \vec{D} = \rho = \nabla(\epsilon_0 \vec{E} + \vec{P}) \quad (31)$$

Les équations (29)-(31) mettent en évidence les sources de chaque champ vectoriel: \vec{P} dû aux charges virtuelles, \vec{D} aux charges libres et \vec{E} à la distribution effective comprenant les deux distributions. On peut donc définir²³ \vec{D} comme étant égal à:

$$\vec{D} = \epsilon_0 \vec{E} + \vec{P} = \epsilon_r \epsilon_0 \vec{E} \quad (32)$$

Supposer que les origines du champ de \vec{D} reposent uniquement sur les charges libres peut laisser penser que ce champ ne reflète en aucune manière les propriétés diélectriques du milieu et qu'il suit les lignes du champ électrique \vec{E}_0 produit par ces charges libres dans le vide. Ceci est faux: même si le champ $\epsilon_0 \vec{E}_0$ est aussi une solution de l'équation (31), il n'y a pas de relation simple entre \vec{D} et \vec{E}_0 , qui ne sont pas généralement colinéaires. Au contraire, \vec{D} et \vec{E} sont reliés par l'équation (32), et sont colinéaires dans les milieux *isotropes*. Autrement, le scalaire ϵ_r de l'équation (32) doit être remplacé par un *tenseur*.

iii. *Les conditions limites à l'interface entre deux milieux diélectriques.* Les résultats précédents nous permettent d'approfondir les relations entre les vecteurs du champ électrique à l'interface de deux corps de propriétés diélectriques différentes. Cela revient à l'étude d'un cylindre infiniment petit, ayant comme base un élément d'interface et d'une hauteur suffisamment petite pour se permettre de négliger les flux des

²³ Seule, l'équation (31) n'est pas suffisante pour définir le déplacement électrique d'une façon non ambiguë, parce que elle est également satisfaite par n'importe quel autre champ vectoriel $\vec{F} = \vec{D} + \vec{X}$, à l'unique condition que $\text{div } \vec{X} = 0$

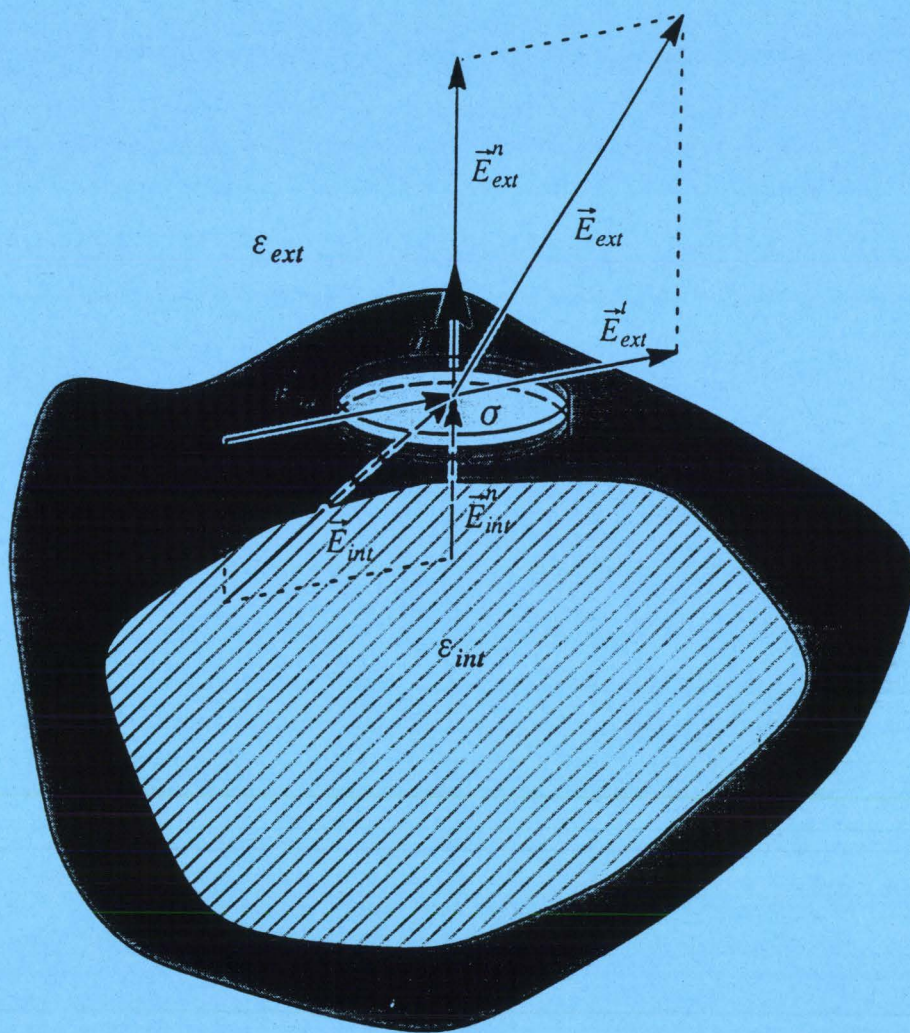


Fig. 2: Les composantes du champ électrique à l'interface entre deux milieux de constantes diélectriques différentes.

vecteurs au niveau de sa surface latérale (Fig. 2). L'application du théorème de Gauss pour ce cylindre revient à prouver qu'il y a une discontinuité de la composante normale de l'intensité électrique en traversant l'interface et que cette discontinuité provient de l'accumulation de la charge virtuelle σ_{virt} à cet endroit (les index *int* et *ext* font référence à l'intérieur et à l'extérieur de l'interface fermée Σ):

$$\epsilon_0 \left(\vec{E}_{ext} - \vec{E}_{int} \right) \cdot \vec{n} = \left(\vec{P}_{ext} - \vec{P}_{int} \right) \cdot \vec{n} = \sigma_{virt} \quad (33)$$

La projection normale du déplacement électrique est continue à l'interface si la densité de charges libres est nulle:

$$\left(\vec{D}_{ext} - \vec{D}_{int} \right) \cdot \vec{n} = \sigma \quad (34)$$

Pour les diélectriques isotropes, ceci implique:

$$\epsilon_{ext} \vec{E}_{ext} \cdot \vec{n} = \epsilon_{int} \vec{E}_{int} \cdot \vec{n} \quad (35)$$

ce qui nous permet d'établir une relation directe entre la densité de charge virtuelle et les constantes diélectriques à l'intérieur et l'extérieur de l'interface:

$$\vec{E}_{ext} \cdot \vec{n} \left(1 - \frac{\epsilon_{ext}}{\epsilon_{int}} \right) = \frac{\sigma_{virt}}{\epsilon_0} \quad (36)$$

A partir de l'équation (36) il est facile de formuler le théorème de Gauss en fonction de la densité de charge virtuelle, en prenant l'intégrale de la projection normale de $\vec{D} = \epsilon_{ext} \epsilon_0 \vec{E}_{ext}$ sur la surface fermée Σ :

$$\int_{\Sigma} \sigma dS = Q_{\Sigma} \left(\frac{1}{\epsilon_{ext}} - \frac{1}{\epsilon_{int}} \right) \quad (37)$$

On peut également montrer que la composante *tangentielle* du vecteur \vec{E} est continue à l'interface, en utilisant la propriété (9) du champ électrique sur un parcours rectangulaire fermé dont les segments tangentiels égaux en longueur sont situés des deux cotés de l'interface, avec des segments normaux traversant l'interface qui sont choisis arbitrairement courts:

$$\left(\vec{E}_{ext} - \vec{E}_{int} \right) \times \vec{n} = 0 \quad (38)$$

iv. *L'énergie électrostatique en présence des diélectriques*: L'équation (13) exprimant l'énergie d'une distribution de charges dans le vide prend en compte le travail nécessaire pour le transport des charges de l'infini vers leur positions finales. La présence des diélectriques pose en plus le problème de l'énergie d'interaction entre le champ électrique et les dipôles du matériau. Le potentiel au sein du système comporte maintenant une contribution de la part des charges virtuelles induites. Néanmoins, la distribution de charges virtuelles ne doit pas être générée explicitement en les amenant de l'infini - elles se forment spontanément au fur et à mesure qu'on met en place la distribution finale de charges *libres* . Autrement dit, on peut à nouveau déduire l'expression de l'énergie électrostatique en suivant le processus de création de la distribution des charges libres. Le potentiel gouvernant les travaux impliqués par le transport de chaque charge fixe inclura cependant les contributions de ces charges virtuelles et va refléter d'une façon implicite l'énergie d'interaction du champ et du matériau diélectrique. Il est évident que dans le vide, l'énergie contenue dans le champ électrique est également une énergie libre (un travail). Par contre, dans les matériaux diélectriques, l'interaction du champ avec les dipôles microscopiques peut être en principe décomposée en un terme enthalpique et un autre entropique. A notre connaissance, ce problème n'est pas clairement discuté dans la littérature, pour la plupart, les auteurs se contentant de parler "simplemment" d'énergie électrostatique.

Considérons qu'à un moment donné pendant la "construction" d'une distribution de charges, on transporte de petites quantités de charge, d'une façon telle qu'on modifie la densité de charge libre $\rho(\vec{r})$ d'un élément de volume par une petite quantité $\delta\rho(\vec{r})$. Pour des raisons de simplicité de raisonnement, on peut considérer ce processus de "chargement" comme la succession des étapes suivantes:

- le transport de la nouvelle quantité de charge de l'infini. Cette charge subirait cependant l'effet exercé par le potentiel Φ des charges libres qui ont été transportées antérieurement et des charges virtuelles issues des effets de polarisation du diélectrique. Ainsi, la variation de la densité d'énergie dans les éléments de volume considérés est $\delta u(\vec{r}) = \delta\rho(\vec{r})\Phi(\vec{r})$ et la variation de l'énergie électrostatique revient à:

$$\delta U = \int_V \delta\rho(\vec{r})\Phi(\vec{r})d^3r \quad (39)$$

- une fois que ces nouvelles charges libres sont amenées à leur positions finales, elles vont perturber les équilibres préexistants des processus de polarisation. Notamment, l'addition des charges exercera un effet *enthalpiquement favorable* sur les dipôles du diélectrique, mais simultanément limitera la liberté de leur mouvement thermique (*entropie défavorable*). La redistribution des charges virtuelles $\delta\rho_{\text{virt}}(\vec{r})$ impliquant une modification du potentiel $\delta\Phi(\vec{r})$ est accompagnée d'un transfert d'énergie du champ électrique en énergie cinétique des molécules: $\delta U_{\text{el}} = -\delta U_{\text{kin}}$. Intuitivement, le champ des nouvelles charges exerce un moment de torsion sur les dipôles du matériel pour les amener à l'orientation optimale, imprimant ainsi un mouvement de rotation aux molécules porteuses de ce dipôles. En conditions *quasi-statiques et isothermes*, cet excès d'énergie cinétique est transformé en chaleur, donc $\delta U_{\text{kin}} = 0$ et $\delta U = \delta U_{\text{el}} = \delta Q$; $\delta S = \delta Q/T$ et la variation globale d'énergie libre $\delta F = \delta U - T\delta S$ est nulle. Pour ces raisons, la variation de l'énergie interne pendant la première étape -l'équation (39)- est égale à la variation d'énergie libre associée à la construction d'une distribution de charges en présence des diélectriques. On va continuer toutefois à utiliser la notation U pour cette énergie libre, comme dans la littérature relative à l'électrostatique.

- le cycle peut être recommencé en introduisant une nouvelle perturbation $\delta\rho'(\vec{r})$, sur laquelle va s'exercer l'influence du potentiel électrostatique $\Phi'(\vec{r}) = \Phi(\vec{r}) + \delta\Phi(\vec{r})$ qui inclut l'effet *global* des charges $\delta\rho(\vec{r})$ introduites antérieurement.

En utilisant $\delta\rho = \nabla \cdot (\delta\vec{D})$ dans (39) et en effectuant une intégration partielle, on obtient:

$$\delta U = \int_v \delta\vec{D} \cdot \vec{E} d^3r \quad (40)$$

En utilisant l'identité $\vec{E} \cdot \delta\vec{D} = 1/2 \delta(\vec{E} \cdot \vec{D})$, on peut exprimer l'énergie libre d'une distribution de charges en présence des matériaux diélectriques d'une façon formellement analogue à l'équation de l'énergie emmagasinée par le champ électrique dans le vide:

$$U = \frac{1}{2} \int_v \vec{E} \cdot \vec{D} d^3r \quad (41)$$

v. Exemple: une sphère chargée dans un environnement diélectrique. Considérons une sphère de rayon R constituée d'un matériel diélectrique de constante ϵ_{int} et située dans un environnement homogène et infini, possédant une constante diélectrique ϵ_{ext} . Une charge positive Q est placée au centre de cette sphère. Ce système simple servira à illustrer le calcul des diverses propriétés électrostatiques en présence des diélectriques.

Prenons tout d'abord le cas $\epsilon_{ext}=1$, la sphère est donc dans le vide. La charge du centre de la sphère induit une densité virtuelle σ à la surface²⁴, ce qui correspond à une charge totale de polarisation $Q_p = 4\pi R^2\sigma$ représentant l'effet net d'orientation des dipôles dans la sphère (il y a plus de dipôles dont la "tête" positivement chargée pointe vers la surface). La charge totale étant conservée, il résulte qu'une quantité $-Q_p$ doit se retrouver dans le centre de la sphère, où s'accumule l'excès des "têtes" dipolaires négativement chargées.

En appliquant (37) en sachant qu'à cause de la symétrie sphérique la densité de charge superficielle est constante, le calcul de σ est immédiat:

$$\sigma = \frac{Q}{4\pi R^2} \left(\frac{1}{\epsilon_{ext}} - \frac{1}{\epsilon_{int}} \right) \quad (42)$$

Implicitement, la charge effective au centre de la sphère devient $Q' = Q - Q_p = Q/\epsilon_{int}$. Il est maintenant facile de calculer l'intensité du champ électrique généré par une charge ponctuelle Q' entourée d'une distribution sphérique de densité σ dans le vide, ce qui revient à notre problème d'origine, une charge Q dans une sphère diélectrique. Evidemment, il n'y a qu'une composante radiale (normale à la surface) de ces vecteurs:

$$E_{ext} = \frac{Q / \epsilon_{ext}}{4\pi\epsilon_0 r^2} \quad (43)$$

$$E_{int} = \frac{Q / \epsilon_{int}}{4\pi\epsilon_0 r^2} \quad (44)$$

$$D = \epsilon_{int}E_{int} = \epsilon_{ext}E_{ext} = \frac{Q}{4\pi\epsilon_0 r^2} \quad (45)$$

Le potentiel dans ce système se calcule facilement en se rappelant qu'une distribution sphérique de charge génère à l'extérieur de la sphère le même

²⁴ on abandonne l'index « virt » pour simplifier la notation; Q est la seule charge libre du système

potentiel qu'une charge ponctuelle centrale et un potentiel constant en tout point situé à l'intérieur de la sphère. On obtient:

$$\Phi = \begin{cases} \frac{Q / \epsilon_{ext}}{4\pi\epsilon_0 r} & (r > R) \\ \frac{Q / \epsilon_{int}}{4\pi\epsilon_0 r} + \frac{Q (1/ \epsilon_{ext} - 1/ \epsilon_{int})}{4\pi\epsilon_0 R} & (r \leq R) \end{cases} \quad (46)$$

Il est aussi très facile d'exprimer l'énergie électrostatique d'une telle distribution de charge dans le vide, en considérant l'interaction de la charge centrale Q' avec le potentiel produit par la couche de charge sphérique dans le centre de la sphère:

$$U^* = Q' \cdot \frac{Q_p}{4\pi\epsilon_0 R} = \frac{Q^2}{4\pi\epsilon_0 R} \cdot \frac{(1/ \epsilon_{ext} - 1/ \epsilon_{int})}{\epsilon_{int}} \quad (47)$$

L'équation (47) décrit-elle l'énergie électrostatique d'une charge située au centre d'une sphère diélectrique? La réponse est *non*: confondre les charges libres et virtuelles dans un calcul d'énergie est une erreur. L'énergie U^* représente le travail nécessaire pour construire notre distribution-modèle uniquement à partir de charges libres, en transportant la charge Q_p de l'infini sur la surface de la sphère autour de Q' . En choisissant donc les grandeurs Q_p et Q' égales aux charges effectives de la sphère diélectrique il est possible de construire une distribution de charges dans *le vide* dont l'intensité du champ électrique et du potentiel sont absolument identiques à ceux produits par la sphère diélectrique chargée. Cependant, dans les deux cas, les *énergies électrostatiques* sont différentes, car la *nature* des distributions de charge impliquées est différente. Evidemment, le vecteur D provenant des charges libres, est aussi différent.

En se basant sur le fait que le travail en présence d'un diélectrique est uniquement dû au transport des charges libres, on obtient la valeur correcte de l'énergie nécessaire pour apporter la charge Q au centre de la sphère, sous influence du potentiel produit par la couche de charges virtuelles émergentes:

$$U = \int_0^Q dq \cdot \Phi(q) = \frac{(1/ \epsilon_{ext} - 1/ \epsilon_{int})}{4\pi\epsilon_0 R} \int_0^Q q \cdot dq = \frac{Q}{8\pi\epsilon_0 R} \left(\frac{1}{\epsilon_{ext}} - \frac{1}{\epsilon_{int}} \right) \quad (49)$$

L'équation (49) représente le premier terme (le "monopole") dans une série d'expressions de plus en plus complexes de l'énergie électrostatique des multipôles (dipôle²⁵, quadripôle, octopôle...) électriques centrés dans une sphère diélectrique. C'est le fondement de l'expression de Born⁵ pour l'énergie de solvation des ions sphériques, le premier modèle continu de solvant, qui s'écrit comme la variation de l'énergie U pendant le transfert de l'ion du vide ($\epsilon_{\text{ext}} = 1$) vers l'eau ($\epsilon_{\text{ext}} = 78$).

Si la sphère n'est pas entourée de vide, mais d'un diélectrique quelconque, le bilan des distributions de charges n'est pas changé en ce qui concerne le centre de la sphère ($Q' = Q/\epsilon_{\text{int}}$), alors que l'équation (42) est toujours valable et exprime la dépendance de la densité de charge interfaciale par rapport aux propriétés du diélectrique externe. Cependant $Q' + Q_p \neq Q$, ce qui pose apparemment un problème de bilan de charges dans le système. Ceci est dû au fait que pour l'instant l'existence d'une deuxième interface diélectrique, celle entre le diélectrique "externe" et le vide, a été ignoré. Cette interface peut se situer pratiquement "à l'infini" mais elle existe néanmoins et porte la charge de polarisation $Q'_p = Q \left(1 - \frac{1}{\epsilon_{\text{ext}}} \right)$, ce qui rétablit le bilan correct des charges. La discussion concernant cette charge de polarisation à l'infini peut paraître purement "académique" mais elle met en évidence une conséquence directe de la loi de Gauss: la quantité totale de charge virtuelle induite par une charge libre à l'interface diélectrique ne dépend pas de la distance entre cette charge et l'interface (ni d'ailleurs de la forme de l'interface). La quantité Q'_p à l'infini n'affecte aucunement notre discussion antérieure, donc les équations (43)-(49) s'écrivent de la même façon pour n'importe quel ϵ_{ext} .

D. Théories microscopiques du comportement diélectrique

Jusqu'à ce point, nous avons évoqué le concept de la polarisation macroscopique, en montrant que les charges virtuelles émergentes aux interfaces diélectriques peuvent décrire complètement les propriétés diélectriques. On a décrit la réponse diélectrique comme la résultante macroscopique des processus d'interactions entre les molécules et le champ

²⁵ L. Onsager, *J. Am. Chem. Soc.* 58,1486 (1936)

électrique, mais nous n'avons pas pour l'instant approfondi le problème de la relation quantitative entre les descripteurs macroscopiques du caractère diélectrique et les propriétés de polarisabilité moléculaire. La difficulté centrale dans ce type d'approche est liée au fait que les molécules subissent une influence complexe aussi bien de la part des champs macroscopiques que de la part des champs intermoléculaires électriques et d'autres natures. Ces derniers effets ne peuvent pas être étudiés facilement, d'une part parce que les ordres de grandeur des gradients microscopiques de potentiel sont immenses (10^{10} V/m) par rapport aux champs macroscopiques et d'autre part parce que l'agitation thermique introduit un élément chaotique. Il est donc nécessaire d'introduire des hypothèses simplificatrices concernant le comportement moléculaire dans le but d'obtenir une approche analytique du problème. Le premier modèle dans ce sens a été proposé par Debye¹⁶.

i. La théorie de Debye de la permittivité statique: En supposant que dans un système composé de molécules de moment dipolaire permanent $\vec{\mu}$, le champ électrique *microscopique* moyen ressenti par les dipôles moléculaires est \vec{E} et que il n'y a pas d'autres forces de nature non-électrostatique qui influencent l'orientation des dipôles, la méthode de Langevin peut être appliquée pour déterminer la projection moyenne des dipôles dans la direction du champ. L'énergie d'interaction dipôle-champ étant donnée par l'équation (17), la probabilité de trouver un dipôle dans l'élément d'angle solide $d\omega$ dont l'inclinaison par rapport au champ \vec{E} est définie par l'angle θ s'écrit:

$$dp(\theta) = \frac{\exp\left(\frac{\mu e \cos \theta}{kT}\right) \cdot d\omega}{\int_{(4\pi)} \exp\left(\frac{\mu e \cos \theta}{kT}\right) \cdot d\omega} \quad (50)$$

Cela représente en fait la probabilité Boltzmannienne de trouver le dipôle dans un état d'énergie $\vec{\mu} \cdot \vec{E}$ dans le contexte d'agitation thermique à la température T. En développant l'expression de l'élément d'angle solide $d\omega = 2\pi \sin \theta d\theta$, la projection moyenne du dipôle $\vec{\mu}$ dans la direction \vec{E} devient:

$$\langle \mu_e^{or} \rangle = \mu \langle \cos \theta \rangle = \int_0^\pi \mu \cos \theta \cdot dp(\theta) \quad (51)$$

L'intégrale (51) peut être résolue analytiquement, ce qui permet d'exprimer la contribution d'orientation des dipôles permanents comme:

$$\langle \mu_e^{or} \rangle = \mu \cdot L(\mu e/kT) = \mu [\coth(\mu e/kT) - (\mu e/kT)] \approx \frac{\mu^2 e}{3kT} \quad (52)$$

Si l'interaction entre les dipôles et le champ électrique est beaucoup plus faible que l'énergie cinétique correspondante à la température T ($x = \mu e/kT \ll 1$), la "fonction de Langevin" $L(x)$ peut être développée en série Taylor et approximée- voir la relation (52).

La contribution de la polarisation électronique peut être considérée comme proportionnelle à l'intensité du champ électrique microscopique \vec{e} et est toujours colinéaire à ce champ (les nuages électroniques peuvent se "réarranger" très rapidement, à la différence des dipôles moléculaires dont la réorientation implique une rotation de la molécule qui possède une inertie non négligeable et éventuellement des barrières énergétiques à franchir). Cette constante de proportionnalité, la polarisabilité électronique moléculaire α , dépendant uniquement des propriétés électroniques des molécules (un tenseur de polarisabilité est nécessaire pour décrire des molécules anisotropes), mais non de la température. Ceci nous permet d'écrire tout de suite la norme du dipôle total induit dans la direction du champ \vec{e} :

$$\langle \mu_e \rangle = \langle \mu_e^{or} \rangle + \langle \mu_e^{el} \rangle = \left(\alpha + \frac{\mu^2}{3kT} \right) e \quad (53)$$

Pour déterminer ce champ microscopique \vec{e} , on peut de nouveau analyser le diélectrique entre les plaques d'un condensateur et diviser ce champ en trois contributions différentes:

a.) la contribution macroscopique \vec{E} qui est présente en tout point du matériel.

b.) les contributions provenant des interactions avec les molécules du proche voisinage. On peut donc envisager une sphère de rayon R autour d'un dipôle pour délimiter les molécules voisines du reste du matériau. A l'intérieur de cette sphère, les interactions intermoléculaires doivent être

traitées explicitement, comme un problème classique d'électrostatique dans le vide. Ces effets ne peuvent pas être quantifiés d'une manière générale. Leurs moyennes par rapport à l'agitation thermique sont cependant nulles dans les cas simples des gaz idéaux ou des réseaux cristallins cubiques.

c.) les contributions des autres molécules du système. En dehors de cette sphère, on peut de nouveau faire appel aux propriétés diélectriques du matériel et donc définir une *interface diélectrique* (voir Fig. 3) séparant l'intérieur de la sphère avec $\epsilon_{\text{int}}=1$ du reste du matériau polarisable. L'interaction entre le dipôle étudié et les charges virtuelles portées par l'interface représente donc l'effet global des dipôles lointains sur celui ci.

Cette "interface" est bien sûr artificielle, les propriétés du matériau sont les mêmes à l'intérieur et à l'extérieur de la sphère. Les différences résident uniquement dans les formalismes utilisés, qui sont qualitativement différents mais néanmoins complètement analogues. En principe, le comportement diélectrique peut être retrouvé en simulant le comportement d'un ensemble suffisamment large de molécules individuelles dans le vide et pendant un temps suffisant. De plus, ce type de modèle "mixte" faisant appel à la description explicite des molécules voisines mais appliquant un modèle continu du solvant lointain a été développé récemment²⁶.

Pour évaluer l'interaction entre le dipôle central et la densité de charge de l'interface, il faut d'abord noter que celle ci dépend de l'inclinaison de chaque élément de surface par rapport au vecteur \vec{P} . La polarisation \vec{P} correspond à une densité σ sur une surface orthogonale à la direction de \vec{P} - voir équation (22), ce qui implique une densité $\sigma \cdot \cos\theta$ sur la surface de la sphère (Fig. 3). Par conséquent, le champ au centre de la sphère dû à cette charge virtuelle devient:

$$\vec{E}_{\text{pol}} = \frac{\sigma}{\epsilon_0} \vec{u} \cdot \int_0^\pi \cos^2\theta \sin\theta d\theta = \frac{\sigma}{3\epsilon_0} \vec{u} = \frac{\vec{P}}{3\epsilon_0} \quad (54)$$

²⁶ A.H. Juffer, E.F.F. Botta, B.A.M van Keulen, A. van de Ploeg & H.C. Berendsen, *J. Comp. Phys*, 97,8 (1991)

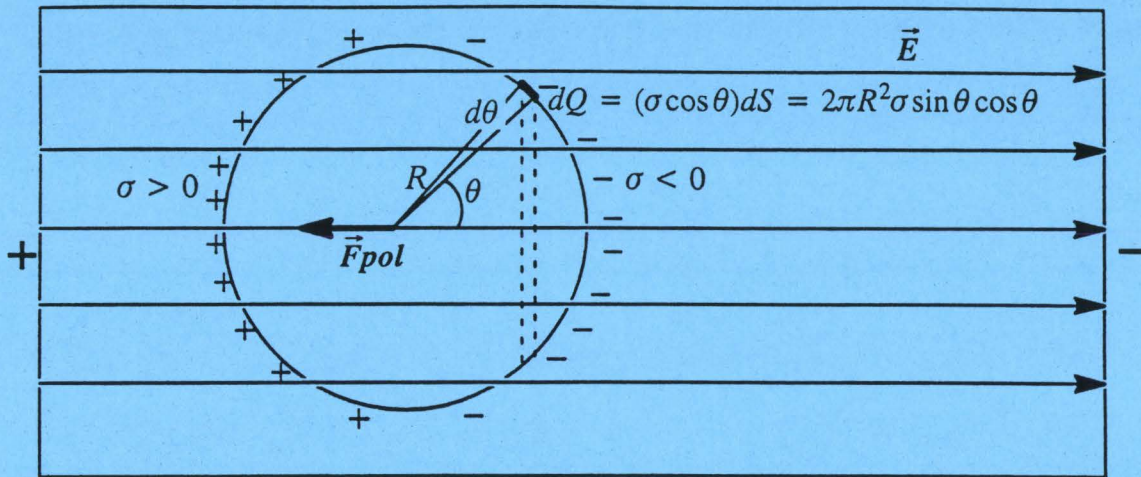


Fig. 3. Calcul du champ microscopique effectif au niveau d'un dipôle moléculaire. A l'intérieur de la sphère, la molécule est considérée explicitement, tandis qu'à l'extérieur, le milieu est supposé homogène et caractérisé par une constante diélectrique relative ϵ . A l'interface avec la sphère, la densité de charge de polarisation virtuelle est égale à $\sigma \cos \theta$, à cause de l'inclinaison de cette surface par rapport aux lignes de champ. Le champ E_{pol} au centre de la sphère est le résultat de toutes les contributions dues aux charges dQ , localisées sur les éléments de surface dS .

et, en considérant les situations où l'influence des molécules voisines est nulle:

$$\vec{e} = \vec{E} + \vec{E}_{pol} = \vec{E} + \frac{\vec{P}}{3\epsilon_0} \quad (55)$$

En appliquant la relation (24), ceci revient à:

$$\vec{e} = \frac{\epsilon_r + 2}{3} \vec{E} \quad (56)$$

Enfin, rappelons que le vecteur \vec{P} représente la densité volumique des dipôles induits, conformément à l'équation (22), ce qui nous permet de démontrer en utilisant également les relations (53) et (56) que:

$$\vec{P} = n \left(\alpha + \frac{\mu^2}{3kT} \right) \cdot \frac{\epsilon_r + 2}{3} \cdot \vec{E} = (\epsilon_r - 1) \epsilon_0 \vec{E} \quad (57)$$

A partir de (57), l'équation de Debye pour la permittivité statique des diélectriques peut s'exprimer en fonction de la masse moléculaire M et de la densité d du matériel, en sachant que $n = N_A d / M$, où N_A est le nombre d'Avogadro:

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} \cdot \frac{M}{d} = N_A \cdot \frac{\alpha + \mu^2/3kT}{3\epsilon_0} \quad (58)$$

Cette équation s'applique donc pour les systèmes idéaux où les interactions intermoléculaires sont en moyenne nulles, ce qui permet une réponse diélectrique *linéaire*. Les limites d'applicabilité de cette approche peuvent être illustrées en considérant la température:

$$T_{CW} = \frac{N_A \cdot d \cdot \mu^2}{9Mk\epsilon_0} \quad (59)$$

appelée le "point Curie-Weiss" en dessous duquel la constante diélectrique relative devient infinie et les dipôles s'orientent spontanément dans leur propre champ électrique, atteignant un état *ferroélectrique*. L'absence de modèles corrects pour décrire les forces intermoléculaires dans la théorie de Debye cause une surestimation importante de cette température (1100 K pour l'eau!). Des théories plus élaborées essayant de prendre en compte ces phénomènes peuvent en effet augmenter les prédictions des points Curie-Weiss¹⁹.

Development and parametrization of continuum solvent models. I. Models based on the boundary element method

D. Horvath

Institut Pasteur de Lille, Service de Chimie des Biomolécules URA CNRS 1309, 1 Rue Prof. Calmette, 59000 Lille, France, Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, P2, Avenue F. D. Roosevelt 50, B-1050 Bruxelles, Belgium, and "Babes-Bolyai" University of Cluj-Napoca, Dept. of Organic Chemistry, Str. Arany Janos Nr.11, 3400 Cluj-Napoca, Romania

D. van Belle

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, P2, Avenue F. D. Roosevelt 50, B-1050 Bruxelles, Belgium

G. Lippens

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, P2, Avenue F. D. Roosevelt 50, B-1050 Bruxelles, Belgium, and Institut Pasteur de Lille, Service de Chimie des Biomolécules URA CNRS 1309, 1 Rue Prof. Calmette, 59000 Lille, France

S. J. Wodak

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, P2, Avenue F. D. Roosevelt 50, B-1050 Bruxelles, Belgium

(Received 27 February 1995; accepted 22 January 1996)

A series of different simplifications of the boundary element method (BEM) for solving the Poisson–Boltzmann equation is investigated in an effort to obtain an accurate and fast enough treatment of electrostatic effects to be incorporated in Monte-Carlo and molecular dynamics simulation methods. The tested simplifications include increasing the size of Boundary Elements, decreasing the surface dot density, and ignoring the interactions between the polarization charges. Combined with terms describing the nonelectrostatic solvation effects, the simplified BEM polarization terms were built into expressions for the solvation potential. The solvation potential is treated as empirical consistent force field equations. The intervening parameters, including atomic and probe radii, are derived by different fitting strategies of calculated vs experimental vacuum to water transfer energies of 173 charged, polar, and nonpolar small molecules. These fits are shown to yield very good correlations (rms ~ 1.4 kcal/mol), even when the interactions between the polarization charges are neglected, proving that the most time-consuming step in BEM, which involves solving the linear system, can be successfully avoided. Finally, the computing efficiency of the method is tested on macromolecules and is found to be convenient for implementation in molecular dynamics or Monte Carlo simulations. © 1996 American Institute of Physics. [S0021-9606(96)50816-2]

I. INTRODUCTION

Protein folding,^{1–3} ligand binding,^{4–6} and enzymatic reactions⁷ are believed to result from a combination of electrostatic and hydrophobic effects. One of the major challenges of computer simulations has been to adequately model these processes. Given that the surrounding water solvent plays a major role in the two types of interactions, it has become clear that theoretical predictions cannot be brought in agreement with the experimental data unless the solvent effects on the behavior of the biomolecules are correctly represented.^{8–13} An adequate treatment of solvation effects is particularly crucial for the evaluation of *pK* values of ionizable groups, which often play a key role in these processes.^{14–17}

Two major approaches have been used to model solvent effects in systems of biological and organic molecules. One is the *microscopic* approach,⁹ which relies on the assumption that the system described by the force field can be represented by an explicit description of both the solute and the

solvent molecules (including counterions). The solvation effects are then given by the sum of solute–solvent and solute–solute terms of the potential. However, though in most current force fields parameters have been refined in order to improve agreement with experimental data,^{18,19} significant shortcomings persist, in particular, with regard to the dynamic properties of liquid water and protein–water solutions (see Ref. 20 and references therein). While it is clear that the microscopic models are well suited for evaluating quantities linked to local and transport properties of the solvation shell,²¹ entropy-dependent solvation properties remain difficult to compute since they require very long simulations.^{22–24}

An alternative approach is based on the derivation of expressions for the *potential of mean force* exerted by the average solvation shell on the solute. It considers both the solvent and solute as homogeneous structureless media and is therefore termed the continuum approach.¹³ This potential of mean force contains two main terms: an electrostatic term, which is a function of the solute *charge distribution*, the

internal and the solvent *dielectric constants* and the *interface between the two dielectric media*, and a term which represents other contributions such as the energy of creating a cavity in the solvent and the solute-solvent van der Waals interactions, collectively referred to as the hydrophobic effect.^{25,26}

The electrostatic term is usually expressed in the framework of classical electrostatics of dielectric media. In particular, much work has been devoted to describing the electrostatic part of the solvation process in terms of the Poisson-Boltzmann equation (PBE).²⁷⁻³³ Most of the approaches have been concerned with finding efficient ways for solving the PBE, and use rather complex mathematical tools. The finite difference method (FDM²⁷⁻³⁰) uses volume integrals based on a 3D grid description of the dielectric properties of the system. The boundary element method (BEM³¹⁻³³) employs surface integrals on the dielectric boundary. However, irrespective of the method used to solve the PBE, the results of the calculations critically depend on choice of the dielectric boundary position, on the description of the solute polarization and on the set of atomic partial charges. Furthermore, it has been found that the results also depend on the parametrization of the nonelectrostatic term that must be added to reproduce experimental solvation energies.³⁴ Several works were concerned with the optimal choice of the parameters of continuum solvent models, none of which could avoid a fitting step in order to improve the results. In older works, the number of fitted parameters was still small—e.g., rescaling the radii or adjusting the value of the internal dielectric constant or rescaling the polarization energies. In one,³¹ the choice of the atomic radii defining the dielectric interface in terms of electron-density profiles around ions was discussed. In another,³⁵ several descriptions of the solute polarizability terms were tested. More recently, extensive fitting procedures include adjustment of atomic radii, charges or bond dipoles.^{36,37} To our knowledge, the only extensive parametrization of a BE-based solvation model was done by Cramer and Truhlar.³⁶ Their semiempirical quantum approach uses a very large number of parameters and cannot be applied directly to macromolecules. A parametrization of a finite difference solvation term which completely reassigns atomic charges from fitted bond dipoles appeared recently in the literature.³⁷

Several attempts have been made to incorporate the continuum solvent formalism into the molecular mechanics Hamiltonian.^{29,33,34,38,39} They mainly involved introducing hypotheses allowing a simplified and fast evaluation of the volume integrals of the FDM integration of PBE. Again, the parametrization issues were not fully addressed and the predictive power of these models has not been tested.

This notwithstanding, some of the proposed continuum solvent models were found to be of the same accuracy as the much more time-consuming free energy perturbation simulations with an explicit solvent description.^{34,40} For ionic compounds however, both methods are often in error by 5–10 kcal/mol with respect to experimental values.

This paper presents a simplified boundary element approach to the calculation of the polarization energies of mol-

ecules, which when combined with an appropriate term for the nonelectrostatic solvation effects constitutes a relatively simple solvent model that can be readily implemented in procedures that involve iterative energy calculations. To derive our model, different levels of simplification of the BE method³¹ are investigated, to determine the best compromise between rigor and computational speed. The approach consists in deriving parameters such as the atomic and probe radii defining the dielectric boundary and the values of the “dielectric constants” by fitting calculated energies to experimental vacuum-to-water transfer energies of several learning sets of (64 and 173) small molecules.^{13,34,36,41} The predictive power of the optimized force field in reproducing the experimental values is then tested by a cross validation procedure. The influence of the atomic charges on the parameters of the model is also investigated by testing three different atomic partial charge sets, respectively Biosym,^{42,43} MOPAC-AM1,⁴⁴ and MOPAC-AM1 (ESP option).⁴⁵ The method has been tested on macromolecules and its precision and computational performance are monitored in function of different setup parameters.

II. METHODS

The potential of mean force representing solvation effects may be considered as containing two terms: an electrostatic (*polarization*) term and a *surface area* term:⁴⁶

$$F_{\text{ext}} = F_{\text{ext}}^{\text{elec}}(\epsilon_{\text{int}}, \epsilon_{\text{ext}}, \Sigma_{\text{ext}}) + F_{\text{ext}}^{\text{surf}}(\kappa_{\text{ext}}, \Sigma_{\text{ext}}), \quad (1)$$

with ϵ_{int} , ϵ_{ext} representing respectively the internal and external dielectric constants, κ_{ext} is the surface tension associated to the exterior medium and Σ_{ext} denotes the separating interface, whose geometry depends on the molecular conformation, and the exterior medium ext.

To a first approximation, the vacuum-to-solvent transfer free energy can be expressed as the variation of the potential of mean force when “switching” the external medium ext from vacuum to solvent, with $\epsilon_{\text{ext}}=1$ becoming $\epsilon_{\text{solvent}}$, and $\kappa_{\text{ext}}=0$ becoming κ_{solvent} (as illustrated in Fig. 1). It represents in fact the difference of the average *total* energies of two *ensembles* of molecular conformations, that in vacuum and that in solvent. This energy may be calculated by adding appropriately parametrized potentials of mean force to the conventional molecular force field in *both* vacuum and solvent, and computing the difference between the Boltzmann averages of the corrected energies in the respective environments. This, to our knowledge, is a more rigorous approach than that traditionally taken,^{32,34} where the description of the vacuum state does not include polarization effects and a transfer energy term is added to the vacuum terms to describe the solvated state. Furthermore, the fact that potentials of mean force are used to describe the polarization phenomena of the molecule in each of the external environments (vacuum, solvent) make it possible to consider more subtle aspects, as for example different definitions of the molecular interface in vacuum and solvent.

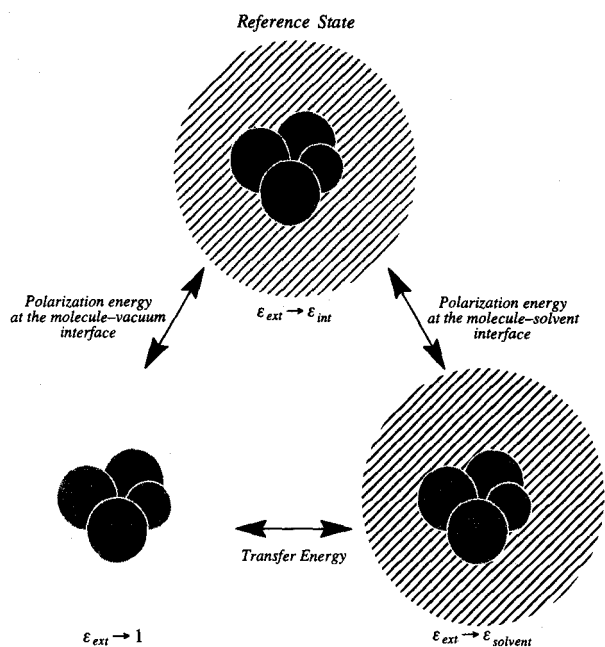


FIG. 1. The thermodynamic cycle describing the contribution of the polarization free energy to the gas phase-to-solvent transfer process. In the reference state, the energy is equal to the Coulomb term computed with the internal dielectric constant ϵ_{int} . When the molecule is transferred from the reference state $\epsilon_{\text{ext}} = \epsilon_{\text{int}}$, to either the vacuum $\epsilon_{\text{ext}} = 1$, or the solvent, $\epsilon_{\text{ext}} = \epsilon_{\text{solv}}$, polarization free energies arise from the creation of the dielectric boundary.

A. Modeling the electrostatic effects with the boundary element method

The boundary element method (BEM) is a numerical procedure which allows to compute the polarization charge densities σ that arise at arbitrary shaped solute-solvent boundaries due to the differences in the dielectric properties of the two media. These polarization charge densities are linked to the discontinuity of the normal component of the electric field vector at the boundary:

$$(\mathbf{E}_{\text{ext}} - \mathbf{E}_{\text{int}}) \cdot \mathbf{n} = \sigma / \epsilon_0 \quad (2)$$

where E_{ext} and E_{int} denote the field vectors at the inner and outer side of the dielectric interface respectively, whose normal direction is n . The electric displacement vector D has a continuous normal component at the boundary:⁴⁷

$$\mathbf{D} \cdot \mathbf{n} = \epsilon_{\text{int}} (\mathbf{E}_{\text{int}} \cdot \mathbf{n}) = \epsilon_{\text{ext}} (\mathbf{E}_{\text{ext}} \cdot \mathbf{n}). \quad (3)$$

The first step of the BE model is the definition of the boundary surface Σ . In principle, both the *molecular surface* and the *solvent-accessible surface*⁴⁸ are possible choices (see Tomasi and Persico⁴⁹ for an overview of the advantages and disadvantages of each type of these surfaces). In the present study, we have opted for the solvent-accessible surface, which is straightforward to obtain and showing no reentrant parts. After having chosen the type of surface, the values for the atomic and probe radii are required to complete the definition of the boundary. Of course, the *dot density* of the

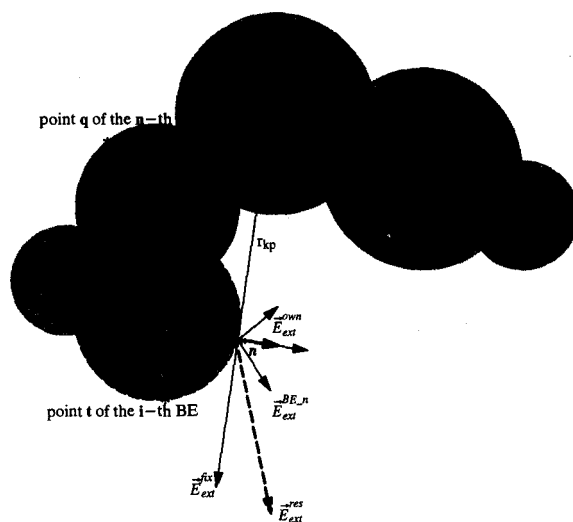


FIG. 2. The various contributions to the electric field at the dielectric boundary. In our approach, the characteristic value of the normal component of the electric field at the i th BE is taken as the average of the local normal components of the electric field vectors $E_{\text{ext}}^{\text{res}}(p:i)$ over all the points p . The notation $p:i$ stands for "point p of the i th BE." $\mathbf{E}_{\text{ext}}^{\text{fix}}(k \rightarrow p:i) = Q_k \mathbf{r}_{kp} / 4\pi\epsilon_{\text{int}} r_{kp}^3$; $\mathbf{E}_{\text{ext}}^{\text{own}}(t:i, p:i \rightarrow p:i) = (\sigma_t / 2\epsilon_0) \mathbf{n} + (\sigma_t dS_t \mathbf{r}_{tp} / 4\pi\epsilon_0 r_{tp}^3)$; $\mathbf{E}_{\text{ext}}^{\text{BE},n}(q:n \rightarrow p:i) = \sigma_n dS_q \mathbf{r}_{qp} / 4\pi\epsilon_0 r_{qp}^3$; $\mathbf{E}_{\text{ext}}^{\text{res}}(p:i) = \sum_k \mathbf{E}_{\text{ext}}^{\text{fix}} + \sum_n \sum_q \mathbf{E}_{\text{ext}}^{\text{BE},n} + \sum_t \mathbf{E}_{\text{ext}}^{\text{own}}$.

boundary, the analog of the "grid spacing" in finite difference approaches, is an important setup parameter of the method.

Once the boundary is defined, the polarization charge densities can be computed numerically, by dividing the boundary surface into small areas [*boundary elements* (BE)], for which σ is assumed to be constant. One strategy has been to define many BEs (hundreds for a molecule of tens of atoms) and take their center as the single representative point for which the electric field is evaluated,³² which is much too time consuming. With the goal of obtaining a solvent model practical enough to be used in molecular dynamics simulations of macromolecules, we take a different approach. We make the hypothesis that *each portion of the boundary surface of an atom can be considered as a single BE*, and take advantage of the fact that it can be represented by a number of equally spaced dots. Naturally, this division into BEs could not have been done on the molecular surface, that cannot be split unambiguously into atomic domains and that requires much more complex tessellation procedures. We then evaluate the normal component of the electric field at all the surface dots contained in the i th BE and calculate its average value $\langle \mathbf{E}_{\text{ext}} \cdot \mathbf{n} \rangle_i$ which is related to the average σ on that BE (Fig. 2).

Consequently, we use the following equations to build up the linear system of the charge densities.

The contribution from the fixed charges at the BE _{i} is

$$\langle \mathbf{E}_{\text{ext}}^{\text{fix}} \cdot \mathbf{n} \rangle_i = \frac{1}{N_{\text{points}}^i} \sum_{p=1}^{N_{\text{points}}^i} \sum_{k=1}^{N_{\text{atoms}}} \frac{Q_k (\mathbf{r}_{kp} \cdot \mathbf{n}_p)}{4\pi\epsilon_{\text{int}} r_{kp}^3}, \quad (4)$$

where N_{points} is the number of considered points on the surface area of the BE $_i$, n_p is the normal vector at point \mathbf{p} of this BE, while r_{kp} is the distance of this point to the center of the atom \mathbf{k} , whose charge is Q_k .

When the distance r_{kp} is large, computer time might be saved by performing the sum (4) over a smaller subset of points $n_{\text{points}} < N_{\text{points}}$ of the boundary elements, in incrementing the counter of points p by steps larger than one. This step size should be an increasing function of the distance d_{ik} between the current atom k and the atom i at the center of the current BE. We have defined this step size as

$$\text{step} = 1 + \frac{d_{ik}^2}{D_{\text{cut}}}, \quad (5)$$

where D_{cut} , the "halving square distance" is the square distance at which integration is performed using only half of the points of a BE. This is a setup parameter which is expected to become important for macromolecules.

The normal projection of the field at each point of each BE produced by the charge density σ at the same point is simply $\sigma/2\epsilon_0$ and is constant within each BE. On the other hand, the points located on the BE feel a different influence from their neighbors as a function of their position. The average field produced by a charge distribution at the points of its own BE $_i$ is given by

$$\langle \mathbf{E}_{\text{ext}}^{\text{own}} \cdot \mathbf{n} \rangle_i = \frac{\sigma_i}{2\epsilon_0} + \frac{1}{N_{\text{points}}^i} \sum_{p=1}^{N_{\text{points}}^i} \sum_{t=1}^{N_{\text{points}}^i} \frac{\sigma_t dS_t (\mathbf{r}_{tp} \cdot \mathbf{n}_p)}{4\pi\epsilon_0 r_{tp}^3} \quad (6)$$

$\sigma_t dS_t$ being the amount of charge located at the point t of the BE $_i$ inducing a field at point \mathbf{p} at a distance of r_{tp} .

The average field produced by the n th BE at the interface of the element i will be

$$\langle \mathbf{E}_{\text{ext}}^{\text{BE}_n} \cdot \mathbf{n} \rangle_i = \frac{1}{N_{\text{points}}^i} \sum_{p=1}^{N_{\text{points}}^i} \sum_{q=1}^{N_{\text{points}}^n} \frac{\sigma_n dS_q (\mathbf{r}_{qp} \cdot \mathbf{n}_p)}{4\pi\epsilon_0 r_{qp}^3}. \quad (7)$$

Due to the discontinuity in the curvature of the exposed surface along the intersection line of two atomic spheres, pairs of points with r_{qp} lower than a given threshold δr (typically in the range of 0.1 Å) must be excluded from the summation in Eq. (7). This "exclusion distance" is therefore the third empirical setup parameter introduced in our approach.

Equations (2)–(7) form a linear system in σ , leading to an initial set of charge densities. The mutual interactions between the polarization charges of different BEs (7) are the source of the nondiagonal terms in the matrix of this linear system. Its inversion, reported to be the most time-consuming step,³² can be avoided by assuming that the interaction between the spatially distributed polarization charges are much weaker than those produced by the atomic point charges and neglecting them. Comparative runs of BEM calculations including or neglecting the terms (7) have been carried out here to check this hypothesis.

The polarization charge densities obtained as solutions of the linear system are affected by inherent errors arising

from the use of finite size BEs. These errors can be compensated by introducing a rescaling step in order to fulfill Gauss' theorem.³¹ The impact of this rescaling on the quality of the results is also investigated in this study.

The *polarization free energy* arises from the interaction of the *fixed charges* with the *polarization charge densities*:

$$F^{\text{pol}} = \sum_{i=1}^{N_{\text{atoms}}} F_i^{\text{pol}} = \frac{1}{2} \sum_{i=1}^{N_{\text{atoms}}} \sum_{k=1}^{N_{\text{BE}}} Q_i \sigma_k \sum_{q=1}^{N_{\text{points}}^k} \frac{dS_q}{4\pi\epsilon_0 r_{iq}}. \quad (8)$$

Since the energy can be derived from the free energy in stating that $U = F - T(\partial F/\partial T)$, we define the "entropy-corrected",³² σ^* that directly leads to the value of U when used in Eq. (8):

$$\sigma^* = \sigma - T \frac{\partial \sigma}{\partial T}. \quad (9)$$

Similarly, we may rescale these quantities according to Gauss' law:

$$\frac{\partial}{\partial T} \int_{\Sigma} \sigma dS = \int_{\Sigma} \frac{\partial \sigma}{\partial T} dS = - \frac{Q_{\Sigma}}{\epsilon_{\text{sol}}^2} \frac{\partial \epsilon_{\text{sol}}}{\partial T}. \quad (10)$$

In the present work, we will refer to experimental *free energies* of transfer. The electrostatic free energy of a molecule embedded in a dielectric medium ext is defined as:

$$F_{\text{elec}}^{\text{ext}} = \frac{1}{\epsilon_{\text{int}}} U_{\text{Coul}}^0 + F_{\text{pol}}^{\text{ext}}. \quad (11)$$

B. Modeling the nonelectrostatic part of the solvation process

The insertion of the solute into the solvent, which is accompanied by a cavity formation in the solvent and the resulting balance of van der Waals interactions, is generally described by a term proportional to the accessible surface of the solute, where the proportionality constants are fitted into the model.^{41,46,50} A recent study on the hydration of hydrophobic solutes⁵¹ suggests that solvent reorientation *in the vicinity of nonpolar groups* is a self-compensating process, which does not contribute significantly to the vacuum-to-water transfer free energy and that hydrophobicity can be mainly attributed to the work of cavity formation. It is unclear however whether this conclusion holds for polar or charged solutes. In this work, the non electrostatic solvation term is expressed as a *linear* function of the *total* area S_{total} (accounting for cavity, dispersion and an average solvent reorganization) and the *hydrophobic* area S_{nonpolar} , accounting for the possibly *different* solvent reorganization contributions around polar and nonpolar groups.

$$F_{\text{surf}}^{\text{solv}} = \kappa_{\text{total}} S_{\text{total}} + \kappa_{\text{nonpolar}} S_{\text{nonpolar}}. \quad (12)$$

Here, S_{total} is the area of the *dielectric interface* and S_{nonpolar} is the sum of the solvent-exposed carbons and hydrogens of aliphatic groups, where the definition of the latter has been made in terms of Biosym-cvff potential types: aliphatic carbons (ca,c1,c2,c3,c4) and hydrogens (h). Aromatic and multiple-bonded carbon atoms have not been included in

the definition, since their interaction with water might be different due to the presence of the π electron clouds. The results will show to what extent these empirical definitions are relevant.

We also report a different approach, specifically ment to avoid the problems raised by this arbitrary division of the molecular surface into "polar" and "nonpolar" parts. Since the BEM calculations readily provide the superficial polarization charge densities σ , it is straightforward to use these values as a measure of the "polarity" of the surface. The *sign* of σ , however, is of not relevant here and therefore we propose the following alternative expression for modeling the cavity and hydrophobic effects:

$$F_{\text{surf}}^{\text{solv}} = \kappa_{\text{total}} S_{\text{total}} + \kappa_{\sigma} \sum_{\text{exposed atoms}} (\sigma_i S_i)^2. \quad (13)$$

Certain authors⁴⁹ suggest that the cavity shape used to describe the electrostatic phenomena might not be appropriate to account at the same time for the "hydrophobic effects" and therefore *different* sets of atomic radii should be used to build these interfaces. Here, we have considered the same molecular cavity to describe both phenomena in order to reduce the number of parameters of the model. We expect the fit of the atomic radii to define this unique "best compromise" cavity.

C. Molecular geometries

To obtain the geometries for the 64 small molecules used in the cross-validation procedure, vacuum molecular dynamics (MD) simulations are carried out for each molecule. The MD simulations are performed using the Discover Package⁴² and consisted of 40 ps equilibration runs at 300 K followed by 200 ps for sampling. Only 100 conformations saved during the last 80 ps of the MD simulation are used in this study.

The aim of the MD simulation is to generate a set of molecular geometries that can be taken as representative statistical ensembles of both solvent and vacuum geometries. The solvation energies for all sampled conformations are estimated *a posteriori* from the vacuum trajectories, and the intrasolute contribution to free energies are represented by the conformational energy computed with Biosym force field.^{42,43} The transfer free energies are computed as the differences between the weighted averages of the free energies of a conformational ensemble of the solute in water and solvent respectively. The same sets of conformations were used together with different values of the atomic partial charges and radii to assess their influence on the predicted transfer energies.

Single geometries minimized by MOPAC-AM1 have been used for the large set of 173 molecules.

D. Setup parameters: Implementation for macromolecules

The choice of the setup parameters (the density of the surface dots N_{dens} , the "halving square distance" D_{cut} and the "exclusion distance" δr) will always reflect a compro-

mise between accuracy and computational effort. Since the former decreases (due to an accumulation of local errors) and the latter increases with the size of the molecule, a convenient implementation of the method for macromolecules will be inherently more difficult to achieve. There is no interest in evaluating the transfer energies of proteins and such an attempt is unrealistic since the accumulations of small systematic errors for the large number of functional groups might easily exceed tens of kilocalories. On the contrary, the method should be able to accurately reproduce *the variations of the polarization energy due to conformational changes or ligand binding*. The role of the setup parameters would be to calculate these changes in energy as precisely as possible but at lower computer time costs.

We have performed high-precision calculations ($N_{\text{dens}}=600$; $D_{\text{cut}}=999$) on a series of 10 geometries of the enzyme barnase⁵² (110 residues, 1700 atoms including aliphatic hydrogens). These geometries were taken from a molecular dynamics simulation of this protein in water,⁵³ the time separation between each geometry being 2×10^{-15} s, which is a conventional integration time step in molecular dynamics of solvated proteins. Such a small integration time step makes the 10 geometries very similar. The results of lower-precision calculations on the same set of geometries were then correlated to the reference energies by means of a linear regression with an imposed slope of 1.0, but with a free intercept. The runtime per geometry has been monitored in function of the setup parameters.

The same procedure has then been repeated on eight geometries of crambin,⁵⁴ retrieved from the Brookhaven Protein Data Bank. These geometries represent eight possible solutions satisfying the NOE-NMR constraints and therefore display more variability than successive conformations taken from a molecular dynamics simulation. We also compare in this case, the rms errors due to low precision BEM calculations. In order to save computer time, a common practice in molecular dynamics and Monte Carlo simulations in the use of a spherical cut-off which neglects the interaction between two atoms if their distance is larger than a certain cut-off distance. The errors introduced by this procedure can be in some way minimized by the use of switching and termination functions multiplying the Coulomb term.^{55,56} The rms errors due to the truncated and modified Coulomb terms are also investigated.

E. Discussion of the fittable parameters of the solvation model

In this work, we considered the following physical parameters commonly used in a continuum solvent model: the *atomic radii* and *probe size*, defining the dielectric interface, the *effective dielectric constants*, and the *surface tensions* characterizing the nonelectrostatic effects. The approach we took to determine their values is by fitting computed solvation energies to experimental measures of transfer free energies from gas phase to water.

An interesting, different approach presented in another study³⁷ considered *fittable bond dipole values from which a*

new set of atomic charges is derived. The use of this set of charges is shown to lead to an almost perfect coincidence between calculated and experimental vacuum to water transfer energies. Our choice to consider the *atomic radii* and the *probe radius* as fittable parameters is the fact that the sharp dielectric interface used in this model has no real physical meaning. These are new parameters, that can be added to any existing empirical force field. The reassignment of atomic charges is subjected to much more constraints since not only the transfer energies, but also the molecular dipole moment must be reproduced. Atomic charges should in principle be derived from the molecular wave function and the charge parameters in force fields might not be altered unless the other terms are changed in order to reestablish its consistency. The use of *different* sets of charges to evaluate the Coulomb and the polarization terms would be a questionable approach. It is easier to find a justification for the use of a new set of radii that are different from the van der Waals values in force fields, since the polarization phenomena described by the first are in no way correlated to the nonbonded interactions described by the latter. From a practical point of view, the number of fittable parameters scales as the square of atom types in case of bond dipoles, but simply equals the number of considered atom types when working with atomic radii. In return, the computed results will certainly be less perfect in the latter.

To our opinion, taking the atomic radii from existing molecular force fields *as they are* does not guarantee the physical consistency. The fact that in this way one can obtain reasonable results suggests that the BEM calculation is a quite robust algorithm that gives correct answers for a fairly large range of parametrizations (vdW radii, crystallographic radii, radii from force fields). Nevertheless, we wished to raise here the question whether—and to what extent—a systematic search for an optimal set of atomic radii may improve the quality of the results. A further incentive for performing such an optimization is the possibility that contributions from effects that are not explicitly treated by this simple continuum model, such as local inhomogeneities of the external dielectric properties or charge redistribution during the solvation process could be mimicked by altering the radii of the involved atoms.

The assignment of atomic radii in function of the potential types might not be sufficiently detailed to ensure good results. The effective radius of an atom can depend on the charge, since, i.e., a charged atom attracts solvent molecules and therefore decreases this radius, while this strong attraction might lead to dielectric saturation in the vicinity of the atom and hence to a locally lower dielectric constant and a larger apparent radius. We cannot tell which effect will prevail, but we can suppose that

$$R_i = R_i^0 + \lambda f(Q_i), \quad (14)$$

where $f(Q_i)$ may be a simple expression like Q_i^n , $|Q_i|$. It is maybe redundant to introduce a dependence on both the potential type *and* the charge of the atom, since the point charges in force fields are themselves assigned in terms of

potential types. Again, the opportunity of introducing the fittable parameter λ must be validated in terms of improved accuracy.

The values of the atomic partial charges used in the calculations are key parameters of any solvation model. To investigate their influence on the solvation energies obtained with our model, the BEM calculations are repeated using the Biosym^{42,43} and MOPAC-AM1⁴⁴ partial charges respectively. A third set of charges is obtained by MOPAC-AM1 calculation using the ESP option.⁴⁵ Since the sets of atomic charges are taken over as such, it is clear that they must be combined with different sets of radii to obtain the same transfer energies.

The continuum description of solvation proved to be successful in predicting vacuum-to-water energies, validating the point of view that in general the description of the solvent as a homogeneous dielectric medium is quite realistic. Nevertheless, the choice of the effective values for the dielectric constants ranges from 1.0–4.0 for the *internal dielectric constant* and from 10.0–80.0 for the *dielectric constant of water*,⁴⁹ such low values being attributed to the bound water molecules at protein surfaces. Furthermore, in a completely rigorous BEM treatment, the Coulomb term should be in $1/\epsilon_{\text{int}}$ and cancels out in the expression of the transfer energy. The polarization terms should be roughly proportional to the *difference in the inverse internal and external dielectric constants*. However, the BEM model studied here makes a series of quite drastic simplifications and we do not know whether it would still produce reasonable values for the transfer energies, whether these values would be completely wrong or whether they would be different, but *correlated* to the experimental values, a situation in which the predictive power of the model might be conserved even if its parameters are no longer equal to the theoretical values (see below).

Therefore, in order to account for the uncertainties in the choice of dielectric constants and for the presumed problems resulting from the drastic simplifications of the BEM model, Eq. (11) is written as a linear combination:

$$F_{\text{elec}}^{\text{ext}} = \alpha_{\text{Coul.}}^{\text{ext}} U_{\text{Coul.}}^0 + \alpha_{\text{pol}}^{\text{ext}} F_{\text{pol}}^{\text{ext}}. \quad (15)$$

Initially, the values of ϵ_{int} , ϵ_{solv} , ϵ_{ext} used to calculate F_{pol} by the BEM are set to 2, 78, and 1 respectively. Given that $F_{\text{pol}}^{\text{ext}}$ depends linearly on $(1/\epsilon_{\text{ext}} - 1/\epsilon_{\text{int}})$, the α_{pol} coefficients represent the ratio between the *effective* and the *initially estimated* differences of inverse dielectric constants. Fitting this quantity indirectly leads to the reassessment of the dielectric constants. Indeed, if the initial choice is reasonable then $\alpha_{\text{pol}}^{\text{vac}} \approx \alpha_{\text{pol}}^{\text{solv}} \approx 1$. The factor multiplying the Coulomb term is theoretically independent of the external dielectric constant, thus $\alpha_{\text{Coul.}}^{\text{vac}} = \alpha_{\text{Coul.}}^{\text{solv}} = 1/\epsilon_{\text{int}}$. However differences between $\alpha_{\text{Coul.}}^{\text{vac}}$ and $\alpha_{\text{Coul.}}^{\text{solv}}$ could in principle compensate for errors in the model.

The nonelectrostatic part of the solvent model described by Eq. (12) contains the two fittable ‘‘surface tensions’’ κ , for which an estimation is hard to make. Quite often,⁵⁷ these parameters are obtained from a separate fit of transfer energies vs surface of *nonpolar* molecules, where the electro-

static term is considered to be negligible. This assumption is questionable, since this electrostatic contribution can be quite important as compared to the nonpolar terms (~30%–40%). Moreover, all the parameters introduced here are intercorrelated and should be optimized simultaneously. In order to have an unified approach, we consider that the same molecular boundary should be used to describe both the *electrostatic* and the *nonelectrostatic* solvent effects. Of course, increasing the number of distinct atomic types should lead to a finer parametrization and implicitly to a better quality of the model. On the other hand, increasing the number of parameters of the model increases the chances of fitting artifacts. The final number of parameters must be chosen as a compromise and it should be reasonably low as compared to the number of molecules in the learning set, while being as well dependent on the simplifying hypotheses introduced in each model: More rigorous models should require less parameters (but more computer time). Statistical methods must be used to validate the significance of the parameters entering the model, however such an approach is highly cumbersome due to the nonlinearity of the involved functions.

In the first class of models, the atomic radii are optimized using as initial values the van der Waals radii of the CVFF—Biosym force field.^{42,43} In contrast to the dielectric parameters ϵ_{ext} , ϵ_{int} , which can be adjusted *after* the calculations of the electrostatic terms because these depend linearly on $1/\epsilon$, the influence of the atomic radii on the polarization term does not follow any simple rule. For each one of the 64 molecules of the first learning set, the BEM calculations were ran on 5 out of 100 sampled conformations from MD simulations. The energies of the individual conformations are then fed into the nonlinear fitting procedure which optimizes the six weighting factors (see below), and returns a cross-validated rms value for the current point. Due to the complexity of the function to be minimized, the applied technique is a systematic grid search, in which the radii, defined for 14 different atom types, plus a probe radius for water (see Table V), are changed by increments at each step.

Based on the conclusions emerging from the previous results, we then proceeded to the development of a more elaborated model that has been calibrated on a large set of 173 molecules,³⁶ referred as the second class of models. Equation (13) is used to describe nonelectrostatic corrections and a charge-dependent term as in Eq. (14), with $f(Q) = Q$ is applied to the radii. A more detailed definition of potential types is introduced, while extending at the same time the parametrization to other chemical elements, not represented in the smaller set of molecules. Since the nonlinear fitting procedure did not bring any significant improvement, single geometries of the molecules have been used instead of ensembles of conformations and the calculated transfer energies were *linearly* related to the experimental data.

The electrostatic weighting factors have been fixed to their theoretical values. The Coulomb term has been dropped from Eq. (15) and the coefficient of the polarization term fixed at 1.0. This led to an overall larger set of 27 parameters, while reducing the ratio of fittable parameters vs experimental points. This fit was performed only with MOPAC

atomic charges, since the set of 173 molecules contained some “exotic” ionic species that failed to be parametrized under the Biosym-CVFF force field.

F. The nonlinear fitting procedure

This nonlinear fitting procedure is designed to assess whether it is justified to calculate vacuum to water energies on the base of a single molecular geometry or whether Boltzmann averaging over solvated and vacuum ensembles of conformations improves the calculated values. The energy of compound i adopting the conformations j in an external medium ext is a linear function of $n = 6$ parameters (in vacuum, the two surface coefficients are set to zero):

$$F_{ij}^{\text{ext}} = E_{i,j}^0 + \sum_{k=1}^n x^{\text{ext},k} A_{i,j}^{\text{ext},k}, \quad (16)$$

where E^0 stands for the solvent-independent intrasolute terms such as the usual covalent and nonbonded contributions and A represent the different calculated molecular properties [Eq. (15) and (12)] weighted by the corresponding fittable parameters x . Given a set of molecular geometries j , the average energy of compound i becomes:

$$\langle F_i^{\text{ext}} \rangle = \frac{\sum_j F_{i,j}^{\text{ext}} \exp(-F_{i,j}^{\text{ext}}/k_B T)}{\sum_j \exp(-F_{i,j}^{\text{ext}}/k_B T)}, \quad (17)$$

with k_B being the Boltzmann constant and T the temperature.

The vacuum-to-solvent transfer energy is then defined as the difference of Eq. (17) between ext=solvent and ext=vacuum; the rms between the experimental values $\Delta F_i^{\text{transf}}$ and this difference is:

$$\begin{aligned} \text{rms} &= \text{rms}(x_k) \\ &= \sqrt{\frac{1}{N_{\text{comp}}} \sum_i^{N_{\text{comp}}} (\Delta F_i^{\text{transf}} - (\langle F_i^{\text{solv}} \rangle - \langle F_i^{\text{vac}} \rangle))^2}. \end{aligned} \quad (18)$$

Calibration of the solvent model consists in finding the set of coefficients x that minimizes the function (18), which is nonlinear in the variables x . The *cross-validated* rms value is obtained by running a series of $n_f = 10$ nonlinear fits with respect to different subsets of compounds referred to as the “learning sets,” and using the obtained coefficients to predict the transfer energies of the $n_e = 6$ or 7 excluded molecules that were not used in the fit. The sets of n_e molecules are chosen such that every molecule is excluded and predicted once during the fit runs. The cross-validated rms value is taken as the root-mean-square deviation between *predicted* and *experimental* transfer energies. The fluctuations of the fittable coefficients upon changing the set of considered molecules are also monitored in order to assess the consistency of the model.

III. RESULTS AND DISCUSSIONS

In this section the various aspects of our solvation model are analyzed. First we evaluate key features of our time saving approximations to the BEM by investigating their ability to reproduce measured transfer free energies. This is fol-

lowed by a discussion of the predictive power of our model when the parameters are fitted in the presence of different sets of atomic radii and partial charges, with and without optimization of the atomic radii, and as a function of a number of technical options.

A. Neglecting the coupling between the polarization charge densities

The first important finding of our analysis is that runs of the BEM calculations including or neglecting the nondiagonal BEM matrix elements respectively lead to practically identical results. These calculations performed on all 64 molecules with the *Biosym* charge set lead to $F_{\text{complete}} = 1.0 F_{\text{diagonal}}$, with a root-mean-square deviation (rms) of 9.1×10^{-3} kcal/mol and a correlation coefficient (r^2) of 1.0. F_{complete} and F_{diagonal} are the polarization free energies computed with and without the nondiagonal terms respectively.

Moreover, there is no significant change in any of the fittable parameters of the model—neither in the linear weighting coefficients, nor in the atomic radii—upon neglecting the nondiagonal terms. Thus, the diagonal BEM algorithm, which is up to an order of magnitude faster than the classical one, entails no trade-off on the quality of the results. Hence in what follows, no distinction will be made between results obtained with the complete or diagonal versions of the BEM method.

We tried to explain this surprising result by considering a simple model system of two interpenetrating atoms of equal radii (1 Å), with an unitary charge located at the center of the first atom. When the interatomic distance is close to zero, the model system degenerates to the case of a single sphere divided into two equal BEs and we can evaluate the normal projection of the field produced by the polarization charge density on the surface spawned by the angle $\theta = [0, \pi]$ at the pole $\theta = 0$:

$$\mathbf{E} \cdot \mathbf{n}_\theta = \frac{\sigma}{2\epsilon_0} \left(1 + \sin \frac{\theta}{2} \right). \quad (19)$$

The term used in our BEM must be somewhat lower than this, since it is averaged over all dots of the BE and not taken only at the pole point. For the considered hemisphere, the field at the pole point is $(\sigma/2\epsilon_0)(1 + \sqrt{2}/2)$, while the field produced by the charge distribution of this BE at a point on its edge is only $(\sigma/2\epsilon_0)(1 + 1/2)$. Therefore, we will consider this "self field" to be equal to $(\sigma/2\epsilon_0)(1 + \alpha)$, where the coefficient α is in the range of 0.5–0.7. This is the effective term used in the BEM when the cross interactions with the σ of the other hemisphere are ignored, otherwise its value would have been σ/ϵ_0 .

Accordingly, the estimated σ would be

$$\sigma = \frac{Q}{4\pi R^2} \frac{2}{\epsilon_{\text{int}}} \left[\frac{\epsilon_{\text{int}} - \epsilon_{\text{ext}}}{(1 - \alpha)\epsilon_{\text{int}} + (1 + \alpha)\epsilon_{\text{ext}}} \right]. \quad (20)$$

The value of the rescaling coefficient is therefore: $(1/2\epsilon_{\text{ext}})[(1 - \alpha)\epsilon_{\text{int}} + (1 + \alpha)\epsilon_{\text{ext}}] \sim (1 + \alpha)/2$, which is in very good agreement with the values reported here. If we increase the interatomic distance, the angle θ , within which the self-field of the polarization charge is accounted for, also in-

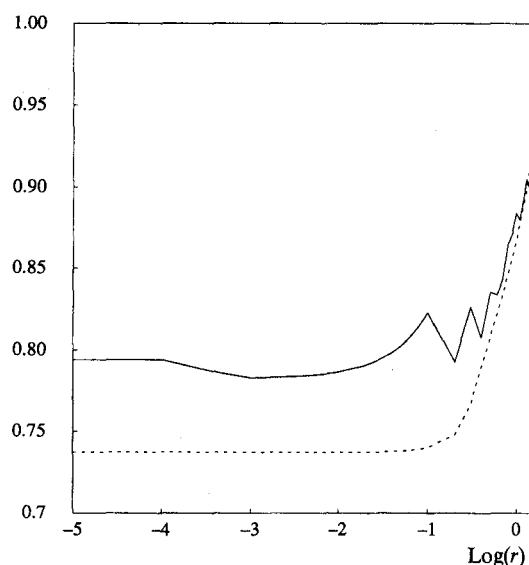


FIG. 3. Rescaling factor for two interpenetrating spheres as a function of the logarithm of the interatomic distance r . Including the nondiagonal BEM matrix elements (solid curve), neglecting the coupling between the polarization charge densities on different BE (dashed curve).

creases. Furthermore, the contribution that was neglected because the integration was not continued up to $\theta = \pi$ is certainly bigger than the influence of the polarization charge of the even more remote surface of the second sphere. Therefore, the rescaling coefficient is expected to increase with the interatomic separation and this is indeed what we find (see Fig. 3).

Repeating now the same procedure with a BEM computation that explicitly includes the influence of polarization charges on the field at other BEs, we would expect much higher (close to unity) rescaling factors. The obtained values are however only slightly higher (Fig. 3) and we ascribe this partly to the fact that certain dot pairs that are very close to the kink at the intersection of the spheres need to be excluded. Due to these kinks of the surface, the cross terms appear to be those that are most sensitive to the chosen dot distribution and introduce important numerical errors. When the separation between the centers of the spheres is increased so that we obtain the typical geometry of a diatomic molecule, the inter-BE influence eventually becomes negligible relative to the self-influence of the σ from the current BE (Fig. 3).

In this case, neglecting the coupling coefficients of the polarization charge densities on different BEs must not be understood in the sense of completely ignoring the electric field produced by them. In the methodology developed here, the BEs have the unusual feature of being large enough so that the dominant influence felt by the polarization charge comes from the polarization charge of the same BE. In other words, the largest part of the nondiagonal terms that would appear in a classical BEM approach based on small BEs are not ignored, but forced into an average diagonal term. Also, if an atom is poorly accessible to solvent, its BE will be small and will hardly contribute to the polarization energy.

TABLE I. The dependence of the computed polarization free energy (kcal/mol) of the protonated Arginine side chain in solvent as a function of the considered dot densities of the solvent-accessible surface and the exclusion distance δr , between dots. In parenthesis, are the required rescaling factors of the surface charge densities.

Surface dot density (points per 3 Å—sphere)	Exclusion distance δr between two points on different BEs	
	0.2 Å	0.5 Å
100	-28.2 (0.73)	-28.2 (0.73)
200	-28.5 (0.71)	-28.5 (0.72)
300	-28.3 (0.74)	-28.3 (0.74)
400	-28.2 (0.72)	-28.3 (0.73)

The coupling phenomena that are effectively neglected and that would have shielded the influence of the atomic charges, add up to the other numerical errors leading to σ values which deviate from Gauss' theorem, are in some way compensated by the rescaling procedure.

B. Setup parameters and numerical errors: Rescaling the charge densities

In agreement with previous observations,³² we find that the BEM algorithm tends to overestimate the charge densities. To obtain agreement with Gauss' theorem, these charge densities must be rescaled by a factor ranging between 0.7 and 0.8 in all the molecules studied. To investigate this issue, we analyzed the influence of charge density rescaling on the polarization free energy in water of a charged solute such as the arginine side chain. Table I illustrates how the calculated polarization free energies and the charge density rescaling factors vary in this case as a function of the surface dot density and the exclusion distance δr between points on different BEs. It shows that not much will be gained by increasing the dot density N_{dens} beyond 400 dots/3 Å radius sphere, or taking δr values smaller than 0.5 Å. The remarkable independence of the results of the *full* BEM calculation on the "exclusion distance" δr actually led us to the important conclusion that the interactions between polarization charges on different BEs could be neglected. For all the small molecules, the halving square distance D_{cut} has been chosen large enough to ensure that all the BE points are always taken into account.

Next we compare the polarization energies F in water computed with and without rescaling the polarization charges, with F values computed as averages over the 100 sampled geometries for each compound. This was done for all 64 compounds, yielding:

$$F_{\text{resc}} = 0.736F \quad (\text{rms} = 0.335; r^2 = 0.999),$$

where F_{resc} and F correspond to the polarization energies with and without rescaling procedure of the charge densities respectively.

Given that the weighting factor applied to the unrescaled F 's in our solvation model [Eq. (15)] is a fittable parameter, the small rms deviation and the excellent correlation coefficient of the fit suggests that either F or F_{resc} could be used.

TABLE II. Cross-validated rms deviations of calculated versus experimental transfer free energies (in kcal/mol) as a function of different combinations for the probe radius in solvent and vacuum. Atomic vdW radii are taken from the Biosym parameter set. To obtain an initial estimate for the probe radii in vacuum and solvent a preliminary fast optimization was done by means of a grid search procedure using Biosym atomic van der Waals radii.

Probe radius solv\vac	-0.3 Å	0.0 Å	0.3 Å	0.6 Å	0.9 Å
0.5 Å	2.44	2.35	2.36	2.43	2.47
0.8 Å	2.44	2.36	2.40	2.48	2.53
1.1 Å	2.46	2.36	2.41	2.51	2.58
1.4 Å	2.48	2.39	2.44	2.55	2.62
1.7 Å	2.50	2.40	2.46	2.57	2.65

This findings show that reliable estimates of the rescaled polarization free energies can be obtained from computations of unrescaled energies. This could have useful applications in cases which involve boundaries made of open surfaces, such as a limited domain of a macromolecule or when introducing a cut-off distance.

C. Optimal sets of parameters

1. Preoptimization of the solvent probe radius

The results of a preoptimization procedure on the small set of molecules in which the electrostatic weighting factors are fitted and the probe radius systematically varied, while the atomic radii are fixed at the values taken from the *Biosym* atomic radii set, are summarized in Table II. They show that in vacuum the optimal probe radius is 0, while in the solvent an estimate of 0.5 Å is obtained instead of the classical 1.4 Å, commonly taken as the water radius. This is a clear indication that the "dielectric boundary" does not coincide with the accessible surface, but is actually much closer to the *molecular* surface.

2. The parameters of the first class of models

The final values of the atomic solvent probe radii obtained by the full optimization within the first class of models (in which the weighting factors as well as all the 14 radii are varied) are listed in Table III. This table lists the results of three independent optimization runs, each one carried out with a different set of partial charges, in order to investigate the dependence of the radii values on the charge distribution. It must be pointed out that the complexity of the functions to be minimized makes a search of an absolute minimum impossible. The fact that the optimization was conducted by moving a single variable at once sometimes blocked the algorithm at points that were not actual minima. Changing the order of moving the variables and/or relaunching the optimization from a slightly different point were therefore used to test whether the obtained points were not optimization artefacts. The results we report here are the best points detected. Another issue of interest is whether these parameters actually increase the predictive power of the model or were artificially fitted to optimize the response of the molecules from

TABLE III. Optimal atomic radii (in Å) for Biosym atom types and for different sets of charges (Biosym, MOPAC and MOPAC-ESP).

Atom	VdWaaals radius (Å)	Biosym atom types	Optimized radii (Å)		
			BIO	MOPAC	ESP
C	1.55	aliphatic sp ³ carbon (c c1 c2 c3 ca cg)	1.26	0.69	0.95
		sp ² and sp carbon (c=ct)	1.40	1.14	1.61
		aromatic and heterocyclic carbon (cp c5 cr)	1.15	1.17	1.51
		carbonyl and carboxyl carbon (c' c ⁻)	1.82	1.84	1.58
H	1.10	aliphatic hydrogen h	1.09	1.28	1.31
		N-bound hydrogen hn	1.44	1.49	1.41
		O-bound hydrogen ho	1.25	0.76	0.87
N	1.35	sp ² nitrogen (n n2 n1 np)	1.10	0.59	1.14
		sp ³ nitrogen (n3 n4)	1.11	1.14	1.13
O	1.40	alcoholic and etheric oxygen (o oh)	1.02	0.97	0.88
		carbonyl oxygen o'	0.79	0.98	0.82
		carboxylate oxygen o ⁻	1.55	1.58	1.44
F	1.30	all fluorine atoms	1.50	1.70	2.18
S	1.81	sulfur in disulfides and thiols	2.72	3.38	3.60
*	0.5	**probe radius in solvent	0.231	0.271	0.237

the learning set. This problem is specifically addressed by the use of a cross-validated comparison of calculated vs experimental values.

Inspection of the values of the atomic radii (Table III) indicates that except in some cases, such as the sp³ carbon, the OH proton, or sulfur atom, very similar radii are obtained with the three partial charge sets. This is indeed verified by the fair correlation obtained between radii optimized with three charge sets respectively:

$$R^{\text{biosym}} = 0.616 R^{\text{MOPAC}} + 0.54 \quad (\text{rms} = 0.18; r^2 = 0.84),$$

$$R^{\text{biosym}} = 0.589 R^{\text{ESP}} + 0.51 \quad (\text{rms} = 0.19; r^2 = 0.81),$$

$$R^{\text{MOPAC}} = 0.902 R^{\text{ESP}} + 0.02 \quad (\text{rms} = 0.24; r^2 = 0.87).$$

This indicates that the chemical context of the defined atom types is preserved across different molecules and upon changing the partial charges.⁵⁸ The weak dependence on the partial charge set can be explained by similarity of the considered sets. The rms deviations of the atomic partials charges are 0.08 units and 0.14 units between *Biosym* and *MOPAC*, *MOPAC-ESP*, respectively. Finally, it is interesting to note that the probe radii optimized together with the atomic radii, are yet smaller (0.23–0.27) than those obtained with the fixed original *Biosym* radii and charge sets.

The resulting fittable weighting coefficients α and κ [Eq. (15)] are listed in Table IV. It can be seen that the α coefficients [which multiply the electrostatic terms of the transfer free energy in Eq. (15)] fluctuate little during the cross validation process. Their rms fluctuations give the magnitude of the error in the corresponding computed energy terms. For example, a fluctuation of 0.003 in $\alpha_{\text{pol}}^{\text{sol}}$ leads to an upper error limit of 0.12 kcal/mol on the polarization term for the ammonium ion whose polarization energy is 40 kcal/mol. The values of the coefficients α_{pol} multiplying the polariza-

tion terms show a very good agreement with the theoretical prediction of 1, except of course in the case of unrescaled polarization charges (scheme I). In other words, these adjustable parameters could be eliminated from the model and fixed at their theoretical value of 1.0. This is done in the second class of models. However, $\alpha_{\text{pol}}^{\text{sol}}$ and $\alpha_{\text{pol}}^{\text{vac}}$ are strongly intercorrelated with the probe radius value. At a probe radius of 0, the interface for both solvent and vacuum calculations would be the same and therefore only the difference $\alpha_{\text{pol}}^{\text{sol}} - \alpha_{\text{pol}}^{\text{vac}}$ would be relevant. In this particular case, the in-

TABLE IV. Optimal cross-validated linear coefficients obtained from different charge sets, probe and atomic radii used in the BEM calculations. $\bar{\alpha}$ is the cross-validated average of the coefficient α (see below) and rms(α) is its standard deviation during the cross-validation procedure. $\alpha_{\text{pol}}^{\text{sol}}$, $\alpha_{\text{pol}}^{\text{vac}}$ are respectively the linear weighting coefficients of the polarization term in solvent and vacuum, and $\alpha_{\text{Coul}}^{\text{sol}}$, $\alpha_{\text{Coul}}^{\text{vac}}$ the Coulomb term coefficients in the solvent and vacuum [see Eq. (15)]. κ_{total} and κ_{nonpolar} are the surface term coefficients as defined in Eq. (12).

	Tested BEM initial conditions					
	I	II	III	IV	V	VI
$\bar{\alpha}_{\text{pol}}^{\text{sol}}$	0.869	0.926	0.969	0.979	1.073	1.024
rms($\alpha_{\text{pol}}^{\text{sol}}$)	0.010	0.008	0.006	0.003	0.002	0.003
$\bar{\alpha}_{\text{pol}}^{\text{vac}}$	0.950	0.926	0.977	0.996	1.103	1.042
rms($\alpha_{\text{pol}}^{\text{vac}}$)	0.006	0.008	0.008	0.003	0.003	0.003
$\bar{\alpha}_{\text{Coul}}^{\text{sol}}$	0.366	0.361	0.381	0.309	0.378	0.361
rms($\alpha_{\text{Coul}}^{\text{sol}}$)	0.003	0.001	0.005	0.001	0.003	0.002
$\bar{\alpha}_{\text{Coul}}^{\text{vac}}$	0.347	0.352	0.333	0.354	0.336	0.303
rms($\alpha_{\text{Coul}}^{\text{vac}}$)	0.003	0.001	0.005	0.002	0.003	0.001
$\bar{\kappa}_{\text{total}}$	-0.032	-0.038	-0.020	-0.041	-0.012	-0.009
rms(κ_{total})	0.003	0.002	0.003	0.003	0.002	0.001
$\bar{\kappa}_{\text{nonpolar}}$	0.047	0.050	0.038	0.064	0.040	0.029
rms(κ_{nonpolar})	0.003	0.003	0.003	0.003	0.002	0.001

ternal dielectric constant cancels out in the expression of the transfer energies.

The exact electrostatic equations state that the internal dielectric constant governing the intramolecular Coulomb term does not depend on the external medium in which the molecule is placed. Accordingly, the Coulomb term should not explicitly appear in the expression of transfer energies and might influence its value only indirectly, as part of the Hamiltonian which in turn determines the probability of appearance of the different conformations. This is only the case when the polarization energies show a strong conformational dependence. The values of the Coulomb weighting factors $\alpha_{\text{Coul.}}^{\text{vac}}$ and $\alpha_{\text{Coul.}}^{\text{solv}}$ could not be established from the nonlinear fit; it is only their difference that has a significance for the calculated transfer energies. The individual coefficients strongly depend on the starting point of the optimization, but the difference between them *does not* and tends to zero in all situations. We conclude that the effect of the Coulomb term on the transfer energies is negligible, since the molecules analyzed here are small and relatively rigid. For the set of molecules this work is based on, the electrostatic weighting factors can be dropped, as done in the second class of models. This however, may no longer be true when complex molecules with internal degrees of freedom are analyzed.

The coefficient of the total area is always negative, while the areas that are labeled as hydrophobic have a larger positive contribution to the solvation term. The κ coefficients of the surface area term (Table IV) fluctuate appreciably, and are thus the less stable parameters of the model. This is less serious for the ionic compounds, where the errors in the electrostatic terms might easily exceed the magnitude of the surface effects. When changing from Biosym to MOPAC charges and using the same set of radii (schemes II and III), the change of the electrostatic coefficients is far less dramatic than the modifications of κ_{total} , whose value is actually halved. This means that on the average the computed energy of van der Waals and cavity creation fluctuates by as much as 3 kcal/mol ($0.02 \text{ kcal/mol}/\text{\AA}^2 \times 150 \text{ \AA}^2$) upon changing parameters that have nothing to do with it. This suggests that the surface area terms in the current solvent models do not represent well-defined physical quantities. However, the fact that errors in the electrostatic term are absorbed in the surface area term is not necessarily an artefact, since there might be some dependence between these errors and the size of the molecule. Using an extra term in S^2 (Ref. 41) slightly improves the quality of the fit, but not enough to justify increasing the number of variables of the model.

3. The parameters of the second class of models

The different choice of atomic types as well as the introduction of the charge dependent radii [Eq. (14)] and the different modeling of nonelectrostatic effects [Eq. (13)] prevent a direct comparison of the radii issued from our second optimization experiment. These "radii" together with the other parameters characterizing the second class of models are listed in Table V. The atomic radii listed in Table V must be corrected by an amount of λQ before using them.

TABLE V. Optimized parameters for the set of 173 molecules, with MOPAC charges, nonpolar term as in Eq. (14), with $f(Q)=Q$.

Atom type	Obtained values
Nonpolar hydrogen	0.91
Polar hydrogen in neutral groups	0.65
Polar hydrogen in cations	0.86
sp ³ carbon atom	1.19
Carbonyl carbon	0.94
Unsaturated carbon in alkenes, alkynes	1.46
Aromatic carbon	1.28
Carbon in carboxylate groups	0.96
sp ³ nitrogen in amines	0.52
Nitrogen in ammonium groups	1.86
Nitrogen with delocalized lone-pair	0.90
Nitrogen in non-aromatic multiple bonds	1.05
sp ³ oxygen	0.73
Oxygen in carboxylate groups	1.08
Oxygen in carbonyl groups	0.90
Oxygen in esters and -COOH groups	2.18
Protonated oxygen	2.21
Sulphur atoms	1.76
Phosphorus atoms	2.15
Fluorine atoms	1.85
Chlorine atoms	0.74
Bromine atoms	0.43
Iodine atoms	0.19
**Probe radius	0.82
Charge dependence factor λ from Eq. (14) (in $\text{\AA}/ e $)	-0.23
κ_{σ} from Eq. (13) (in kcal/mol/ $ e ^2$)	-23.1
κ_{total} from Eq. (13) (in kcal/mol/ \AA^2)	0.013

Interestingly, in this second class of models, the probe radius of 0.8 \AA is much closer to the commonly used value of 1.4 \AA . The *negative* value of $\lambda = -0.23$ implies that the charge dependency correction of the radii might reach up to 0.1 \AA and is qualitatively in agreement with the trend that negative atoms have an excess of electrons and therefore a larger radius. If saturation or compression effects would be exceedingly important, than an alternative model of charge-dependent radii in $\lambda|Q|$ might be more appropriate. This alternative has been tested and is found to be less satisfactory.

The positive value of the total surface coefficient κ_{total} of 0.013 kcal/mol/ \AA^2 is in good agreement with the results of other authors.^{34,57} The negative κ_{σ} counterbalances the unfavorable *total* surface solvation contribution ($\kappa_{\text{total}}S > 0$) for the polar regions of molecules, setting the real surface contribution of an atom to $(\kappa_{\text{total}} - |\kappa_{\sigma}| \sigma^2 S)S$ [see Eq. (13)]. Very surprising is the *strong decrease* of the halogen radii $F > \text{Cl} > \text{Br} > \text{I}$ (the charge dependence will emphasize this trend!). If the calculated MOPAC-AM1 charges on halogens are correct—for example, the charge on the I atom in methyl iodide is practically zero (+0.06 vs +0.11 $|e|$ on hydrogens), this means that the solvation of such species imply phenomenons that are not correctly described by the continuum solvent model. Whatever the nature of such processes—alkyl halides are known for the solvent-induced polarization of the C–X bond, eventually leading to heterolytic dissociation⁵⁹—the fitting procedure tried to take these

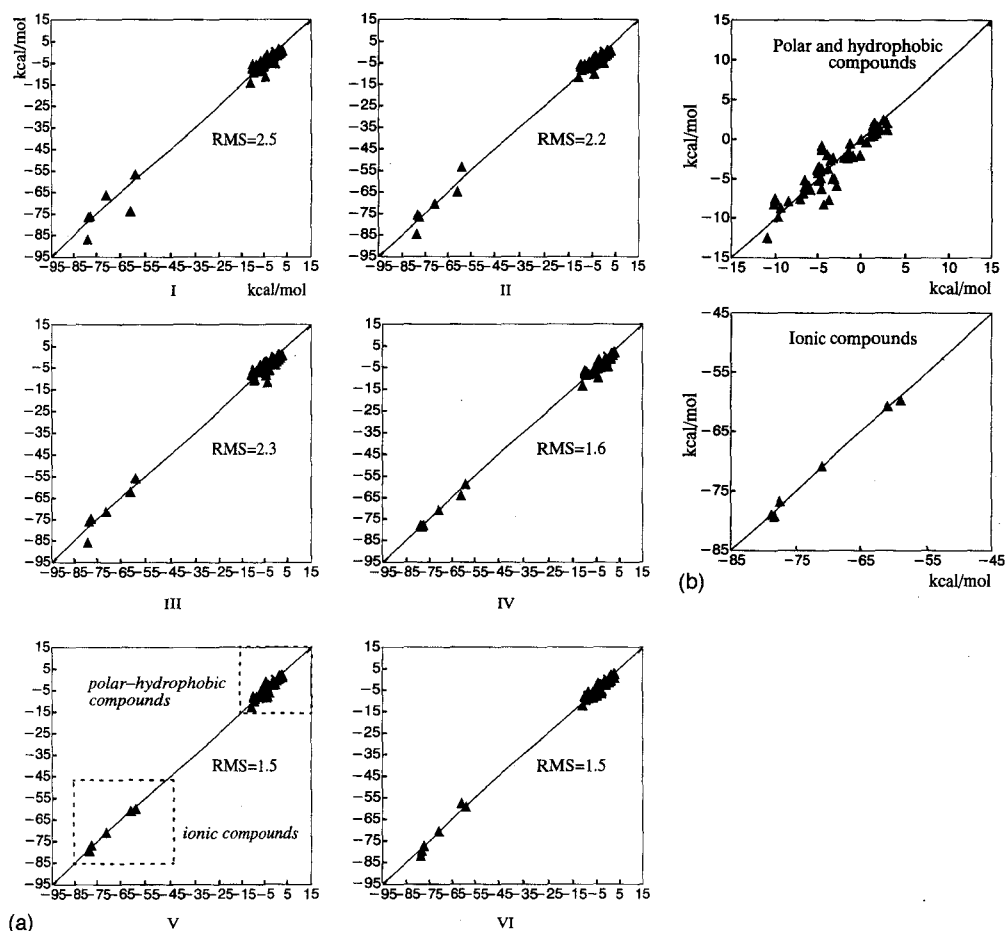


FIG. 4. (a) Predicted vs experimental vacuum-to-water transfer energies using different charge sets, atomic and probe radii for 6 ionic and 58 hydrophobic and polar compounds. rms values are in kcal/mol. (b) Detailed predicted vs experimental vacuum-to-water transfer energies correlation plot for scheme V (rms=1.5 kcal/mol). I—BEM run with Biosym charges and van der Waals atomic radii (probe radius set to 0.4), *without* rescaling charge densities; II—BEM *with* rescaled charge densities, run under the same conditions as I; III—As in II, but using MOPAC-AM1 atomic charges; IV—As in II, but the radii were optimized with the Biosym charge set; V—As in III, but the radii were optimized with the MOPAC charge set; VI—As in II, but the radii were optimized using the MOPAC-AM1 charges obtained with the ESP option. The experimental vacuum-to-water transfer energies were those used by other authors in similar studies (Refs. 13, 34, 41, and 36). Some values for the amino acid side chains were taken from (Ref. 60). The transfer energy for the *protonated* Arginine side chain (~ -61 kcal/mol) was estimated on the basis of a thermodynamic cycle using the deprotonation free energies of the propylguanidinium moiety in vacuum and solvent, and the transfer energy of the *neutral* propylguanidine. The deprotonation free energy in vacuum was calculated by MOPAC-AM1, while the one in solvent is known from the pK_a value of propylguanidine.

effects into account by artificially decreasing the radii of the halogens.

D. The predictive power of the BEM model

1. First class of models

The correlation plots between experimental and computed transfer free energies for the learning set of 64 molecules are displayed in Fig. 4(a)(I–VI). The computed values were obtained using different partial charge and radii sets, with or without radii optimization or charge density rescaling, constituting a total of six initial conditions for the BEM calculations (denoted I–VI).

Inspection of Fig. 4 clearly shows that the computed values correlate better with the experimental ones when optimized radii are used in the calculations [Figs. 4(a)IV–4(a)VI]. The radii optimization under MOPAC charges yields transfer energies in closest agreement with the experi-

mental values (rms=1.53 kcal/mol) as seen from Fig. 4(a) and 4(b). Interestingly, the MOPAC-ESP parametrization (VI) yields a somewhat higher rms value (rms=1.55 kcal/mol). The rms obtained for all six parametrization experiments, ranging from 1.5 to 2.5 kcal/mol are on the whole quite satisfactory. This is particularly true for the ionic compounds, where they are of the same order as the error in the experimental transfer energies.⁶⁰ Table VI is a sample output of predicted vacuum-to-water transfer energies for an extended set of molecules (the compounds that were present in the original learning set are displayed in bold) with Biosym charge parametrization and optimized radii.

2. Second class model

The transfer energies for the 173 molecules calculated with MOPAC charge distribution could be predicted with an rms error of 1.35 kcal/mol by the second class model. This

could be achieved at the expense of only six supplementary fittable parameters, from which four had to be introduced due to the presence of the elements Cl, Br, I, and P which were either absent from the restricted first set of molecules or considered with a radii fixed at the force-field values. The success of this second class model is therefore due to the introduction of the charge dependence parameter λ [Eq. (14)] and to the implicit distinction between polar and apolar surfaces made by the use of the superficial polarization charge σ introduced in the κ_σ correction term [Eq. (13)].

In a recent quantum study made on the same set of molecules,³⁶ the authors reported quite comparable rms values ranging from 0.8 to 1.3 kcal/mol for the neutral compounds only, while we obtain for the same molecules 1.30 kcal/mol. On the other hand, the situation is quite different for the ionic species. We measure for these compounds a rms of 1.9 kcal/mol, a much lower value than the 3.9–5.6 kcal/mol rms published in.³⁶ The simplified solvation model presented here proves to be surprisingly effective in comparison to this quantum study, which can adjust charge distributions to the dielectric reaction field and uses a much more elaborate model with a large number of fittable parameters.

3. Discussion of the worst predicted transfer energies

The optimization of the atomic radii still leads for certain compounds to transfer energies that are far from the experimental measurements. Three explanations can be drawn in such situations.

- (1) The charges taken from the empirical Biosym-cvff force field are a poor description of the real charge distribution of the molecule, but the use of charges from semiempirical MOPAC calculations improves the situation. Within the first class of models, this seems to be the case for several fluorinated compounds and some heterocyclic systems, which are at the same time the main outliers of the regression line. The worst predictions obtained with Biosym charges are for trifluoroethanol, difluoroethane and methylimidazole, where the errors are -5.2 , -4.4 , and $+3.9$ kcal/mol, respectively. Indeed, these errors decrease when using MOPAC charges, and are practically eliminated when using MOPAC-ESP charges.
- (2) None of the charge distributions leads to results that are close to experiment. The problem is not actually linked to the charge distributions, but to the way in which the assignment of radii has been done in terms of Biosym potential types. This can be the problem encountered with the amine molecules and will be discussed in the next section. These errors are typically less than 4 kcal/mol.
- (3) The charge distributions are inaccurate because the electron densities of the solute change when crossing the vacuum–water interface. We already pointed out such a case when discussing the radii of halogens in the second-class model. However, the observed decrease in radii successfully compensates for this effect. Halogenated molecules are quite accurately predicted by the second-class model: the calculated transfer energy of methyl io-

dide reproduces perfectly well the experimental value (-0.9 kcal/mol). It is interesting to note that ethyl bromide is mispredicted by $+0.32$ kcal/mol only, while phenyl bromide by -1.1 kcal/mol. The artificially reduced radius of the Br atom accounts well for the supplementary C–Br bond polarization in water, but is wrong for the deactivated aromatic halide where no such polarization can be evidenced. The same two chlorine derivatives are in error of ~ 0.5 kcal/mol since alkyl–Cl bonds are much less polarizable in water than their brominated or iodated analogs.

4. The total number of fittable radii

The atomic radius of any atom will depend on the chemical context in which this atom is placed. Force fields encode this chemical context information by assigning “potential types” to each atom in a molecule. In order to keep the number of fittable radii as low as possible, we have defined larger chemical contexts as sets of related potential types and we have used a fittable radius for each such context. The definitions of such sets is definitely arbitrary and might lead to artefacts. This could be the case for the bad predictions observed in the series of aliphatic amines. Since in the first class of models, we consider the same types “sp³ nitrogen” and “N-bound hydrogen” for both the neutral amines and the ammonium ions, the highly charged ions that exercise a stronger influence on the overall rms will “dominate” the fitting of these values. Indeed, the final values of the radii perfectly describe the solvation of ammonium ions, while introducing quite important errors for neutral amines. Of course, imposing some values that satisfy the solvation of the amines would have produced much more important errors (probably of the order of tens of kcal/mol) for the ions’ transfer energies. With Biosym parameters, the calculated transfer energy is overestimated ($+1.4$ kcal/mol) for methylamine and ($+1.2$ kcal/mol) for ammonia, is well predicted (-0.12 kcal/mol) for the dimethylamine and is underestimated (-4.2 kcal/mol) for trimethylamine and other tertiary amines. Despite of belonging to the same potential type “n3,” the charge of the N atom in primary, secondary and tertiary amines is -0.5 , -0.58 , and -0.66 |e|, respectively. On the other hand, the radius of the N-bound hydrogen atoms “hn” has been fixed to a quite large value of 1.44 Å, probably in order to accommodate the ammonium ions, where it carries a much larger charge of 0.36 vs 0.14 |e| in the neutral NH groups. The replacement of such hydrogens with methyl carbons of smaller radius (1.26 Å) combined with the increase of their atomic charge produce the spurious results of the calculation. The conclusion is that in this particular situation, the definition of the potential types can be shown to be insufficiently detailed to describe the local solvation features of N-containing molecules. There is no improvement when switching to the MOPAC charge set. On the contrary, in the second-class model, the amine molecules are excellently predicted (errors of $+0.2$ kcal/mol for methylamine and triethylamine and of -0.05 kcal/mol for dimethylamine) probably due to the redefinition of chemical

TABLE VI. Predicted vs experimental transfer energies calculated with the first class of models, using Biosym charge parametrization. The molecules displayed in bold are part of the learning set used to optimize the parameters of the model. Energies are given in kcal/mol.

MOLECULE	Experimental transfer energy	Calculated transfer energy	Absolute error
ethane	1.80	1.27	-0.53
propane	2.00	1.48	-0.52
cyclopropane	0.80	1.04	+0.24
butane	2.10	1.69	-0.41
isobutane	2.30	1.70	-0.60
neopentane	2.50	1.79	-0.71
cyclopentane	1.20	1.01	-0.19
hexane	2.50	2.42	-0.08
heptane	2.60	2.69	+0.09
cyclohexane	1.20	1.66	+0.46
methylcyclohexane	1.70	1.89	+0.19
2,4-dimethylpentane	2.90	2.29	-0.61
1,2-dimethylcyclohexane	1.60	1.53	-0.07
ethene	1.30	-0.29	-1.59
methylpropene	1.20	-0.83	-2.03
propene	1.30	-0.43	-1.73
2-pentene	1.30	-0.35	-1.65
cyclopentene	0.60	-1.03	-1.63
butadiene	0.60	-0.98	-1.58
benzene	-0.90	-2.19	-1.29
toluene	-0.90	-1.74	-0.84
<i>o</i> -xylene	-0.90	-1.23	-0.33
<i>m</i> -xylene	-0.80	-1.29	-0.49
<i>p</i> -xylene	-0.80	-1.23	-0.43
naphtalene	-2.40	-3.44	-1.04
anthracene	-4.20	-4.86	-0.66
ethyne	0.00	-2.79	-2.79
propyne	-0.30	-1.87	-1.57
1-butyne	-0.20	-1.45	-1.25
1-pentyne	0.00	-1.10	-1.10
1-hexyne	0.30	-0.86	-1.16
butenyne	0.00	-2.86	-2.86
ammonia	-4.30	-3.02	+1.28
methyl amine	-4.60	-3.17	+1.43
ethylamine	-4.50	-2.41	+2.09
propylamine	-4.40	-2.28	+2.12
butylamine	-4.30	-2.05	+2.25
aniline	-4.90	-5.52	-0.62
dimethylamine	-4.30	-4.18	+0.12
trimethylamine	-3.20	-7.43	-4.23
pyrrolidine	-5.50	-4.09	+1.41
piperazine	-7.40	-8.54	-1.14
N-methylpiperazine	-7.80	-11.69	-3.89
N,N'-dimethylpiperazine	-7.60	-14.45	-6.85
pyridine	-4.70	-3.40	+1.30
4-methylpyridine	-4.90	-2.90	+2.00
2,6-dimethylpyridine	-4.60	-2.32	+2.28
2-methylpyrazine	-5.50	-4.03	+1.47
acetonitrile	-3.90	-1.10	+2.80
propionitrile	-3.90	-0.66	+3.24
butyronitrile	-3.70	-0.35	+3.35
water	-6.30	-9.41	-3.11
methanol	-5.10	-3.86	+1.24
ethanol	-5.00	-4.10	+0.90
dimethylether	-1.90	-3.33	-1.43
acetaldehyde	-3.50	-5.76	-2.26
acetic acid	-6.70	-7.11	-0.41
methyl formiate	-2.80	-5.48	-2.68
ethanediol	-7.70	-9.07	-1.37
2-methoxyethanol	-6.80	-7.29	-0.49
1-propanol	-4.80	-3.91	+0.89
2-propanol	-4.80	-3.49	+1.31

TABLE VI. (Continued.)

MOLECULE	Experimental transfer energy	Calculated transfer energy	Absolute error
allyl alcohol	-5.00	-5.49	-0.49
propionaldehyde	-3.50	-5.12	-1.62
acetone	-3.90	-4.49	-0.59
propionic acid	-6.50	-6.61	-0.11
metyl acetate	-3.30	-4.73	-1.43
t-butanol	-4.50	-3.10	+1.40
diethylether	-1.60	-2.75	-1.15
1-methoxy-2-propanol	-2.00	-2.63	-0.63
butyraldehyde	-3.20	-4.88	-1.68
butyric acid	-6.40	-5.90	+0.50
butanone	-3.60	-3.80	-0.20
ethyl acetate	-3.10	-3.90	-0.80
methyl propionate	-2.90	-3.70	-0.80
methyl- <i>t</i> -butyl ether	-2.20	-2.41	-0.21
pent-2-one	-3.50	-3.46	+0.04
pent-3-one	-3.40	-3.32	+0.08
methylbutanoate	-2.80	-3.60	-0.80
hept-4-one	-2.90	-2.69	+0.21
non-5-one	-2.70	-2.16	+0.54
tetrahydrofurane	-3.50	-3.01	+0.49
azetidine	-5.60	-11.10	-5.50
1,4-dioxane	-5.10	-6.22	-1.12
phenol	-6.60	-6.65	-0.05
benzaldehyde	-4.00	-7.55	-3.55
acetophenone	-4.60	-5.92	-1.32
<i>m</i> -hydroxybenzaldehyde	-9.50	-11.96	-2.46
<i>p</i> -hydroxybenzaldehyde	-10.50	-12.05	-1.55
hydrogen sulphide	-0.70	-2.79	-2.09
methyl fluoride	-0.20	-3.43	-3.23
methyl trifluoride	0.80	-7.10	-7.90
tetrafluoromethane	3.10	-4.67	-7.77
1,1-difluoroethane	-0.10	-4.75	-4.65
methylindole	-5.50	-5.40	+0.10
p-cresole	-6.10	-6.11	-0.01
acetate ion	-78.50	-78.69	-0.19
propionate ion	-77.70	-77.71	-0.01
tetramethylammonium	-59.00	-59.10	-0.10
methylammonium ion	-71.00	-70.98	+0.02
dimethylammonium ion	-63.00	-64.52	-1.52
ammonium ion	-79.00	-78.21	+0.79

classes "cationic hydrogen" and "polar neutral hydrogen" but also due to the different treatment of hydrophobicity. Nevertheless, the systematic errors of as much as 3 kcal/mol in the prediction of nitriles suggest that the introduction of an independent "triple bound N" type would be required.

The transfer energies of aliphatic hydrocarbons are underestimated by approximately 1 kcal/mol in the second class model. A preliminary fit, ran only on the subset of the 32 hydrocarbons out of the 173 molecules led to a much better rms value of 0.3 kcal/mol. However, when re-refining the resulting set of parameters in presence of all the species, the radii of nonpolar aliphatic C and H decrease, while the κ_{σ} factor increases ten times, from -2.49 to -23.08 kcal/mol/ $|e|^2$. The charge-dependence coefficient λ is stable and changes from -0.23 to -0.24 Å/ $|e|$. The total surface coefficient κ_{total} slightly decreases from 0.013 to 0.010 kcal/Å². The aliphatic atoms in polar molecules must be shrunk in size and the contribution from the superficial

TABLE VII. rms errors (kcal/mol) and CPU time per conformer as a function of the setup parameters for ten successive MD geometries of barnase. N_{dens} , the density of the surface dots is given in number of points per 3 Å radius sphere. D_{cut} (Å²), the halving square distance is defined in Eq. (5).

Dot density N_{dens}	Halving square distance D_{cut}	rms and (r^2)	CPU time (seconds)
600	999	0.00 (1.00)	70.0
600	400	0.46 (0.96)	60.1
600	100	0.34 (0.98)	44.5
600	50	1.21 (0.70)	38.7
400	999	0.68 (0.91)	51.3
400	400	0.61 (0.92)	44.8
400	100	0.94 (0.82)	34.6
400	50	1.58 (0.50)	30.8
200	999	0.82 (0.87)	32.6
200	400	1.22 (0.70)	29.5
200	100	1.61 (0.48)	24.7
200	50	1.56 (0.51)	22.9

charges is enhanced in order to accommodate the polar molecules, but this leads to a negative feedback on the solvation energies of aliphatic hydrocarbons. This is less the case for the aromatic hydrocarbons, where the properties of the special atom type "aromatic carbon" is not much perturbed by the presence of other molecules. The quality of the model could be improved if distinction was made between the carbons in aliphatic chains and the ones bound to heteroatoms. A higher-order scaling as a function of the surface polarization charges κ_{σ} . $(\sigma S)^n$ could also improve the quality of the model.

5. Conclusions regarding the parametrization of the models

In conclusion, we can show that a central requirement for a good understanding of the solvent effects is a realistic charge parametrization of the solute. Charge distributions obtained from semiempirical vacuum calculations appear to perform on the overall slightly better than the ones taken from force fields, and much better for "exotic" molecules

TABLE VIII. rms errors (kcal/mol) and CPU time per conformer as a function of the setup parameters for the 8 NMR structures of crambin. N_{dens} , the density of the surface dots is given in number of points per 3 Å radius sphere. D_{cut} (Å²), the halving square distance is defined in Eq. (5).

Dot density N_{dens}	Halving square distance D_{cut}	rms and (r^2)	CPU (seconds)
600	999	0.00 (1.00)	7.4
600	400	0.19 (1.00)	6.6
600	100	0.47 (0.99)	5.2
600	50	0.47 (0.99)	4.6
400	999	0.88 (0.98)	5.3
400	400	0.96 (0.97)	4.8
400	100	0.90 (0.98)	3.9
400	50	0.84 (0.98)	3.5
200	999	1.10 (0.96)	3.2
200	400	0.86 (0.98)	3.0
200	100	1.44 (0.94)	2.6
200	50	2.32 (0.84)	2.4

TABLE IX. rms errors of the correlation between exact and truncated Coulomb energy term for the same series of the crambin molecule, as a function of the cutoff distance cutoff .

Cutoff (Å)	rms ^a (kcal/mol)	rms ^b (kcal/mol)
5.0	164.9	19.6
10.0	112.0	12.2
15.0	35.3	6.9
20.0	6.9	3.8
25.0	0.4	0.8
30.0	0.0	0.4

^aNeglecting atom pairs that are further apart than the cutoff distance without any correction.

^bNeglecting atom pairs that are further apart than the cutoff distance but multiplying below the cutoff distance the Coulomb term by a termination function $(1 - d/\text{cutoff})^2$ (Ref. 55) (d is the distance between two charges).

for which the force field parametrization seems to be less accurate. The example of the alkyl halides clearly suggests why the optimized radii should be considered as empirical force field parameters. They do not actually define the position of a physical molecule-solvent interface, but merely tell us where the interface must be placed in order to allow the simplified model to retrieve the correct value of energy which results from the extremely complex solute-solvent interaction. Another conclusion is the strong interdependency between the parameters of the solvation force field, making the definitions of chemical types a central issue of any parametrization attempt. The analysis of the results suggests that our model might be improved by the introduction of supplementary parameters. Nevertheless, the performance of the second-class approach is very satisfactory, considering both the excellent overall rms and the very good predictions of solvation energies of some exotic species like the protonated acetamide ion (-66.0 kcal/mol, error=0.0), Me_2OH^+ (-83.0 kcal/mol, error=-0.11 kcal/mol).

E. Setup parameters and computer effort

The analysis of Table VII (barnase) and Table VIII (crambin) shows that a good estimation of conformational energy differences (say an rms of the order of $kT \sim 0.6$ kcal/mol) can be achieved with considerably less computer effort (40%–60% of the computer time required for the precise runs) if the setup parameters are conveniently chosen, although, the noise introduced by a less precise parametrization will increase with the size of the part of the molecule that undergoes the conformational changes. The correlation coefficients measured from the regression line with an imposed slope of 1.0 and a free intercept will moreover depend on the spread of the exact energy values for the current set of geometries. This spread being more important for crambin than for barnase (the conformations used for the analysis are much more similar for barnase than for crambin), the same rms values will reflect a better correlation for crambin than for barnase.

Table IX displays the rms on the Coulomb energy term of crambin introduced by applying a spherical cutoff to truncate the interactions, combined or not with the use of a ter-

mination function (columns *a* and *b* respectively). They show that even in presence of a termination function, the error arising from the introduction of a cutoff distance of 10 to 20 Å is much larger than the lowest accurate solvation model (Tables VII and VIII). In other words, considerable speedup might be obtained by decreasing the precision while still remaining within errors that are commonly judged as acceptable in a molecular dynamics simulation, although this has to be investigated in more details.

The computer times listed in Tables VII and VIII, measured on one processor of an IRIS-Power Challenge machine, probably are of the same order of magnitude or even smaller than the values reported in the literature⁶¹ for the most recent finite-difference multigrid algorithms for solving the Poisson–Boltzmann equation on a Convex C3 machine (CPU times of ~10 to 100 s for the crambin molecule). It is however difficult to compare the computing efficiency of two algorithms running on different systems.

IV. CONCLUSIONS

The empirical force field approaches for the treatment of the solvent effects, developed in this study show a good predictive power with respect to experimentally measured vacuum-to-water transfer energies of a large set of compounds, which includes ionic, polar and nonpolar species. The average overall error of the computed values is of the same order as the imprecision of the experimental data. The coefficients of the nonlinear fit freely adjust to values not far from the ones expected on theoretical grounds, suggesting that this approach is physically consistent. The polarization energy values taken as such, without any weighting correction, provide a reasonable approximation to the electrostatic contribution to the transfer energies. The drastic simplifications of the BEM proposed here are shown to preserve the physical quality of these calculations, with a substantial decrease in computational effort. In particular, our results suggest that the most time-consuming step of the “classical” BEM, which is the BEM matrix inversion, can be successfully avoided. A significant improvement of the predictive power has been achieved by optimization of the atomic radii. However, while it is hazardous to claim that the optimized radii have a physical significance, it is nevertheless clear that the local geometry of the hypothetical dielectric interface considered in the BEM algorithm can be consistently described in terms of the atomic types defined here. The less stable part of the model is the surface area term which adjusts itself in the optimization, so as to compensate for the errors in the electrostatic terms. The continuum solvent model proposed in this study is particularly well adapted for implementations in Monte Carlo algorithms and probably also in the molecular dynamics simulations, provided the calculation of the forces can be sufficiently optimized. It is shown that calculations on macromolecules can be made more efficient by an optimal choice of the setup parameters of the model. The results suggest that further simplifications might significantly reduce the computer times to the typical

values needed for the energy and gradient calculations in a molecular dynamics simulation ran with explicit solvent molecules.

ACKNOWLEDGMENTS

We acknowledge support from the French Centre National de la Recherche Scientifique (CNRS), and the Belgian National Science Funds (FNRS), for support in the framework of the LEA (Laboratoire Européen Associé). Part of this research was also financed by the Belgian program of Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister’s Office for Science, Technology and Culture. D. Horvath thanks Professor A. Tartar, for constant encouragement and acknowledges financial support from Conseil Régional du Nord, France. D. van Belle and S. J. Wodak acknowledge support from European Community Biotechnology BRIDGE Contract No. CT91-0270.

¹ N. Muller, *Acc. Chem. Res.* **23**, 23 (1990).

² T. E. Creighton, *Curr. Opin. Struct. Biol.* **1**, 5 (1991).

³ L. Chiche, L. M. Gregoret, F. E. Cohen, and P. A. Kollman, *Proc. Natl. Acad. Sci. USA* **87**, 3240 (1990).

⁴ J. Åquist, C. Medina, and J. E. Samuelsson, *Protein Eng.* **7**, 385 (1994).

⁵ R. Tan, T. Truong, J. McCammon, and J. Sussman, *Biochemistry* **32**, 401 (1993).

⁶ V. Misra, K. Sharp, R. A. Friedman, and B. Honig, *J. Mol. Biol.* **238**, 245 (1994).

⁷ A. Warshel, *Computer Modelling of Chemical Reactions in Enzymes and Solutions* (Wiley, New York, 1991).

⁸ D. Bashford, *Curr. Opt. Struct. Biol.* **1**, 175 (1991).

⁹ D. L. Beveridge and F. M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.* **18**, 431 (1989).

¹⁰ J. A. McCammon, *Annu. Rev. Biophys. Biophys. Chem.* **1**, 196 (1991).

¹¹ K. Sharp, *Annu. Rev. Biophys. Biophys. Chem.* **1**, 171 (1991).

¹² J. Moulton, *Annu. Rev. Biophys. Biophys. Chem.* **2**, 223 (1992).

¹³ C. J. Cramer and D. G. Truhlar, *Rev. Comp. Chem.* (in press).

¹⁴ C. Tanford and J. G. Kirkwood, *J. Am. Chem. Soc.* **79**, 5333 (1957).

¹⁵ D. Bashford, M. Karplus, and G. W. Canters, *J. Mol. Biol.* **203**, 507 (1988).

¹⁶ M. K. Gilson and B. H. Honig, *Proteins* **3**, 32 (1988).

¹⁷ J. Antosiewicz, J. A. McCammon, and M. K. Gilson, *J. Mol. Biol.* **238**, 415 (1994).

¹⁸ A. Ben-Naim and Y. Marcus, *J. Chem. Phys.* **81**, 2016 (1984).

¹⁹ D. van Belle, I. Couplet, M. Prévost, and S. J. Wodak, *J. Mol. Biol.* **198**, 721 (1987).

²⁰ D. van Belle, M. Froeyen, G. Lippens, and S. J. Wodak, *Mol. Phys.* **77**, 239 (1992).

²¹ R. M. Brunne, E. Liepinsh, G. Otting, K. Wüthrich, and W. F. van Gunsteren, *J. Mol. Biol.* **231**, 1040 (1993).

²² M. Prévost, S. J. Wodak, B. Tidor, and M. Karplus, *Proc. Natl. Acad. Sci. USA* **88**, 10880 (1991).

²³ J. Gao, K. Kuczera, B. Tidor, and M. Karplus, *Science* **244**, 1069 (1989).

²⁴ B. Tidor and M. Karplus, *Biochemistry* **30**, 3217 (1991).

²⁵ L. R. Pratt and D. Chandler, *J. Chem. Phys.* **67**, 3683 (1977).

²⁶ B. Lee, *Biopolymers* **31**, 993 (1991).

²⁷ I. Klapper, R. Hangstrom, R. Fine, K. Sharp, and B. Honig, *Proteins* **1**, 47 (1986).

²⁸ M. K. Gilson and B. Honig, *Proteins* **3**, 32 (1988).

²⁹ M. K. Gilson, M. E. Davis, B. A. Luty, and J. A. McCammon, *J. Phys. Chem.* **97**, 3591 (1993).

³⁰ K. Sharp and B. Honig, *J. Phys. Chem.* **94**, 7684 (1990).

³¹ A. A. Rashin, *J. Phys. Chem.* **94**, 1725 (1990).

³² A. A. Rashin and K. Nambodiri, *J. Phys. Chem.* **91**, 6003 (1987).

³³ A. H. Juffer, E. F. F. Botta, B. A. M. van Keulen, A. van der Ploeg, and H. J. C. Berendsen, *J. Comp. Phys.* **97**, 8 (1991).

- ³⁴W. C. Still, A. Tempczyk, R. C. Hawley, and T. F. Hendrickson, *J. Am. Chem. Soc.* **112**, 6127 (1990).
- ³⁵K. Sharp, A. Jean-Charles, and B. Honig, *J. Phys. Chem.* **96**, 3822 (1992).
- ³⁶C. J. Cramer and D. G. Truhlar, *J. Comp. Aid. Mol. Des.* **6**, 629 (1992).
- ³⁷D. Sitkoff, K. A. Sharp, and B. Honig, *J. Phys. Chem.* **98**, 1978 (1994).
- ³⁸M. K. Gilson and B. Honig, *J. Comp. Aided. Mol. Des.* **5**, 5 (1991).
- ³⁹K. Sharp, *J. Comp. Chem.* **12**, 454 (1991).
- ⁴⁰A. Jean Charles, A. Nicholls, K. Sharp, B. Honig, A. Tempczyk, T. F. Hendrickson, and W. C. Still, *J. Am. Chem. Soc.* **113**, 1454 (1991).
- ⁴¹T. Simonson and A. T. Bruenger, *J. Phys. Chem.* **98**, 4683 (1994).
- ⁴²Discover User Guide, version 3.1., Biosym Technologies, San Diego, 1993.
- ⁴³A. T. Hagler, S. Lifson, and P. Douber, *J. Am. Chem. Soc.* **101**, 5122 and 6842 (1979).
- ⁴⁴M. S. J. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.* **107**, 3209 (1985).
- ⁴⁵B. H. Besler, K. M. Mertz, Jr., and P. A. Kollman, *J. Comp. Chem.* **11**, 413 (1990).
- ⁴⁶D. Eisenberg and A. D. McLahan, *Nature* **319**, 199 (1986).
- ⁴⁷J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975).
- ⁴⁸B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
- ⁴⁹J. Tomasi and M. Persico, *Chem. Rev.* **94**, 2027 (1994).
- ⁵⁰T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **84**, 3086 (1987).
- ⁵¹B. Lee, *Enthalpy-Entropy Compensating Processes* (to be published).
- ⁵²Y. Mauguen, R. W. Hartley, E. J. Dodson, G. G. Bricogne, C. Chothia, and A. Jack, *Nature* **297**, 162 (1982).
- ⁵³S. J. Wodak, D. van Belle, and M. Prevost, *Computer Modelling in Molecular Biology*, edited by J. M. Goodfellow (VCH, Weinheim, 1995), pp. 61–102.
- ⁵⁴A. M. J. J. Bonvin, J. A. C. Rullmann, R. M. J. N. Lamerichs, R. Boelens, and R. Kaptein, *Proteins. Struct. Funct.* **15**, 385 (1993).
- ⁵⁵C. L. Brooks, B. M. Pettitt, and M. Karplus, *J. Chem. Phys.* **83**, 5897 (1985).
- ⁵⁶M. Prévost, D. van Belle, G. Lippens, and S. J. Wodak, *Mol. Phys.* **71**, 587 (1990).
- ⁵⁷D. J. Tannor, B. Marten, R. Murphy, R. A. Friesner, D. Sitkoff, A. Nicholls, M. Ringnalda, W. A. Goddard III, and B. Honig, *J. Am. Chem. Soc.* **116**, 11875 (1994).
- ⁵⁸E. Clementi, *J. Chem. Phys.* **74**, 578 (1979).
- ⁵⁹J. D. Roberts and M. C. Caseiro, *Basic Principles of Organic Chemistry* (Benjamin, New York, 1964).
- ⁶⁰A. Radzicka and R. Wolfenden, *Biochemistry* **27**, 1664 (1988).
- ⁶¹M. J. Holst and F. Saied, *J. Comp. Chem.* **16**, 347 (1994).

Development and parametrization of continuum solvent models.

II. A unified approach to the solvation problem

D. Horvath

Institut Pasteur de Lille, Service de Chimie des Biomolécules URA CNRS 1309, 1 Rue Prof. Calmette, 59000 Lille, France; and Department of Organic Chemistry, Babes-Bolyai University, 11 Arany Janos Street, 3400 Cluj-Napoca, Romania; and Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, P2, Avenue P. Héger, B-1050 Bruxelles, Belgium

G. Lippens

Institut Pasteur de Lille, Service de Chimie des Biomolécules URA CNRS 1309, 1 Rue Prof. Calmette, 59000 Lille, France and Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, P2, Avenue P. Héger, B-1050 Bruxelles, Belgium

D. van Belle

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, P2, Avenue P. Héger, B-1050 Bruxelles, Belgium

(Received 30 January 1996; accepted 24 May 1996)

In order to avoid the computational expense required to obtain precise numerical solutions of the Poisson–Boltzmann equation, different hypotheses have been introduced which led to less rigorous, but fast and simple solvation models. However, few systematic studies of the predictive features of such models have been reported. Comparisons between different continuum models are made difficult by the large variety of simplifying hypotheses or thermodynamic reference states used by their authors, so that the relationships between the proposed solvation terms is not straightforward. In the present work we consider various continuum models whose common feature is the description of the solvation process in terms of *displacement* of the high dielectric solvent by the low dielectric molecular bulk. We adapt these different models to work within the frame of the nonlinear fit algorithm developed in our previous study. This leads to a benchmark allowing to run rigorous comparative tests between all the different schemes. The unification of these allows us to derive new solvent models that combine the advantages of the original ones. We address also the problem of the optimal parametrization of the solvation terms and propose a new strategy of assignment of the atomic radii. It is shown that the exact solvation energies, which are in theory linearly related to the solvent displacement terms, cannot be expressed as a linear combination of the displacement effects evaluated with the usual simplifications. Nevertheless, *nonlinear* empirical models based on these simplified displacement terms are found to yield high quality predictions of vacuum-to-water transfer energies. © 1996 American Institute of Physics. [S0021-9606(96)50733-6]

I. INTRODUCTION

One of the most serious problems of molecular mechanics calculations is the description of the solvation process.^{1–3} The explicit inclusion of water molecules in simulations is indeed a common practice, although this approach raises a series of difficulties and limitations: the problem of accurate force field parametrization,^{4–6} the artifacts related to the boundary conditions and truncating the long-range interactions,⁷ and the computational burden related to the size of the resulting system.

The solvation process cannot be understood as the interaction of the solute with well localized water molecules, but merely as the averaged influence exercised by a continually changing solvation shell on the solute molecule.^{8–10} This continual change in the orientation of solvent molecules led to the idea of ignoring the details of the solvent structure and replacing it with a description of the averaged effect of the water molecules on the solute.^{11,13}

Continuum solvent models^{14–16} rely on the common assumption that on time scales that are long enough to allow an

effective averaging of the solvation shell, the *electrostatics* of the solvation process can be reduced to an interaction between the considered charge distribution of the solute, which can be either obtained from quantum electron density calculations or approximated by atomic point charge distributions as used in empirical force fields, with a surrounding dielectric continuum representing the solvent. *Nonelectrostatic contributions* such as cavity creation and solvent–solute van der Waals interactions or more subtle entropic effects such as the hydrophobic effect are not fully understood^{17–20} and are usually assumed to be proportional to the solvent-accessible or molecular surface of the solute.^{18–21} The sum of these two contributions defines the influence of the solvent^{19,21} in terms of a *potential of mean force* resulting from the averaged actions of solvent molecules in all possible orientations on the solute molecule. The *electrostatics of solvation* was intensely studied by means of the equations of macroscopic electrostatics.^{11–13,22–25} Many of these approaches are mathematically rigorous but much too complex to be included into molecular dynamics and Monte Carlo calculations.

Other approaches, aimed to be parts of the force field are necessarily empirical and the different approximations used by the authors make it sometimes difficult to trace all these models back to their common origin, the Poisson equation.

In a previous paper, we have studied a simplified method for modeling the electrostatic effects based on the evaluation of superficial polarization charge densities by the boundary element method (BEM^{24,27}). An empirical approach derived by Still,²¹ despite its simplicity, shows surprisingly good results. Another approach²⁸ has been to correct the atomic fractional charges by the local polarization charges predicted by the solvent model. Gilson and Honig²⁹ proposed a force-field term based on the effect of the displacement of the high dielectric solvent by the low dielectric surrounding atoms. An even more empirical approach is the treatment of solvation (Eisenberg and Scheraga) in terms of accessible surfaces of different groups without considering any explicit polarization term.^{18,20} Other reported approaches—Stouten's exponential solvation term^{30,31} which is highly similar to Gilson and Honig's function or Scheraga's "solvation-volume-energy-density" approach³² will, however, not be detailed in the present work.

This study analyzes several *empirical* solvent models that are not based on rigorous electrostatic equations, but on different simplifying hypotheses leading to different formalisms which, at first sight, appear to have few points in common. We introduce here a generalized treatment based on the evaluation of the energetics of the *displacement* of the external dielectric by the surrounding atoms of the molecule, in the frame of which several of the existing empirical solvent models naturally emerge as particular cases. This allows us to clearly outline the electrostatic approximations that are committed for each one of these terms, to point out which of the intervening parameters might be improved by fitting, and eventually to derive new, improved expressions of solvation terms. There is little information concerning the predictive quality of these models and the possible optimizations of the parameters they imply.¹⁵ Therefore, we have conceptually unified and adapted several of the schemes for calculating the *electrostatic terms of solvation* (taken as such from the original works^{18,21,29} and generalized by us) to be compatible with a nonlinear fitting procedure developed previously.²⁷ The nonelectrostatic part of the model is the same (cavity formation, dispersion, and hydrophobic effects) as the one used in our previous paper, even if certain authors did not use these terms in their original models. The nonlinear fitting procedure²⁷ used to relate the solvent-corrected conformational energies of a set of small test molecules to their experimental vacuum-to-water transfer free energies provides a benchmark allowing a comparative test of the predictive features of all the proposed solvation terms under identical conditions of parametrization. Finally, the models showing the best predictive power were subjected to full parameter set optimization. Two different parametrization strategies are used to assign the atomic radii: the classical parametrization in terms of the chemical contexts of the atoms and a new approach using charge-dependent atomic radii.

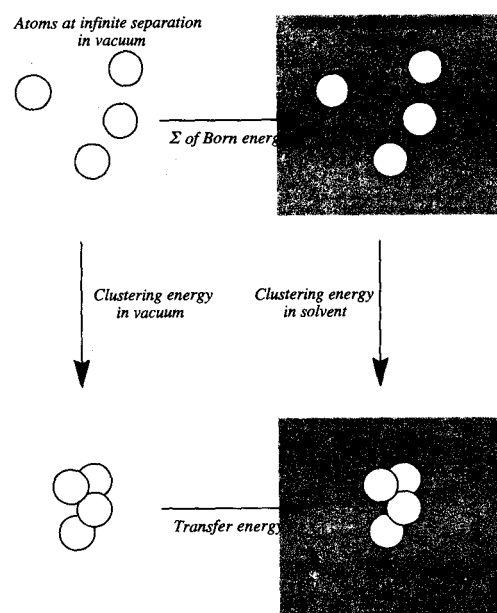


FIG. 1. Thermodynamic cycle describing the vacuum-to-solvent transfer energy in terms of "clustering" infinitely separated atoms. The energy of the reference state equals the sum of the energies stored in the radial fields around each of the isolated atoms. After soaking the separated atoms into the solvent, this energy equals the energy of the reference state divided by the solvent dielectric constant.

II. MATERIALS AND METHODS

A. The analysis of the different electrostatic models

The electrostatic part of the solvation process can be described by a large variety of contributions. This is the result of the different approximations and mathematical approaches applied to solve the Poisson equation^{22,24,25} and of the different thermodynamic reference states chosen to define the thermodynamic cycle of the solvation process. In our previous paper,²⁷ we have expressed the electrostatic energy of the system as a sum of Coulomb and polarization terms arising from the interaction between superficial polarization charges and fixed atomic charges. This corresponds to a thermodynamic cycle based on the process of *creation of the dielectric interfaces*.

In the present work, we will focus on another approach to the thermodynamics of solvation, used in several solvation models (Gilson and Still), in which the reference state contains infinitely separated atoms and the change in energy when building a molecule in a homogeneous external medium is evaluated (Fig. 1).

In principle, this work can be evaluated by integrating the electrostatic energy density stored in the electric field,³³ which is the product of the charge and dielectric distributions

$$\frac{dU}{dV} = \frac{1}{2} \epsilon E^2 = \frac{1}{2} \mathbf{E} \cdot \mathbf{D}. \quad (1)$$

U is the electrostatic free energy (including an entropic contribution from solvent reorientation), E and D are the electric

field and displacement vectors, and ϵ is the dielectric constant which is a function of the spatial coordinates.

The *energy of clustering* is defined as the electrostatic energy needed to bring the atomic spheres from infinity to their final positions in the external medium of given dielectric constant ϵ_{ext} (Fig. 1). The atoms are treated as dielectric spheres of ϵ_{int} having their charge located at the center. In the reference state, where the atoms are infinitely separated, the field outside the atomic dielectric spheres is radial and equals in the solvent $1/\epsilon_{\text{solv}}$ of its value in vacuum. The self-energy "contained" in this electric field is $Q^2/8\pi\epsilon_{\text{ext}}R$, as given by the integral of Eq. (1) over all space outside the sphere of radius R . The difference of these terms for $\epsilon_{\text{ext}}=1$ and $\epsilon_{\text{ext}}=\epsilon_{\text{solv}}$ leads to the well-known Born solvation term for a spherical ion.³⁴

Now, we want to calculate the electrostatic interactions between the approaching atoms as the molecule is formed. For systems with spherical symmetry, the dielectric constant ϵ only influences the magnitude of the vector E with respect to its value in vacuum

$$\mathbf{E}_\epsilon = \frac{1}{\epsilon} \mathbf{E}_{\text{vac}}. \quad (2)$$

The volume integral in Eq. (1) can be decomposed into contributions from the distinct dielectric regions, where the corresponding dielectric constants appear as simple proportionality factors. This no longer holds for arbitrary geometries of the dielectric interface: The field lines will be distorted and a change in dielectric constant does not imply a proportional change in energy. The initial radial fields will interfere to generate the final, irregular pattern of field lines, for which the integral of Eq. (1) cannot be solved analytically.

The first type of models we discuss here are called "displacement" models. These are based on the following hypotheses:

(a) The boundary distortion effects are ignored and Eq. (2) is claimed to stand in all situations.

(b) The charge-charge interaction and the dielectric damping of the electric field can be treated separately. It is supposed that the total electrostatic energy can be written as a sum of these two terms.

(c) The dielectric damping effect of the electrostatic energy can be split into atomic contributions assuming that each atom gives rise to an *undisturbed spherical field*: All other charges are set to zero while evaluating the dielectric contribution of an atom (the rest of the molecule plays only the role of a dielectric). We obtain

$$\begin{aligned} U_i^{\text{diel}} &= \frac{Q_i^2}{32\pi^2} \left\{ \frac{1}{\epsilon_{\text{int}}} \int_{V_{\text{int}}} \frac{dV}{r_i^4} + \frac{1}{\epsilon_{\text{ext}}} \int_{V_{\text{ext}}} \frac{dV}{r_i^4} \right\} \\ &= \frac{Q_i^2}{32\pi^2} \left\{ \frac{1}{\epsilon_{\text{int}}} I_i^{\text{int}} + \frac{1}{\epsilon_{\text{ext}}} I_i^{\text{ext}} \right\}. \end{aligned} \quad (3)$$

I_i^{int} and I_i^{ext} are the volume integrals evaluated over the dielectric regions *inside* and *outside* the molecule, excluding the atomic sphere of radius R_i itself. The integral of dV/r^4

outside a sphere of radius R_i yields $4\pi/R_i$, which is equal to the sum of I_i^{int} and I_i^{ext} . The precise evaluation of these volume integrals has been performed in polar coordinates centered on the nucleus of the current atom, with a constant step of $d\theta \approx 0.1$, while the magnitude of $d\phi$ is chosen such as to ensure a constant solid angle $d\phi = 0.01/(\sin\theta d\theta)$. The radius r is incremented from R_i to an R_{max} for which the integration sphere encompasses the whole molecule. Furthermore, choosing $dr = 0.01r^2$ insures a constant contribution of $dV/r^4 = r^2 dr \sin\theta d\theta d\phi/r^4 = 10^{-4}$ from each explored point around the atomic sphere, independently of its position. All the points beyond R_{max} are in the external medium and contribute to I_i^{ext} by $4\pi/R_{\text{max}}$. The accumulated numerical errors can be ruled out by rescaling the two calculated integrals, applying the constraint that their sum must yield exactly $4\pi/R_i$. A faster method involving the analytical estimation of the atomic surface with an increasing probe size has been reported.²¹

(d) A simple Coulomb term is considered to account for the charge-charge interaction, as if the medium was homogeneous. An effective damping factor $\epsilon_{\text{ext}}^{\text{eff}}$ is used to mimic the real behavior of the charge-charge interactions in the inhomogeneous system; the subscript "ext" emphasizes that this value depends on both ϵ_{int} and ϵ_{ext} . Different models of distance-dependent "dielectric constants" have been proposed.³⁵⁻³⁷ Those are based on the fact that the separation between two charges will modulate the relative contributions of the external and internal media to the interaction damping. However, the factors determining the effective dielectric constant cannot be reduced to the simple pairwise relative atomic distances, but are related to the overall shape of the solute.

According to the previous hypotheses, the electrostatic energy of a system of atoms embedded in an external dielectric medium of dielectric constant ϵ_{ext} is

$$\frac{1}{2} \int_V \mathbf{E} \cdot \mathbf{D} dV \approx U_{\text{Coul}} + U_{\text{diel}} = \frac{1}{\epsilon_{\text{ext}}^{\text{eff}}} \sum_{i < j} \frac{Q_i Q_j}{4\pi r_{ij}} + \sum_i U_i^{\text{diel}}. \quad (4)$$

When atoms are brought from infinity to build a molecule, the change in energy due to the *displacement of the external medium* by the dielectric molecular bulk is

$$\Delta U_{\text{disp}} = \left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}} \right) \sum_i \frac{Q_i^2}{32\pi^2} I_i^{\text{int}}. \quad (5)$$

The sum of U_{Coul} and ΔU_{disp} represents the *clustering energy* within the frame of the "displacement" models. According to the thermodynamic cycle (Fig. 1) and switching the external medium ext from vacuum to water, we obtain for the electrostatic (polarization) contribution to the vacuum-to-water transfer energy

$$\Delta F_{\text{pol}}^{\text{transfer}} = \left(\frac{1}{\epsilon_{\text{solv}}} - \frac{1}{\epsilon_{\text{vac}}} \right) \sum_i \frac{Q_i^2}{32\pi^2} \frac{4\pi}{\alpha_i} + \left(\frac{1}{\epsilon_{\text{sol}}^{\text{eff}}} - \frac{1}{\epsilon_{\text{vac}}^{\text{eff}}} \right) \sum_{i < j} \frac{Q_i Q_j}{4\pi r_{ij}}. \quad (6)$$

In Eq. (6) we have introduced the "generalized Born radii" α_i , which are related to the effective atomic radii R_i by the relation

$$\alpha_i = R_i \left(1 + \frac{f_i^{\text{int}}}{f_i^{\text{ext}}} \right). \quad (7)$$

We can also define the "generalized Born energy" for a molecule embedded in an external medium ext

$$F_{EQ\text{-Born}}^{\text{ext}} = \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}} \right) \sum_i \frac{Q_i^2}{8\pi\alpha_i} = \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}} \right) (U_{\text{Born}}^0 - \Delta U_{\text{disp}}^0), \quad (8)$$

where superscript 0 will be used throughout the paper to denote the energetic terms without their multiplicative factors in $1/\epsilon$.

While it would have been possible to follow our approach used in Ref. 27, where atomic radii that differed by a given offset (the "probe" radius) were considered to describe the dielectric boundaries in vacuum and in solvent, a single common interface will be defined in the present work. This is due to the fact that the terms that will be studied furtheron have been originally defined *without* accounting for a shift of the dielectric interface when passing from vacuum into solvent.

Gilson and Honig²⁹ proposed to replace the integral in Eq. (5) by a discrete sum over all neighboring atoms

$$\Delta U_{\text{disp}} = \left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}} \right) \bar{V} \sum_i \frac{Q_i^2}{32\pi^2} \sum_j \frac{1}{r_{ij}^4}. \quad (9)$$

This term is always positive in water, stating that charged atoms tend to interact favorably with the solvent. The authors have used Eq. (9), assuming that all atoms occupy the same average volume \bar{V} as found in the packed structure of a protein.

We can also introduce a variant of Eq. (9) where the average atomic volume \bar{V} is replaced by the precise individual volumes V_j .

$$\Delta U_{\text{disp}} = \left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}} \right) \sum_i \frac{Q_i^2}{32\pi^2} \sum_j \frac{V_j}{r_{ij}^4}. \quad (10)$$

We have used an empirical approach to rapidly estimate the atomic volumes required in Eq. (10). Considering the simple case of two overlapping atoms as depicted in Fig. 2, the shaded volume cut out by the atom j from atom i is given by

$$V_{ij} = \frac{\pi}{3} h_i^2 (3R_i - h_i), \quad (11)$$

where the parameter h_i can be easily expressed in function of the known radii R_i and R_j and of the interatomic distance

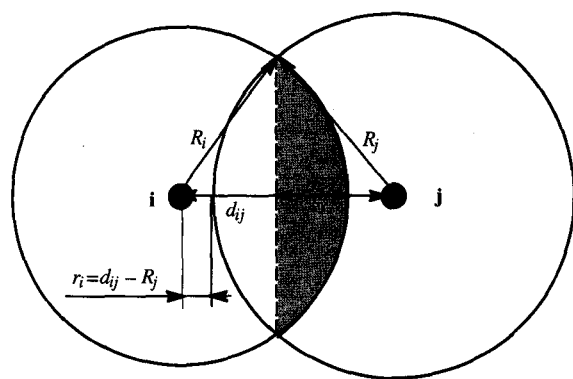


FIG. 2. Approximative evaluation of atomic volumes. The shaded part represents the volume of atom i cut out by the neighbor atom j .

d_{ij} . The value $r_i = d_{ij} - R_j$ is the radius of a sphere centered on the nucleus i , which is not cut by atom j . When atom i has more than one close neighbor, we use

$$V_i^u = V_i^0 \prod_j \left(1 - \frac{V_{ji}}{V_i^0} \right), \quad (12)$$

where V^u is the estimated "uncut" volume based on the probabilities that a given point is not cut out by any of the neighbors j . V^0 is the volume of the sphere of radius R_i . Equation (12) underestimates the uncut volume of atom i , because it relies on the hypothesis that the neighbors j are randomly distributed around i . Therefore, we use a different estimation of the lower limit of the volume of atom i , based on the "average radius of the uncut inner sphere" R_i^{avg}

$$R_i^{\text{avg}} = \frac{1}{N_n} \sum_j^{N_n} (d_{ij} - R_j), \quad (13)$$

from which

$$V_i^{\text{min}} = \frac{4\pi}{3} R_i^{\text{avg}^3}. \quad (14)$$

The maximum volume between V^u and V^{min} is taken as the final atomic volume of i .

Within the concept of the "displacement" solvent models, we have also tried out a corrective term in function of the fraction of accessible surface. Instead of evaluating the discrete sum appearing in Eq. (10), we assumed that the dielectric constant is a simple function of the solid angle around the center of a given atom, which is part of a molecule. The dielectric constant is set to ϵ_{ext} within a solid angle delimited by the accessible fraction of the atomic surface and to ϵ_{int} in all the rest of the space. We obtain thus

$$\Delta U_{\text{disp}} = \left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}} \right) \sum_i \frac{Q_i^2}{8\pi R_i} (1 - X_i), \quad (15)$$

where X_i denotes the accessible fraction of the atomic area and R_i is the Born radius of atom i .

Still's model based on the "generalized Born" approach²¹ consists in combining the Coulomb and solvent displacement terms of Eq. (6) in order to obtain a function

that reproduces the analytical formulas for the electrostatic energy in simple atomic arrangements and is expected to lead to reasonable results for real molecular geometries

$$U_{\text{Still}} = \left(\frac{1}{\epsilon_{\text{solv}}} - \frac{1}{\epsilon_{\text{vac}}} \right) \sum_{ij} \frac{Q_i Q_j}{8\pi f(\alpha_i, \alpha_j, r_{ij})}. \quad (16)$$

The empirical function f was chosen to approximate the Onsager expression³⁸ at small interatomic distances, while yielding the Coulomb+Born terms for atoms at large distances

$$f(\alpha_i, \alpha_j, r_{ij}) = \sqrt{r_{ij}^2 + \alpha_i \alpha_j e^{-r_{ij}^2 / 4\alpha_i \alpha_j}}. \quad (17)$$

At this point, we will introduce a modification in Still's approach, based on a more general definition of the generalized Born radii. The previous definition of these radii (α_i) was chosen so as to allow the solvent displacement effects accompanying the vacuum-to-water transfer of a molecule to be broken up into atomic contributions that can be formally written as a classical Born term in Eq. (8). However, in inhomogeneous media, the electric field lines no longer maintain the spherical symmetry implied when evaluating the volume integrals I^{int} and I^{ext} , but tend to "concentrate" in the high dielectric regions (Fig. 3). The evaluation of these field line distortions requires an exact solution of the Poisson equation, which cannot be solved within the present conceptual frame. The assumption adopted here is that the electrostatic energy stored in the electric field *inside* the molecule is also a function of a weighting coefficient c^{ext} , depending on the external medium. This weighting coefficient is larger for the molecule in vacuum than for the solvated molecule, since $\epsilon_{\text{solv}} > \epsilon_{\text{int}} > \epsilon_{\text{vac}}$. The introduction of such a coefficient is consistent with the Born terms describing solvation at infinite interatomic separation, where I^{int} is zero. Therefore, an atomic contribution to U^{diel} can be written as

$$U_i^{\text{diel}} = \frac{Q_i^2}{32\pi^2} \left(\frac{1}{\epsilon_{\text{ext}}} I_i^{\text{ext}} + \frac{c^{\text{ext}}}{\epsilon_{\text{int}}} I_i^{\text{int}} \right). \quad (18)$$

The generalized Born radius of this atom can be expressed from the change in U^{diel} during the transfer from vacuum to water

$$\frac{4\pi}{\alpha_i} = I_i^{\text{ext}} \frac{c^{\text{solv}} - c^{\text{vac}}}{\epsilon_{\text{solv}} - \epsilon_{\text{vac}}} \frac{\epsilon_{\text{solv}} \epsilon_{\text{vac}}}{\epsilon_{\text{int}}} I_i^{\text{int}} = I_i^{\text{ext}} + (1 - \lambda) I_i^{\text{int}}, \quad (19)$$

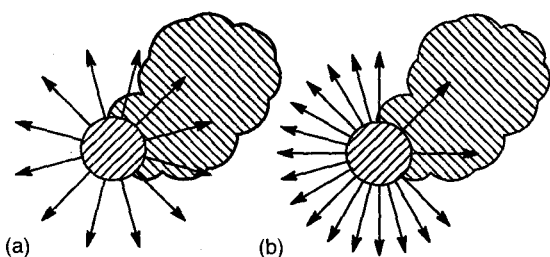


FIG. 3. The alteration of the field line density as a consequence of inhomogeneous dielectric medium. (a) The homogeneous field line distribution used for calculating the integrals I_{int} and I_{ext} ; (b) The field lines are distorted and concentrate in the high dielectric region in inhomogeneous media.

which can be rewritten as

$$\alpha_i = R_i \left(1 + \frac{\lambda I_i^{\text{int}}}{\frac{4\pi}{R_i} - \lambda I_i^{\text{int}}} \right), \quad (20)$$

where $\lambda < 1$ since $c^{\text{solv}} < c^{\text{vac}}$. This last expression no longer has the same limits as the original one, because it leads to finite radii even if the atom is completely buried ($I^{\text{ext}}=0$). Therefore, we have also investigated an alternative expression for the generalized Born radii

$$\alpha_i = R_i \left(1 + \lambda \frac{I_i^{\text{int}}}{I_i^{\text{ext}}} \right). \quad (21)$$

This definition maintains the limit values for α , leading to $\alpha_i = R_i$ for a completely isolated atom and to $\alpha_i = \infty$ for a completely buried atom.

The most time-consuming step in the Still method is the evaluation of the volume integrals I^{int} and I^{ext} . These integrals (and hence the generalized Born radii) are weakly dependent on small conformational changes, so that they do not have to be reevaluated at each step of a MD simulation.²¹ Since the fluctuations of these integrals due to small changes in molecular geometry can be ignored in Still's approach, we could as well tolerate an error of the same order of magnitude in their initial estimation. The volume integrals could be replaced by faster, but less accurate approximations. Gilson's approximation suggests to replace I^{int} by a discrete sum on interatomic distances as found in Eq. (10). Introducing a common average volume, Eq. (20) can be used to evaluate the generalized Born radii with $\lambda I_i^{\text{int}} = \lambda \bar{V} \sum r_{ij}^{-4}$, where $\lambda \bar{V}$ is a new fittable parameter. We will refer to this scheme as the "Gilson-Still" model. The main drawback of this approach lies in the use of the "average atomic volume." In reality the packing density of the atoms is different in small molecules (where the packing is merely of covalent nature) than in biomolecular aggregates (where van der Waals forces are governing the packing). Therefore, the parameters derived for small molecules might not be directly transferable to solvated macromolecules.

B. The link between the "volume" and the "surface" models

We can conclude that the main methodological difference between the solvation models reviewed here and the BEM approach presented in the previous paper²⁷ lies in the direct use of the volume integration, while the BEM method is based on surface integrals. This difference is not only mathematical, but also of a conceptual nature, illustrating two ways of building a solute molecule in the solvent. We have seen that the polarization energy, as defined in terms of interactions between polarization charge densities and fixed charges, can be directly related to the electrostatic (polarization) contribution of the vacuum-to-water transfer energy

$$\Delta F_{\text{pol}}^{\text{transfer}} = \Delta F_{\text{pol}}^{\text{vac} \rightarrow \text{solv}}, \quad (22)$$

where $\Delta F_{\text{pol}}^{\text{vac} \rightarrow \text{solv}}$ stands for the difference between these values in solvent and vacuum. This is the same $\Delta F_{\text{pol}}^{\text{transfer}}$ as

given in Eq. (6). However, Eq. (22) is much simpler and it is tempting to derive the equivalent expressions for the polarization energy F_{pol} in terms of the "volume" models that led to Eq. (6). Comparing the two equations, we obtain the following expression:

$$F_{\text{pol}}^{\text{ext}} = \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}} \right) \sum_i \frac{Q_i^2 4\pi}{32\pi^2 \alpha_i} + \left(\frac{1}{\epsilon_{\text{ext}}^{\text{eff}}} - \frac{1}{\epsilon_{\text{int}}} \right) \sum_{i < j} \frac{Q_i Q_j}{4\pi r_{ij}}, \quad (23)$$

which is a linear combination of the "generalized Born energy" and Coulomb terms

$$F_{\text{pol}}^{\text{ext}} = \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}} \right) F_{EQ\text{-Born}}^0 + \left(\frac{1}{\epsilon_{\text{ext}}^{\text{eff}}} - \frac{1}{\epsilon_{\text{int}}} \right) U_{\text{Coul}}^0. \quad (24)$$

Equation (23) can be further rewritten in function of the displacement term according to Eq. (8)

$$F_{\text{pol}}^{\text{ext}} = \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}} \right) (U_{\text{Born}}^0 - \Delta U^{\text{disp}}) + \left(\frac{1}{\epsilon_{\text{ext}}^{\text{eff}}} - \frac{1}{\epsilon_{\text{int}}} \right) U_{\text{Coul}}^0. \quad (25)$$

From its original derivation, it can be seen that Still's term [Eq. (16)] gives directly the transfer energy: $U_{\text{still}} = \Delta F_{\text{pol}}^{\text{transfer}}$. This term has not been introduced as a difference of the polarization energies between the solvent and vacuum media, but as a correction that can be applied to the vacuum energy in order to obtain the energy in solvent. However, to make Still's model compatible with our nonlinear fitting procedure, we formally decompose this term into a vacuum and solvent polarization energy

$$\begin{aligned} U_{\text{Still}} &= \left(\frac{1}{\epsilon_{\text{solv}}} - \frac{1}{\epsilon_{\text{vac}}} \right) U_{\text{Still}}^0 \\ &= \left(\frac{1}{\epsilon_{\text{solv}}} - \frac{1}{\epsilon_{\text{int}}} \right) U_{\text{Still}}^0 - \left(\frac{1}{\epsilon_{\text{vac}}} - \frac{1}{\epsilon_{\text{int}}} \right) U_{\text{Still}}^0 \\ &= U_{\text{Still}}^{\text{solv}} - U_{\text{Still}}^{\text{vac}}. \end{aligned} \quad (26)$$

C. The fittable parameters of the electrostatic part of solvation

Given the unification of the different solvent terms, it is straightforward to introduce these into a nonlinear fitting procedure developed previously,²⁷ in which the electrostatic component of total free energy of the solute in an environment ext has been expressed as a linear combination of the Coulomb and polarization energies, with fittable coefficients

$$F_{\text{electro}}^{\text{ext}} = \alpha_{\text{Coul}}^{\text{ext}} U_{\text{Coul}}^0 + \alpha_{\text{pol}}^{\text{ext}} F_{\text{pol}}^{\text{ext}}. \quad (27)$$

While that study had shown that the F_{pol} terms obtained from the BEM computations are accurate and therefore the fitted α values of Eq. (27) have been found close to the theoretical values of $\alpha_{\text{Coul}}=0$ and $\alpha_{\text{pol}}=1$, this must not necessarily apply to the simpler terms which will be tested here. Substituting F_{pol} with the expressions derived previously, Eq. (27) can be rewritten as linear combinations of other energetic terms

$$F_{\text{electro}}^{\text{ext}} = \beta_{\text{Coul}}^{\text{ext}} U_{\text{Coul}}^0 + \beta_{EQ\text{-Born}}^{\text{ext}} F_{EQ\text{-Born}}^{\text{ext}}. \quad (28)$$

Equation (8) shows that $F_{EQ\text{-Born}}$ contains two contributions: the "classical" Born term and the displacement energy. For the same reasons invoked when deriving alternative definitions for the generalized Born radii, we will replace the difference appearing in Eq. (8) by a linear combination

$$F_{EQ\text{-Born}}^{\text{ext}} = \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}} \right) (U_{\text{Born}}^0 - \gamma \Delta U_{\text{disp}}^0). \quad (29)$$

This must be done in any case when Eq. (10) (Gilson) is used with an unknown average atomic volume which is incorporated into the parameter γ .

D. Modeling the nonelectrostatic part of the solvation process

The same terms as those used in a previous study,²⁷ proportional to the total and hydrophobic molecular surface, are introduced to account for the cavity formation energies in solvent, solute-solvent van der Waals interactions, and "hydrophobic" effects. While certain contributions to the hydrophobic effect, as the cavitation and the solute-solvent van der Waals interactions, do practically not depend on the local polarity/hydrophobicity of the molecular surface and should therefore correlate to the total exposed area S_{tot} , there is still an open debate on whether entropic solvation effects that may display such a dependence are of crucial importance to the overall hydrophobic effect²⁶

$$F_{\text{non-el}}^{\text{solv}} = \sigma_{\text{tot}} \cdot S_{\text{tot}} + \sigma_{\text{hphob}} \cdot S_{\text{hphob}}. \quad (30)$$

The term in hydrophobic surface has not been considered for the displacement models that count a supplementary parameter γ , in order to maintain the same number of fittable coefficients throughout the study

$$F_{\text{non-el}}^{\text{solv}} = \sigma_{\text{tot}} \cdot S_{\text{tot}}. \quad (31)$$

Adding this term to the electrostatic energy $F_{\text{electro}}^{\text{solv}}$ yields the expression of the total solvation potential of a molecule

$$F_{\text{tot}}^{\text{solv}} = F_{\text{electro}}^{\text{solv}} + F_{\text{non-el}}^{\text{solv}}. \quad (32)$$

Of course, in vacuum no surface-dependent contributions need to be added to the electrostatic energy component.

Interestingly, Eisenberg *et al.*¹⁸ and Ooi *et al.*²⁰ have introduced an alternative model, that extends the use of such linear terms of the surfaces of different types of functional groups to account for both nonelectrostatic solvent effects and the polarization of the solvent due to these groups. The central hypotheses of this model are that a given functional group of well-defined charge, which is completely exposed to the solvent will have a constant contribution to the solvation energy, independently of the rest of the charge distribution in the molecule and that this contribution linearly decreases with the degree of solvent accessibility for partially buried groups. The total solvation potential will include only the Coulomb term as an explicit electrostatic contribution and the sum over the considered types of surfaces

$$F_{\text{tot}}^{\text{solv}} = \frac{1}{\epsilon_{\text{solv}}^{\text{eff}}} U_{\text{Coul}}^0 + \sum_{\text{Types}} \sigma_i S_i. \quad (33)$$

We did not follow Eisenberg's choice of surface types, since the set of molecules we considered was different (his study did not include any charged species). The atom types defined in this work are (1) carboxylate oxygens, (2) ammonium hydrogens, (3) polar/neutral groups, (4) aromatic carbon atoms, (5) nonpolar hydrogens, and (6) total area. Of course, the coefficients σ , as well as the inverse dielectric constants, have to be derived by fit and are expected to be strongly dependent on the choice of surface types.

E. Optimization of the set of atomic radii

The optimization of the set of atomic radii is poorly relevant for models displaying a low predictive power. Among the schemes discussed here, only the model of Still justifies the effort. Since the evaluation of Still's volume integrals is highly time consuming, we have used the cross-validated rms value of a linear regression as the quality index of the set of atomic radii during the optimization of the atomic radii. In this linear regression, the intercept was forced to zero and the experimental vacuum-to-water transfer energies were correlated with the Coulomb energies, polarization terms, total and hydrophobic areas of the minimized geometries of the 64 small molecules tested. When using the empirical Eq. (20) and (21), the λ parameter, appearing in the generalized Born radii [Eq. (21)], is fitted simultaneously with the radii. As discussed previously, no probe radius is taken into account as an explicit optimizable parameter, since a single dielectric interface is used in this model. Sulfur and fluorine atoms were no longer included in the list of optimizable radii parameters. All the other definitions of the types of radii have been kept as previously reported.

A different parametrization strategy, introducing charge-dependent radii³⁹⁻⁴¹ has been also applied to the Still model. It has been suggested that the position of the dielectric interface is dependent on the local charge distribution. Highly charged atoms attract more the solvent dipoles and hence tend to "compress" the low dielectric molecular volume (reducing the effective atomic radius). On the other hand, such a strong interaction can provoke "dielectric saturation" effects since the tightly bound water molecules are restricted in movement, so that the effective dielectric constant is locally lowered (this is equivalent to an increased atomic radius). Besides these physical reasons, the introduction of charge-dependent atomic radii could be a valuable parametrization strategy, since the atomic charges actually are measures of the chemical contexts of the atoms. Instead of considering a large set of types of radii in function of the individual chemical environments, we could also express the radius of an atom in function of the *chemical element* and *its fractional charge*, which is related to the local environment of the atom. We have used a simple linear dependence of the *absolute* value of the atomic point charge

$$R_i = R^{\text{chem element}} + \chi |Q_i|, \quad (34)$$

where the parameter χ , $R^{\text{chem type}}$ and λ values have to be optimized concomitantly (the different chemical elements considered for optimization were C, N, O, H, S, and F. The other types were taken with their standard van der Waals radius⁴²).

III. RESULTS AND DISCUSSIONS

A. Linear dependence between different solvation terms: Checking the accuracy of the established theoretical inter-relations

Tests are carried out in order to evaluate to which extent the theoretical relations are verified by the computed electrostatic terms. Least-square linear relations between these terms are derived and compared with the expected theoretical ones. The energetic terms are taken as nonweighted averages over 100 MD conformations (80 ps) for each of the 64 molecules included in this study (a detailed list of which can be found in our previous publication on continuum solvent models²⁷). Whenever the theoretical linear dependence is expected to have a zero intercept, a forced regression with a zero free energy term is applied. The comparisons are made with energy terms calculated with the same dielectric interface used in the BEM calculations in solvent (defined by the Biosym^{42,43} CVFF van der Waals radii + a probe radius of 0.4) and charge parameter set (Biosym CVFF charges). The relations between the approximate and exact displacement energies provide an evaluation of the error that is introduced by the various approximations applied when calculating the volume integrals.

We evaluated the regression slopes (with forced null intercept), root mean square deviations (rms in kcal/mol) and r^2 between the displacement term based on the exact volume integrals (taken as the explained variable Y) and its approximations (used as explicative variable X).

The approximate displacement energy defined by Eq. (9) relates to the "exact" term defined in Eq. (5) as $Y = 1.187X$, with a rms of 2.6 kcal/mol and $r^2 = 0.896$. With the terms according to Eq. (10), we obtained $Y = 0.344X$, a rms of 2.62 kcal/mol and $r^2 = 0.888$, and finally, Eq. (15) led to $Y = 0.799X$, rms = 5.19 kcal/mol and $r^2 = 0.593$.

In the first approximation scheme [Eq. (9)], the coefficient of the regression should be equal to the average volume of the atoms, while in the other two situations [Eqs. (10) and (15)] it should be equal to 1. The coefficient of 1.187 in the first approximation is much too small to be a reasonable estimate of the average atomic volume. None of the approximate terms can be considered to reach a quantitative agreement with the displacement energy calculated with the exact integration procedure. Nevertheless, they are well correlated and this feature can be exploited in empirical models with fittable parameters to replace the computationally expensive integrals with *weighted* approximative terms. The term appearing in Eq. (15), which is a function of the accessible surface area fractions, is expected to overestimate the actual low dielectric volume since it implies that the depth of the molecular dielectric layer on the hidden surface of the atoms is infinite. Therefore, we can conclude that solvation cannot

TABLE I. The linear dependence between the polarization energy as obtained from BEM calculations and solvent displacement terms (linear regressions with forced zero intercepts).

Displacement term	Coefficient of the Coulomb term	Coefficient of the Born term	Coefficient of the displacement term	r^2 (cross-validated)	rms (cross-validated) kcal/mol
Eq. (5)	-0.32	-0.60	0.05 ($\gamma=0.09$)	0.799	4.3
Eq. (9)	-0.34	-0.51	-0.50 ($\gamma=-1.0$)	0.836	3.9
Eq. (10)	-0.34	-0.35	-0.35 ($\gamma=-1.01$)	0.854	3.7
Eq. (15)	-0.33	-0.61	0.05 ($\gamma=0.08$)	0.775	4.6
No displacement term	-0.32	-0.59		0.802	4.3

be properly described in terms of the contributions of the solvent-accessible atoms only and that the buried charges cannot be neglected unless they are located at large distances from any point of the surface. Gilson's terms are in fact a measure of the burial depth of an atom, since a buried atom has a larger number of neighboring atoms that bring large contributions to the sums of inverse interatomic distances.

There is excellent agreement between the polarization energies provided by Still's approach and the results of the simplified BEM, irrespective of the formula used for the evaluation of the generalized Born radii [the parameter λ appearing in Eq. (20) was set to 0.5]

$$F_{\text{BEM}} = -0.477F_{\text{Still}}$$

with a rms of 0.569 kcal/mol and $r^2=0.996$.

The theoretical slope should be equal to -0.487 [see Eq. (26)]. Despite the large conceptual differences between the two methods, their results appear to be highly comparable. This speaks for the consistency of these models.

If dielectric boundary effects can be neglected, then the solvent displacement terms should be linearly related to the polarization energy, as can be easily seen by substituting the generalized expression of $F_{\text{EQ-Born}}$ [Eq. (29)] into Eq. (24). The linear relations between the polarization energy as obtained from the BEM calculations and the Coulomb, Born and different displacement terms are listed in Table I.

It can be seen that the displacement energies, even when evaluated by precise integration, cannot be linearly correlated to the polarization energies. There is a strong correlation between the polarization energies and the Coulomb and Born terms since these terms taken together are able to account for the pattern of charge distribution and therefore correctly predict the order of magnitude of the solvent effects. The regression coefficients of the Born and Coulomb terms are compatible with the inverse dielectric constant differences appearing in Eq. (24).

The displacement terms are not good descriptors of the polarization energy, irrespective of the approach used for the evaluation of the volume integrals. In the best cases, the displacement terms only slightly increase the correlation coefficients but enter into the relation with the wrong sign (the parameter γ should be +1). We conclude that the hypotheses

introduced when defining the displacement terms are not a valid framework to describe the energetics of building up a molecule from infinitely separated atoms. These terms provide nevertheless information about the electrostatic properties of the molecule. The relations discussed previously can be considered as empirical QSAR equations⁴⁴ relating a complex molecular property (the polarization energy) to simple electrostatic descriptors. Still's empirical function is a nonlinear relation based on these terms, which is able to circumvent the limitations introduced by the hypotheses leading to the linear expressions. Our results, however, do not invalidate the utility of the linear Gilson-Honig model in the kind of problems it has been originally designed for: the prediction of the solvation energy differences upon conformational change of a macromolecule or ligand binding phenomena.²⁹ The description of the displacement phenomena adopted here is actually accurate if the distance separating the intervening atoms is large enough. The failure to reproduce the absolute values of the polarization energies is due to very large, but unreliable contributions of the atom pairs that are brought from infinity to overlapping distances of their spheres. Good correlations between conformational polarization energy differences and conformational displacement energy differences have been established; they will be reported elsewhere.⁴⁵

Since the same dielectric boundary was considered in this study, the polarization terms in vacuum and solvent dif-

TABLE II. Comparison between the results of a linear and nonlinear Boltzmann-weighted solvation model, using a Still polarization term. The linear fit coefficient should be approximatively equal to the difference of solvent-vacuum coefficients.

Coefficient	Solvent	Vacuum	Linear
Coulomb	0.72	0.69	0.03
Still term	-0.54	0.52	-1.06
Total area	-0.04	0.00	-0.04
Hydrophobic area	0.05	0.00	0.05
rms	2.2 kcal/mol		2.6 kcal/mol
Maximal error	6.5 kcal/mol		7.0 kcal/mol
r^2	0.988		0.985

TABLE III. The different continuum solvent approaches tested. The optimization strategies of the radii are (A) assignment of the radii in function of the chemical environment of the atom, as reported in Ref. 27; (B) assignment of radii in terms of the chemical type and charge-dependent correction, according to Eq. (34); No refers to no radii optimization (original vdWals Biosym radii+0.4 Å offset are used).

Nr.	Type	Evaluation of disp. integrals	Generalized Born radii	Polarization energy term	Optimization of radii	Charge set	Surface dependent terms
1	Disp.	Eq. (5)		Eq. (24)	No	Biosym	Eq. (31)
2	Disp.	Eq. (9)		Eq. (24)	No	Biosym	Eq. (31)
3	Disp.	Eq. (10)		Eq. (24)	No	Biosym	Eq. (31)
4	Disp.	Eq. (15)		Eq. (24)	No	Biosym	Eq. (31)
5	Still	Eq. (3)	Eq. (7)	Eq. (26)	No	Biosym	Eq. (30)
6	Still	Eq. (3)	Eq. (7)	Eq. (26)	A	Biosym	Eq. (30)
7	Still	Eq. (3)	Eq. (20)	Eq. (26)	A	Biosym	Eq. (30)
8	Still	Eq. (3)	Eq. (21)	Eq. (26)	A	Biosym	Eq. (30)
9	Still	Eq. (3)	Eq. (21)	Eq. (26)	B	Biosym	Eq. (30)
10	Gilson-Still	Eq. (9)	Eq. (21)	Eq. (26)	A	Biosym	Eq. (30)
11	Still	Eq. (3)	Eq. (20)	Eq. (26)	A	Mopac ESP	Eq. (30)
12	Gilson-Still	Eq. (9)	Eq. (21)	Eq. (26)	B	Mopac ESP	Eq. (30)
13	Eisenberg			Eq. (33)	No	Biosym	Eq. (33)

fer from each other only by their ϵ -dependent factor, being proportional to each other and to the polarization energy difference. Negelecting the effect of the Boltzmann averaging upon the families of conformations used in our nonlinear fit, linear models can be established between the vacuum-to-water transfer energies and the unweighted averages of various calculated solvation terms

$$\Delta F_{\text{tot}}^{\text{transfer}} = a_{\text{pol}} F_{\text{pol}}^0 + a_{\text{Coul}} U_{\text{Coul}}^0 + a_{\text{tot}} S_{\text{tot}} + a_{\text{hphob}} S_{\text{hphob}}$$

In this study we have used small rigid molecules, displaying no large conformational changes. The Boltzmann weighting is not supposed to lead to large differences with respect to the unweighted averages and therefore, the a coefficients in Eq. (35) should be close to the differences between the solvent and vacuum values of the corresponding terms in the nonlinear fit. A typical example is shown in Table II. The agreement between the two classes of coefficients is a proof that the nonlinear fit converged properly.

TABLE IV. Predictive power of the tested continuum models. Root mean square and maximal error are given in kcal/mol. Model numbering refers to Table III.

Model	rms (cross-validated)	r^2 (cross-validated)	Maximal error	Mispredicted compounds and other remarks
1	5.8	0.918	18.6	Trifluoroethanol; ammonium ions mispredicted by ~17 kcal/mol
2	5.2	0.934	13.4	Trifluoroethanol; ammonium ions mispredicted by ~17 kcal/mol
3	6.2	0.912	24.3	Trifluoroethanol; ammonium ions mispredicted by ~17 kcal/mol
4	7.7	0.863	25.6	Trimethylammonium; this model cannot handle buried charges
5	2.2	0.988	6.5	Trifluoroethanol; NH_4^+ mispredicted by ~5 kcal/mol
6	1.6	0.993	4.6	Trifluoroethanol; NH_4^+ is correctly predicted (-79.2 kcal/mol)
7	1.4	0.995	4.7	Difluoroethane
8	1.5	0.995	4.8	Difluoroethane
9	1.5	0.994	3.1	Neutral Arg side chain; the fluorine compounds are well predicted when using the charge-dependent radii
10	1.6	0.994	5.7	Difluoroethane
11	1.2	0.996	3.4	Arg+side chain
12	1.8	0.992	4.9	Nitroethane
13	6.5	0.901	42.2	Trimethylammonium; this model cannot handle buried charges

TABLE V. Averages and rms fluctuations of the coefficients in the nonlinear model using *displacement terms*. In parentheses, the rms fluctuations. Model numbering refers to Table III.

Model	Total surface coeff.	Coulomb coeff. in solvent	Born coeff. in solvent	Born coeff. in vacuum	Coulomb coeff. in vacuum	Weighting factor of displacement term (γ)
1	0.06 (0.00)	-0.63 (0.03)	-0.67 (0.03)	0.68 (0.03)	0.63 (0.03)	0.05 (0.14)
2	0.07 (0.00)	-0.66 (0.02)	-0.58 (0.01)	0.58 (0.01)	0.66 (0.02)	-0.95 (0.11)
3	0.06 (0.01)	-0.66 (0.08)	-0.56 (0.05)	0.45 (0.03)	0.66 (0.08)	-0.48 (0.15)
4	0.07 (0.01)	-0.42 (0.12)	-0.38 (0.09)	0.46 (0.05)	0.42 (0.12)	-0.88 (0.31)

B. Predictive power of the models: Optimal parameters

Table III displays the different models and resumes the hypotheses on which these are based. The predictive power of the nonlinear fits is displayed in Table IV, while the corresponding optimal weighting factors are given in Tables V, VI, and VII. As expected, the Still models perform much better than the simple displacement or surface-dependent ones. An optimization of the parameters of these schemes has been performed, the results are shown in Table VIII. It has already been discussed that the polarization energies cannot be rigorously obtained from a linear combination of Coulomb, Born, and displacement terms. Part of the low predictive power of the displacement models can be ascribed to the missing term in hydrophobic exposed areas, which was neglected in order to keep six fittable parameters in the model. The best displacement model turns out to be the one using Gilson's term of Eq. (9) and not the one calculated by exact volume integrals of Eq. (5). However, it can be seen that the weighting coefficients, and especially the γ value, are quite unstable during the cross-validation procedures (i.e., γ in

scheme 2 adjusts to values between -0.65 and -1.05 in function of the set of compounds excluded from the fit). The average values of the coefficients in the solvent are in good agreement with the results of the linear correlations of Table I, except for scheme 4 which uses fractions of accessible surface. Here, the fitted coefficients of both the Coulomb and the Born term should equal in theory $[(1/\epsilon_{\text{ext}})-(1/\epsilon_{\text{int}})]$.

It is interesting to compare the solvation energies reported by Still in his original paper with the results of our personal implementation of this method. Of course, there is an overall good agreement between both models and the experimental data. The small differences appearing in Table IX arise from the different parametrization of atomic charges and radii (OPLS and Biosym, respectively) and from the use of adjustable parameters in our model. It can be seen that Still's original results are comparable to data for the uncharged species, despite the fact that he uses a single corrective term in function of the total exposed area in contrast to the two surface terms (total and hydrophobic areas) applied in the current work. The OPLS parameters are maybe more adopted to describe uncharged molecules. On the contrary, model 5 is largely superior in the prediction of the transfer energies of charged species. The overall rms of calculated vs. experimental data listed in Table IX is 2.1 kcal/mol for model 5 and 4.5 kcal/mol for the original Still's values, due to the large misprediction of the ammonium ion in the latter. The improvement of predictive power in our study can be partly ascribed to the use of fittable parameters, although these have spontaneously adopted values that are very close to the theoretical ones. From Table VI we obtain the global contribution of the Still term to the transfer energy in model 5 as $(-0.540-0.525)U_{\text{Still}}=-1.065U_{\text{Still}}$, to be compared with $-(1-1/78)U_{\text{Still}}=-0.987U_{\text{Still}}$ in Still's original calculation. The small difference of 0.08 in the weighting coefficient of Still's term is nevertheless very important for charged species. For the ammonium ion with $U_{\text{Still}} \approx 80$ kcal/

TABLE VI. Averages and rms fluctuations of the coefficients in the nonlinear model using *Still polarization terms* (in parentheses, the rms fluctuations). Model numbering refers to Table III.

Model	Total surface coeff.	Hydrophobic surface coeff.	Coulomb coeff. in solvent	Polarization coeff. in solvent	Polarization coeff. in vacuum	Coulomb coeff. in vacuum
5	-0.039 (0.002)	0.048 (0.003)	0.360 (0.001)	-0.540 (0.009)	0.525 (0.009)	0.349 (0.002)
6	-0.014 (0.002)	0.036 (0.002)	0.355 (0.001)	-0.516 (0.001)	0.476 (0.001)	0.350 (0.002)
7	-0.0142 (0.0006)	0.0247 (0.0007)	0.3533 (0.0006)	-0.637 (0.004)	0.598 (0.004)	0.352 (0.001)
8	-0.0141 (0.0006)	0.0258 (0.0006)	0.3534 (0.0006)	-0.613 (0.004)	0.578 (0.003)	0.352 (0.001)
9	-0.0040 (0.0004)	0.0128 (0.0008)	0.3542 (0.0002)	-0.508 (0.001)	0.467 (0.001)	0.3512 (0.0005)
10	-0.020 (0.001)	0.0288 (0.0009)	0.3479 (0.0007)	-0.578 (0.001)	0.539 (0.001)	0.3580 (0.0007)
11	-0.0067 (0.0006)	0.0229 (0.0006)	0.3582 (0.0007)	-0.565 (0.001)	0.525 (0.001)	0.3474 (0.0007)
12	0.017 (0.002)	-0.011 (0.002)	0.3456 (0.0006)	-0.456 (0.002)	0.415 (0.002)	0.3480 (0.0004)

TABLE VII. Averages and rms fluctuations of the coefficients in the nonlinear model using the *solvent-accessible surfaces* of different functional groups (in parentheses, the rms fluctuations).

Electrostatic terms		Coefficients multiplying the surfaces of functional groups				
Coulomb coeff. in solvent	Coulomb coeff. in vacuum	Carboxylate oxygens	Ammonium hydrogens	Aromatic carbons	Nonpolar hydrogens	Total area
0.281 (0.003)	0.286 (0.004)	-1.428 (0.018)	-1.39 (0.05)	0.10 (0.01)	0.10 (0.01)	-0.11 (0.01)

mol, this can affect the final result by ~ 6 kcal/mol. The difference of $0.3602 - 0.3485 = 0.012$ in the Coulomb weighting factors has a somehow less important influence. However, Still's large error in the prediction of the ammonium ion is certainly caused by the OPLS parameters, leading to a much too negative value of $U_{\text{Still}} = -91.2$ kcal/mol. The same tendency is observed for all charged species under the OPLS force field.

The comparison of the transfer energies of ethane and octane (see Table IX) may serve as an example of the difficulties that may arise when trying to give a detailed, physical interpretation of the results obtained from empirical force field calculations. Considering the van der Waals radii of 1.1 Å for H and 1.55 Å for C and a probe radius of 0.4 Å (e.g., effective radii of 1.5 and, respectively, 1.95, as used in model 5, the calculated van der Waals area ethane is 85.5 \AA^2 , compared to 202 \AA^2 for octane. The surface of these molecules is—by definition—entirely “hydrophobic” and therefore it is straightforward to calculate an “experimental” hy-

drophobic surface coefficient equal to $(2.9 - 1.8) \text{ kcal}/(202 - 85.5) \text{ \AA}^2 = 9.44 \text{ cal}/\text{\AA}^2$. In Still's parametrization, the hydrocarbon atoms are considered to be uncharged and thus the polarization contribution is zero: the only contribution to the transfer energies of hydrocarbons arises from the surface effects. Still's surface term of $7.2 \text{ cal}/\text{\AA}^2$ is slightly smaller, but in good agreement to this “experimental” value. In our model based on CVFF point charges, we predict a difference in transfer energies of 1.8 kcal/mol instead of the experimental 1.1 kcal/mol. However, our polarization contributions are no longer negligible, but amount -0.12 kcal/mol for ethane and $+0.49$ kcal/mol for octane. Therefore, according to our model, the difference of the *hydrophobic* contributions to the solvation is again only 1.2 kcal/mol. The actual hydrophobic surface coefficient in our work—see Table V—equals the sum of the “total area” coefficient ($-39 \text{ cal}/\text{\AA}^2$) and the “hydrophobic area” coefficient ($+48 \text{ cal}/\text{\AA}^2$), e.g., $9 \text{ kcal}/\text{\AA}^2$. In other words, our larger error in predicting the transfer energies of nonpolar fragments is not due to an in-

TABLE VIII. Optimal sets of radii (in Å) and displacement weighting factors for the tested *Still models*. Model numbering refers to Table III.

Symbol	Definition of the atomic types Chemical context	Model						
		6	7	8	9	10	11	12
C	Aliphatic	1.586	2.204	2.234	1.883	1.883	1.829	1.995
	Unsaturated sp^2 and sp	1.706	2.255	2.258	2.176	2.176		
	Aromatic	1.894	2.500	2.458	2.105	2.105		
	Carbonyl and carboxylate groups	2.747	3.218	3.277	2.807	2.807		
H	Aliphatic	1.237	1.693	1.680	1.584	1.584	1.119	1.149
	N bound	1.328	1.780	1.820	1.658	1.658		
	O bound	3.254	4.632	4.417	3.748	3.748		
N	sp^2	1.214	1.504	1.523	1.557	1.557	1.011	1.625
	sp^3	1.003	1.267	1.296	1.391	1.391		
O	sp^3	1.476	1.744	1.741	1.935	1.935	1.223	1.169
	Carbonyl	1.388	1.453	1.553	1.806	1.806		
	Carboxylate	1.759	2.130	2.144	2.141	2.141		
S	Thiols and thioethers			fixed at 2.21			5.310	1.748
F	Fluorocarbons			Fixed at 1.70			2.913	1.779
	Displacement weighting factor λ	Fixed at 1.0	0.346	0.492	0.569	0.569	0.387	0.327
	Charge-radius coupling parameter χ						0.785	0.234

appropriate parametrization of the surface terms, but due to the important contributions of the electrostatic terms, arising from the large partial charges in the CVFF force field. Furthermore, the *positive* polarization contribution of 0.49 kcal/mol for octane is obviously wrong, since the transfer of any charge distribution from vacuum into a high dielectric is *favorable*. This is an artefact of the empirical function of Still, which does not offer any guarantees of physical consistency despite its overall success in reproducing vacuum-to-water transfer energies. Indeed, the polarization contribution calculated for octane with the boundary element method²⁷ is negative, as expected, and is counterbalanced by a larger contribution from the hydrophobic areas (15–20 cal/Å²). Two main conclusions therefore emerge from the previous discussion

—The “hydrophobic coefficients” strongly depend on the issue of whether the polarization contributions arising from the partial charges of the hydrocarbon atoms are implicitly included in the surface terms or treated explicitly.

—The fitted parameters used in continuum solvent models are context-specific empirical force field parameters. Only the sum of the different contributions to the transfer energies has been calibrated to fit the experimental values and this does not imply that each one of these contributions correctly represents the particular effects it is supposed to account for.

The optimization of the atomic radii is found to improve significantly both the predictive power and physical consistency of the coefficients. The difference between the fittable coefficients of the Still term in model 6 (−0.991) is practically equal to the theoretical $1/\epsilon_{\text{solv}} - 1 = -0.987$ while the difference between the coefficients of the Coulomb term is almost zero, as expected. However, some of the atomic radii, especially those of the O-bound hydrogen atom, systematically appear to be unphysically large. Since in a previous study,²⁷ an optimization of the radii used in connection with boundary element calculations of the solvation energies of strictly the same set of 64 small molecules lead to reasonable values for the radii of this chemical type (0.8–1.2), we believe that the currently obtained radii are due to the empirical character of Still’s function in contrast to the better founded BEM approach. We have argued in our previous study that these atomic radii should be regarded rather as empirical force field parameters than as physically founded atomic properties, since there is no sharp dielectric interface separating the molecule from its environment. As far as the method used for the calculation of the transfer energies is based on a set of physically reasonable hypotheses, the resulting atomic radii should remain close to what is generally considered to be the correct atomic “size.” However, the only measure of the “correctness” of such parameters in the context of Still’s empirical function is their success in reproducing experimental vacuum-to-water transfer energies, which is sensibly improved when using the “counterintuitive” optimal radii.

The introduction of the weighting factors of the displacement term (λ) further increases the predictive power of the models. This might simply be a consequence of the increased

TABLE IX. Comparison between the original values of transfer energies (in kcal/mol) reported by Still and the results of model 5 (see Table III), using the Biosym CVFF fractional charge set; the atomic radii were taken as Biosym CVFF van der Waals radii +0.4 Å offset.

Compound	Still’s value	Predicted by model 5	Experimental
Ethane	+1.3	0.7	1.8
<i>n</i> -octane	+2.9	2.5	−2.9
Methanol	−6.2	−3.7	−5.1
Ethanol	−5.2	−3.7	−5.0
Acetone	−3.2	−4.2	−3.8
Acetic acid	−6.5	−6.7	−6.7
Acetamide	−10.6	−8.7	−9.7
Benzene	−1.0	−2.9	−0.9
Phenol	−6.3	−5.8	−6.6
Ammonium	−90.8	−84.1	−79.0
Methylammonium	−80.4	−71.0	−71.0
Trimethylammonium	−63.1	−55.4	−59.0
Acetate	−82.9	−75.6	−78.5

number of fittable parameters. Unfortunately, the optimization of the λ parameters was done simultaneously with that of the radii and was not subjected to a cross-validation test that could confirm their significance. The condition of $\lambda < 1$ is fulfilled in all cases. Furthermore, the ratio of the weighting parameters of the exact displacement term in model 8 and the approximate Gilson term in model 10 reproduces the least-square relations between these terms as appearing above [see regression equation of the approximation defined by Eq. (9)]. On the other hand, the weighting coefficients of the Still terms in Table VI are seen to deviate more from the theoretical values for the models 7 and 8. Both empirical equations (20) and (21) used to define the generalized Born radii show comparable performances. In particular, they lead to very close sets of optimized radii, but display somehow different λ values.

The main interest in introducing the supplementary λ parameter resides in the possibility of coupling the readily available Gilson approximation with the Still model. The resulting model 10 shows a comparable (slightly better) predictive power than model 6 (based on Still’s original term) while being computationally fast enough to allow recalculation of the displacement terms at every integration step of a molecular dynamics run.

The comparison of models 9 and 12, both using *charge-dependent* radii, shows that model 9 (using the Biosym charge set) performs better than model 12, which uses Mopac-ESP atomic charges. In the first model, the atomic fractional charges are explicitly assigned in terms of the chemical contexts of the atoms according to the CVFF rules, while in the second the correlation between the chemical context and the atomic charge is somehow weaker. All this suggests that the charge dependence of radii plays a more important role in accounting for the chemical contexts of atoms than in reflecting the physical effects such as dielectric saturation or cavity compression. The positive χ values that were obtained cannot be used as evidence that the dielectric

saturation effects are stronger than the cavity compression in the vicinity of a charged atom.

The main methodological interest of this strategy lies in the much smaller number of parameters (one standard radius value per atomic type plus the coefficient of the radii-charge dependence) whereas the number of different chemical contexts is practically unlimited. The use of the charge-dependent radii allows to reduce the number of parameters to 8 instead of 13, as seen in Table VIII while including at the same time the radii of sulfur and fluorine atoms into the optimization procedure. These were kept fixed during the optimization of the radii in function of the chemical context of the atoms, where the number of variables was already large. Although the function used to express this dependence is a simple empirical guess, see Eq. (34), the results are quite encouraging. Eliminating five supplementary fittable parameters only increases the rms of the model by less than 0.1 kcal/mol (under Biosym charge parametrization). In our previous work, the radius of fluorine was optimized according to a chemical context strategy. The fluorine radius has not been significantly modified during the optimization run and the fluorinated species continued to be the worst predicted compounds by the BEM calculation. Within the strategy of charge-dependent radii, the prediction of the transfer energies of fluorinated compounds is significantly improved (errors of only 0.5 kcal/mol for trifluoroethanol and 2.3 kcal/mol for 1,1-difluoroethane). This suggests that this latter strategy can be successful in situations where the definition of a new chemical context is required.

IV. CONCLUSIONS

Still's empirical solvation term has been built into the general frame of continuum solvation models based on a nonlinear fit procedure. Despite its conceptual and mathematical simplicity, this term turns out to be an accurate description of the polarization effects, having at least the quality of the more elaborated and physically realistic boundary element method which has been previously investigated. Still's model has been studied in the general context of "displacement" solvation models and we have shown that combining the displacement terms to the polarization energy in a nonlinear way is crucial for any accurate prediction. Different alternative modifications including a new fittable parameter have been proposed for the estimation of the "generalized Born radii" used in Still's formalism. This class of models leads to better results than the original derivation. Gilson's very simple approach for evaluating the displacement terms is not accurate enough to be used as such in quantitative evaluations of solvent effects, although it can be easily coupled to the Still formula, leading to a very fast model of high predictive power. Its implementation into a molecular dynamics algorithm should be straightforward, since it utilizes only the interatomic distances as variables. The optimization of the atomic radii defining the molecular interface turns again to be important for obtaining models of high predictive power. A new radii optimization strategy assuming the dependence of the atomic radii on the partial

charge of the atom has been proposed. This parametrization strategy is found to lead to quite satisfactory results, and provided its context independence, can be easily adapted to work with any molecular force fields.

ACKNOWLEDGMENTS

D. Horvath thanks Professor A. Tartar, Institut Pasteur de Lille, for the position offered in his laboratory, and to all organizers of the Joint European Laboratory between Institut Pasteur de Lille and Université Libre de Bruxelles, who have created a proper framework for this collaboration. He also acknowledges the financial support from Conseil Régional du Nord, France. D. Van Belle acknowledges support from the European Community Biotechnology BRIDGE Contract (No. CT91-0270).

- ¹D. Bashford, *Curr. Op. Struct. Biol.* **1**, 175 (1991).
- ²D. L. Beveridge and F. M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.* **18**, 431 (1989).
- ³J. A. McCammon, *Annu. Rev. Biophys. Biophys. Chem.* **1**, 196 (1991).
- ⁴E. Clementi, *J. Chem. Phys.* **74**, 578 (1979).
- ⁵D. van Belle, I. Couplet, M. Prevost, and S. J. Wodak, *J. Mol. Biol.* **198**, 721 (1987).
- ⁶D. van Belle, M. Froeyen, G. Lippens, and S. J. Wodak, *Mol. Phys.* **77**, 239 (1992).
- ⁷R. O. Watts, *Mol. Phys.* **28**, 4 (1974).
- ⁸A. A. Rashin and B. Honig, *J. Phys. Chem.* **89**, 5588 (1989).
- ⁹B. Roux, H. A. Yu, and M. Karplus, *J. Phys. Chem.* **94**, 4683 (1990).
- ¹⁰A. Jean Charles, A. Nicholls, K. Sharp, B. Honig, A. Tempczyk, T. F. Hendrickson, and W. C. Still, *J. Am. Chem. Soc.* **113**, 1454 (1991).
- ¹¹M. K. Gilson and B. Honig, *Proteins* **3**, 32 (1988); **4**, 7 (1988).
- ¹²M. K. Gilson and B. H. Honig, *Proteins* **3**, 32 (1988).
- ¹³M. K. Gilson, M. E. Davis, B. A. Luty, and J. A. McCammon, *J. Phys. Chem.* **97**, 3591 (1993).
- ¹⁴J. Moulton, *Annu. Rev. Biophys. Biophys. Chem.* **2**, 223 (1992).
- ¹⁵C. J. Cramer and D. G. Truhlar, *Rev. Comput. Chem.* **6**, 1 (1995).
- ¹⁶K. Sharp, A. Jean-Charles, and B. Honig, *J. Phys. Chem.* **96**, 3822 (1992).
- ¹⁷K. Sharp, *Annu. Rev. Biophys. Biophys. Chem.* **1**, 171 (1991).
- ¹⁸D. Eisenberg and A. D. McLahan, *Nature (London)* **319**, 199 (1986).
- ¹⁹T. Simonson and A. T. Bruenger, *Proteins* (in press).
- ²⁰T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **84**, 3086 (1987).
- ²¹W. C. Still, A. Tempczyk, R. C. Hawley, and T. F. Hendrickson, *J. Am. Chem. Soc.* **112**, 6127 (1990).
- ²²K. Sharp and B. Honig, *J. Phys. Chem.* **94**, 7684 (1990).
- ²³A. H. Juffer, E. F. F. Botta, B. A. M. van Keulen, A. van der Ploeg, and H. J. C. Berendsen, *J. Comput. Phys.* **97**, 8 (1991).
- ²⁴A. A. Rashin and K. Nambodiri, *J. Phys. Chem.* **91**, 6003 (1987).
- ²⁵A. A. Rashin, *J. Phys. Chem.* **94**, 1725 (1990).
- ²⁶B. Lee, *Biopolymers* **31**, 993 (1991).
- ²⁷D. Horvath, D. v. Belle, G. Lippens, and S. J. Wodak, *J. Chem. Phys.* **104**, 6679 (1996).
- ²⁸K. Sharp, *J. Comput. Chem.* **12**, 454 (1991).
- ²⁹M. K. Gilson and B. Honig, *J. Comput. Aided. Mol. Des.* **5**, 5 (1991).
- ³⁰P. F. W. Stouten, C. Froemmel, H. Nakamura, and C. Sander, *Mol. Simulation* **10**, 97 (1993).
- ³¹B. A. Luty, R. Wasserman, P. F. W. Stouten, and C. N. Hodge, *J. Comput. Chem.* **4**, 454 (1995).
- ³²G. Nemethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.* **87**, 1883 (1983).
- ³³J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975).
- ³⁴M. Born, *Z. Phys.* **1**, 45 (1920).
- ³⁵E. L. Mehler and T. J. Solmajer, *Biochemistry* **23**, 3887 (1984).
- ³⁶E. L. Mehler and T. J. Solmajer, *Protein Eng.* **4**, 903 (1991).
- ³⁷E. L. Mehler and T. J. Solmajer, *Protein Eng.* **4**, 911 (1991).
- ³⁸L. Onsager, *J. Am. Chem. Soc.* **58**, 1486 (1936).
- ³⁹C. J. Cramer and D. Truhlar, *J. Comput. Chem.* **13**, 1089 (1992).

- ⁴⁰C. J. Cramer and D. Truhlar, *J. Comput.-Aided Mol. Design* **6**, 629 (1992).
- ⁴¹S. W. Rick and B. J. Berne, *J. Am. Chem. Soc.* **116**, 3949 (1994).
- ⁴²A. T. Hagler, S. Lifson, and P. Douber, *J. Am. Chem. Soc.* **101**, 5122, 6842 (1979).
- ⁴³*Discover User Guide*, Version 3.1. (Biosym Technologies, San Diego, 1993).
- ⁴⁴C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir, *Nature (London)* **194**, 178 (1962).
- ⁴⁵D. Horvath, *J. Med. Chem.* (submitted).

Introduction à la théorie générale du champ de forces empirique, à la Mécanique et Dynamique moléculaire et au DOCKING de substrats enzymatiques

Cette introduction présente une brève discussion des techniques de modélisation moléculaire appliquées utilisées actuellement. Le concept de "champ de force empirique" décrivant les interactions interatomiques ainsi que la méthodologie associée aux techniques de simulation basées sur une telle description des interactions, de même que le rôle de certains paramètres intervenant dans les simulations (p. ex. la distance de troncature des interactions à longues portées) seront développés. Le formalisme permettant de relier les propriétés microscopiques à leur équivalent macroscopique (mesurable) sera brièvement abordé. Nous terminerons ce chapitre par les problèmes d'évaluation de l'affinité de ligands pour une enzyme ainsi que du placement d'un ligand dans le site actif d'une enzyme ("docking").

1. L'approche empirique du traitement des interactions moléculaires.

Par opposition aux méthodes de calcul quantique qui résolvent les équations d'onde de Schrödinger à différents niveaux d'approximation (plus ou moins) simplificatrices, les méthodes de mécanique (MM) et de dynamique moléculaire (MD) utilisent une description classique des interactions atomiques. Chaque atome est considéré comme un point matériel qui génère autour de lui un champ de forces en accord avec ses propriétés chimiques. Evidemment, aucune équation physique exacte ne peut décrire un tel champ de forces, la physique classique n'étant pas applicable au niveau moléculaire. Néanmoins, on peut imaginer des équations empiriques décrivant de façon satisfaisante les interactions atomiques. Cette idée est bien antérieure au développement récent des moyens informatiques, moyens qui ont ensuite permis d'en tester la validité. Ainsi, Bjerrum¹ a proposé en 1914 deux types de champ de forces empirique : le *champ de forces de valence* qui contient des termes d'élongation de liaison et de déformation angulaire et le *champ de forces centrales* fondé sur des interactions atomiques dirigées selon les axes internucléaires. Il a utilisé ces deux modèles pour expliquer le spectre de

¹ N.Bjerrum, *Deut.Physik.Ges.Ber.*, 16,737 (1914)

vibration du dioxyde de carbone. Urey et Bradley² ont introduit un terme quadratique de correction pour les interactions entre atomes géminaux. Après 1960, les moyens informatiques ont permis la diversification et le raffinement des différents modèles de champ de forces. Pour une revue sur ce sujet, voir^{3,4}.

Les champs de forces utilisés aujourd'hui sont pratiquement tous des hybrides des prototypes proposés par Bjerrum. Ils contiennent autant de termes "de valence" décrivant les déformations des liaisons et des angles, que de termes d'interactions centrales entre atomes non-liés. Concrètement, un champ de forces est constitué par un ensemble de fonctions qui décrivent les changements en énergie potentielle moléculaire dus aux variations des positions atomiques. Dans un tel champ de forces, l'énergie moléculaire est décomposée sous la forme d'une somme de contributions rattachées aux différentes coordonnées internes, elles-mêmes fonctions des coordonnées atomiques. Il reste à définir ces coordonnées internes et à établir leurs contributions respectives à l'énergie potentielle.

a) la déformation des liaisons de valence (*bond stretching*) peut être en première approximation assimilée à la déformation d'un ressort (oscillateur harmonique): la contribution à l'énergie potentielle va dans ce cas augmenter proportionnellement au carré de la déviation de la longueur de liaison b par rapport à la longueur de référence b_0 , pour laquelle l'énergie est nulle.

$$V_b = k_b (b - b_0)^2 \quad (1a)$$

b) de la même façon, nous pouvons traiter la déformation des angles de valence (*angle bending*):

$$V_\phi = k_\phi (\phi - \phi_0)^2 \quad (1b)$$

c) pour décrire la variation du potentiel due à la rotation autour d'une liaison s , une fonction périodique doit être employée. Dans (1c) n représente la multiplicité de la barrière de rotation. On introduit également des potentiels de torsion qui empêchent la rotation autour d'une double liaison (potentiel nul pour $\theta=0, 180^\circ$ et maximal pour 90°).

² H.C.Urey and C.A.Bradley, *Phys.Rev.*, 38,1969 (1931)

³ S.R.Niketic and K.Rasmussen, "The Consistent Force Field" Springer, New York, 1977

⁴ N.L.Allinger and Y.H.Yuh, "MM2 Manual", San Diego Supercomputer Center, GA Technologies, San Diego (CA), (1988)

$$V_{\theta} = k_{\theta} (1 + \cos n.\theta) \quad (1c)$$

Des termes "mixtes"⁵ (*cross-terms: stretch-bend, bend-torsion...*) sont employés dans les champs de force très élaborés pour aboutir à expliquer les spectres de vibration des composés organiques.

d) l'interaction entre deux atomes non liés est décrite par une fonction qui prend en compte l'augmentation d'énergie lors du rapprochement trop important de deux atomes (interpénétration des nuages électroniques). A longue distance, l'interaction est attractive (potentiel négatif) et devient répulsive (potentiel positif) au fur et à mesure que r diminue (forces van der Waals). Les fonctions (1d) satisfont ces exigences (A, B et c sont des constantes positives dépendant des types d'atomes en interaction et r représente la distance qui les sépare). Ceci reste vrai uniquement si la puissance n de r dans le terme répulsif est supérieure à celle du terme attractif. On utilise ainsi plusieurs fonctions de ce type (L-J; Lennard-Jones) avec une puissance de r pour le terme répulsif variant de 9 à 12, mais aussi des termes répulsifs exponentiels (B; Buckingham). Le terme attractif est toujours en r⁶, ceci est l'expression théorique déduite par London⁶ pour les forces dispersives:

$$V_{L-J} = \frac{A}{r^n} - \frac{B}{r^6} \quad (1d)$$

$$V_B = A e^{-cr} - \frac{B}{r^6}$$

e.) enfin, la distribution de charges dans une molécule implique des interactions de type électrostatique, qui ne peuvent être discutées sans faire référence aux problèmes de solvation, principalement dans le cas de solvants polaires. Evidemment, les interactions électrostatiques ne sont pas le seul type d'interaction à intervenir dans la description de la solvation⁷, mais elles représentent la contribution majeure et la mieux connue des effets de solvant. Malgré l'apparente simplicité du terme Coulombien:

$$U_{Coul} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r.r} \quad (1e)$$

⁵ J.R. Maple, T.S.Thacher, U. Dinur, A.T. Hagler, *Chemical Design Automation News*, 5 (9),5 (1990)

⁶ P.W. Atkins, "Physical Chemistry", Oxford University Press, London (1982)

⁷ voir la section traitant des effets de solvation pour une discussion plus détaillée.

le traitement de ces interactions pose un grand nombre de problèmes dans la modélisation moléculaire. Ceci est dû notamment au caractère à longue portée des interactions entre les charges q dans un environnement hétérogène ou il n'est pas évident de définir la constante diélectrique effective ϵ_r .

L'ensemble des termes précédents ne représente pas toutes les interactions à prendre en considération pour l'évaluation de l'énergie moléculaire, mais uniquement les contributions qui peuvent être exprimées sous la forme des équations très simples (1a)-(1e). D'autres effets importants, notamment les effets de polarisation⁸ et l'effet "hydrophobe"⁹ ne sont pas couramment inclus dans des champs de forces, en raison de leur complexité ou du manque de compréhension théorique du phénomène.

2. Simulations dans le vide ou dans le solvant. Conditions aux bords ("Boundary Conditions") imposées par la nature à longue portée des interactions électrostatiques - rayon de troncature ("cutoff distance" d_{cut})

Dans le cas d'une molécule isolée dans le vide, l'énergie électrostatique revient au calcul du terme de Coulomb entre les charges atomiques portées par chaque atome. Les simulations dans de telles conditions ne décrivent pas correctement le comportement des molécules en solution, ce qui limite drastiquement leur domaine d'applicabilité.

En accord avec notre discussion antérieure, le caractère diélectrique du solvant résulte implicitement des mouvements de ses dipôles. L'agrégat supermoléculaire comprenant le soluté plus les molécules de solvant et les ions environnants sera donc simulé avec une constante diélectrique relative unitaire. C'est au champ de force "classique" (constitué uniquement des termes discutés précédemment) de décrire simultanément les interactions inter et intramoléculaires.

Des raffinements de certains termes du champ de force décrivant la molécule d'eau sont impératifs pour assurer de bons résultats. Par exemple, le traitement explicite de la polarisabilité électronique des molécules d'eau

⁸ K. Sharp, A. Jean-Charles and B. Honig, *J. Phys. Chem.* 96, 3822 (1992)

⁹ B. Lee, *Biopolymers*, 31, 993 (1991)

peut être inclus comme un terme supplémentaire dans un champ de force empirique¹⁰.

Nous avons déjà discuté les avantages et désavantages d'un traitement explicite du solvant par rapport aux modèles continus. D'un point de vue purement technique, l'approche explicite pose des problèmes de gestion du très grand nombre d'interactions qui varie en principe comme le carré du nombre des atomes considérés. Nous aurons à choisir entre des temps de simulation raisonnables (peu de molécules de solvant) et un réalisme physique indispensable (qui exige un grand nombre de ces molécules). Le caractère *fini* du système résultant est en contradiction avec la nature à longue portée des interactions électrostatiques. Ceci est discuté dans le paragraphe suivant.

2.i. *Rayon de troncature ("cut-off distance")*. Nous pouvons définir un rayon de troncature d_{cut} de façon à négliger les interactions interatomiques au delà de cette distance. Ceci revient à réduire drastiquement le nombre d'interactions à évaluer pendant le calcul de l'énergie. De plus, ce nombre d'interactions augmente proportionnellement au nombre N d'atomes de la molécule, au lieu de croître comme N^2 en l'absence de troncature.

Ce sont évidemment les interactions Coulombiennes en $1/r$ qui imposent la valeur de ce rayon, les interactions de van der Waals s'annulant beaucoup plus rapidement en fonction de la distance ($1/r^6$). Une valeur typique pour d_{cut} se situe entre 8...12 Å pour les simulations des macromolécules solvatées. L'énergie d'interaction entre deux charges unitaires situées à 12 Å est égale à 27 kcal/mol (!), ce qui pour deux charges fractionnaires atomiques typiques (0.1...0.5 unités électroniques) revient à 3...7 kcal/mol¹¹. Cette erreur en elle même ne serait pas importante si elle était *constante* pendant la simulation; ce n'est pas la valeur absolue de l'énergie moléculaire qui nous intéresse, mais ses variations. Cependant, si une paire d'atomes à $d < d_{cut}$, dont la contribution est bien prise en compte dans le calcul de l'énergie, s'éloigne à $d > d_{cut}$ pendant la simulation, cette contribution sera désormais ignorée, ce qui provoquera de grandes fluctuations de l'énergie.

Il y a deux façons d'éviter ce type de problème¹²:

¹⁰ D. van Belle, M. Froeyen, G. Lippens & S.J. Wodak, *Molecular Physics*, 77,239 (1992)

¹¹ D.H. Kitson, A.T. Hagler, *Biochemistry*, 27, 5246 (1988)

¹² C.L. Brooks, B. Montgomery Pettitt, M. Karplus, *J.Chem.Phys.* 83,5897 (1985)

2.i.a) ne pas "couper" les fragments moléculaires électriquement neutres. Pour chaque charge portée par un atome d'une molécule neutre, il doit y avoir une charge de signe opposé sur un groupe d'atomes avoisinants. L'exemple le plus évident sont les dipôles de liaison où deux atomes liés portent des charges égales et de signe opposé (+q,-q). L'interaction d'un troisième atome avec ce dipôle est forcément faible parce que +q et -q sont toujours très proches et donc les très fortes interactions individuelles avec l'autre charge se compenseront mutuellement en grande partie. L'intensité des interactions charge-dipôle ($1/r^2$) et dipôle-dipôle ($1/r^3$) diminue plus rapidement en fonction de la distance. Si on se situe dans le cas où la distance séparant cet atome de +q est inférieure à d_{cut} , tandis que celle vers -q dépasse cette valeur, le fait de négliger la dernière interaction provoquera une importante erreur parce que cette compensation ne se fait plus. Il est donc mieux de négliger les deux interactions. Autrement dit, une bien meilleure stratégie d'application d'un rayon de troncature consiste à *diviser* la molécule en un maximum de fragments neutres, vérifier si la distance entre deux fragments A et B est *inférieure* à d_{cut} et si c'est le cas inclure toutes les paires d'atomes ($a \in A, b \in B$) dans la liste des interactions à évaluer.

2.i.b.) modifier l'expression du potentiel Coulombien en le multipliant par une fonction de "terminaison" qui est proche de l'unité pour $d \ll d_{cut}$, modifiant donc peu l'énergétique des interactions proches, et qui décroît de manière continue pour devenir nulle à $d = d_{cut}$. De cette façon, les sauts brusques du potentiel à $d = d_{cut}$ sont remplacés par une variation continue, ce qui réduit largement les fluctuations¹³.

2.ii. *Conditions aux bords*: Ceci concerne la façon de construire des agrégats polymoléculaires en y ajoutant un nombre minimal de molécules de solvant autour du soluté afin d'obtenir une description réaliste de la couche de solvation. Vu le caractère à longue portée des interactions soluté-solvant et solvant-solvant, il est évident qu'une couche mono- ou bimoléculaire de solvant n'est pas suffisante pour entourer un soluté. L'épaisseur de cette couche doit excéder les dimensions du rayon de troncature. Le comportement des molécules d'eau en contact direct avec le soluté est très important et il est donc crucial d'assurer que non seulement le soluté, *mais également les premières couches de solvation*, soient entourés par un nombre suffisant de molécules d'eau pour assurer un comportement réaliste de *l'ensemble* des molécules de solvant. La

¹³ M. Prévost, D. van Belle, G. Lippens and S. Wodak, *Molecular Physics*, 71,587 (1990)

simulation doit reproduire la structure microscopique de la solution incluant les couches proches du soluté et un nombre suffisant de molécules de solvant au cœur de la solution.

Lorsque le problème du choix du nombre de molécules de solvant a été résolu, la "cellule" de simulation (c'est à dire le soluté entouré de solvant confiné dans un volume régulier - cubique, sphérique..) ainsi obtenue ne peut pas être placée simplement dans le vide. Pour éviter un comportement irréaliste des molécules de solvant situées à l'interface, il est nécessaire d'imposer des conditions spéciales aux bords de la cellule de simulation, ce qui peut se faire de deux façons suivantes:

2.ii.a) "geler" les molécules d'eau à l'interface, c'est à dire les considérer comme immobiles pendant la simulation. Les molécules intérieures ne peuvent pas pénétrer cette couche les séparant du vide. Cependant, une telle approche nécessite une correction artificielle pour remplacer les interactions ignorées de ce système fermé au reste du monde. Notamment, des impulsions stochastiques sont appliquées aux molécules mobiles à proximité de cette couche fixe pour simuler les chocs qui auraient normalement lieu avec les autres molécules du solvant. Cette procédure est communément appelée les conditions *stochastiques* aux bords (Stochastic Boundary Conditions)¹⁴

2.ii.b) générer des copies identiques de la cellule initiale de solvation (considérée cubique pour simplifier la discussion) dans les trois directions de l'espace et remplacer l'interface entre la cellule et le vide par une interface entre la cellule et une de ses copies. L'évolution de chaque cellule-copie va refléter le comportement de la cellule originale; ainsi, si une molécule de solvant quitte la cellule en traversant une face du cube, la copie de cette molécule (une molécule "fantôme") va rentrer simultanément dans la cellule par la face opposée. On applique des conditions *périodiques* aux bords (Periodic Boundary Conditions, PBC^{15/16}), et ceci peut se faire en plusieurs variantes:

¹⁴ C.L. Brooks, B. Montgomery Pettitt, M. Karplus, "Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics" , J. Wiley & Sons, (1988)

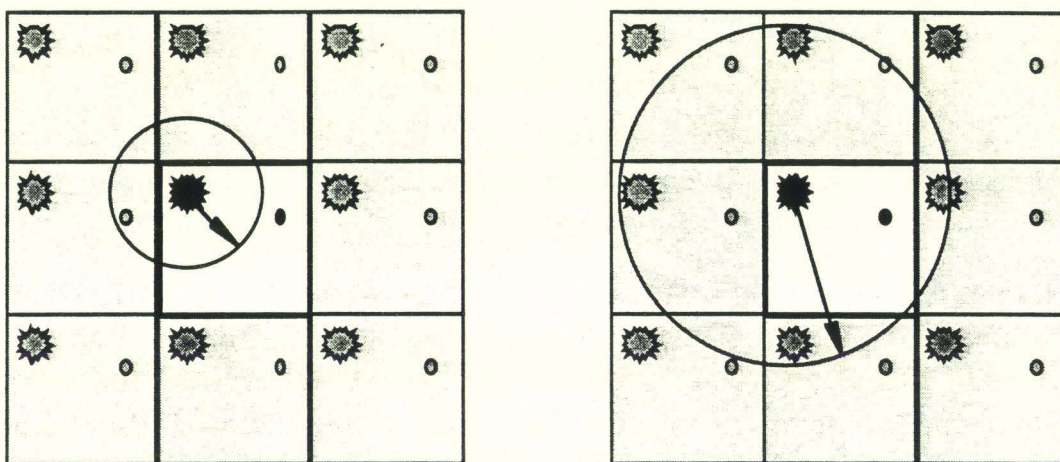
¹⁵ W.F. van Gunsteren & H.J.C. Berendsen, *J. Mol. Biol.*, 176, 559 (1984)

¹⁶ D.N. Theodorou and U.W. Sutter, , *J.Chem.Phys.* 82,995 (1985)

- prendre en compte l'interaction d'un atome i avec l'atome j et toutes ses copies $j'_1, j'_2 \dots j'_n$ qui se trouvent à moins de d_{cut} par rapport à i ("image explicite")

- considérer uniquement l'interaction de i avec une seule des images $j, j'_1, j'_2 \dots j'_n$ - celle qui se trouve à la distance minimale par rapport à i ("image minimale"). Ceci implique que d_{cut} est inférieure à la moitié de la dimension de la maille périodique utilisée.

- appliquer la méthode d'Ewald^{17,18} pour évaluer analytiquement les sommes des influences de l'infinité d'images périodiques du système. De nouvelles techniques^{19,20} utilisant cette méthode de base ont été développées récemment.



A.

B.

Fig. 1. Conditions aux bords: A. - l'image minimale: une molécule réelle (en noir, carée central) interagit au plus avec une seule des images des autres molécules. Le rayon de troncature doit être inférieur à 1/2 de la dimension de la maille périodique; B- l'image explicite: une molécule interagit avec toutes l'images (y compris les images « fantôme » de soi-même) situées à l'intérieur de la sphère définie par le rayon de troncature.

¹⁷ M.P. Tosi, *Solid State Physics*, 16,107 (1964)

¹⁸ N. Kasarawa, W.A. Goddard, *J. Phys. Chem.* 93,7320 (1989)

¹⁹ K.E. Schmidt, M.A. Lee, *J. Stat. Phys.*, 63, 1223 (1991)

²⁰ H.Q. Ding, N. Karasawa, W.A. Goddard, *J. Chem. Phys.*, 97, 4309 (1992)

Le problème essentiel des conditions périodiques vient du fait que pour la simulation des macromolécules, la dimension de la maille périodique ne peut pas être prise beaucoup plus grande que la taille du soluté. Ceci implique que le système des cellules périodiques ressemble plus à une organisation cristalline, où le soluté peut interagir avec ses images, ce qui n'est pas le cas des solutions diluées.

3. Simulation implicite de la présence du solvant. Modèles de constantes diélectriques effectives.

A l'échelle microscopique, la constante diélectrique n'est pas une grandeur bien définie. Cependant, l'idée de pondérer le terme Coulombien pour essayer de tenir compte de la présence du solvant de manière implicite et sans grands efforts de calcul est une approche courante dans les simulations de systèmes biologiques. Ceci représente en effet une première et très simple catégorie de modèles continus de solvation. Divers modèles de constante diélectrique dépendant de la distance ont été introduits dans les champs de forces²¹. Plus la distance entre deux charges est importante, plus le nombre d'atomes (de la molécule ou de solvant) pouvant venir s'intercaler va augmenter, et va de ce fait faire fortement chuter l'intensité des interactions électrostatiques. L'expression (2) du potentiel effectif d'interaction entre deux ions dans une atmosphère ionique idéale (une solution très diluée d'un composé ionique entièrement dissocié) s'écrit selon la théorie de Debye et Hueckel²²

$$w(r) = \frac{q_i q_j}{4\pi\epsilon_0} \cdot \frac{1}{\epsilon_{sol} \cdot e^{\kappa r}} \quad (2)$$

ce qui suggère que les effets de contre-ions peuvent être modélisés en utilisant une constante diélectrique effective croissant d'une manière exponentielle en fonction de la distance entre les ions, ou le coefficient exponentiel κ dépend de la *force ionique* de la solution.

²¹ E.L. Mehler & T.J. Solmajer, *Biochemistry* 23,3887,(1984); *Protein Eng.* 4,903,(1991); *ibid.* 4,911, (1991)

²² D.A. McQuarrie, "Statistical Mechanics" Harper Collins Publishers Inc., N.Y. (1976)

Ce n'est pas par hasard que la première partie de ce travail a été dédié entièrement à l'étude des problèmes de solvation. En effet, l'approche basée sur l'utilisation d'une constante diélectrique variable, bien que préférable à l'ignorance totale des effets de solvant, n'est pas conforme à la théorie générale de l'électrostatique, tandis que les solutions exactes des équations différentielles de l'électrostatique dans le cadre des modèles continus de solvant sont trop coûteuses en temps calcul. L'introduction des termes fournissant une description rapide et correcte des effets électrostatiques dans les champs de forces actuels représente une des directions modernes du développement de la modélisation moléculaire.

4. Le concept du champ de forces "consistant" (Consistent Force Field CFF)

On peut se demander si une description empirique du potentiel moléculaire est possible. Si oui, comment faire pour obtenir les coefficients qui interviennent dans les équations (1a)-(1e) (k , b_0 , A , B , charges atomiques fractionnaires...). Certains paramètres sont obtenus expérimentalement. La spectroscopie IR permet par exemple la détermination des constantes de force des liaisons et des angles de valence. Les valeurs "naturelles" des longueurs de liaisons et des angles de valence sont accessibles par radiocristallographie (diffraction aux rayons X). Les charges atomiques fractionnaires sont obtenues par calcul quantique sur de petites molécules modèles et ensuite transférées aux systèmes plus complexes. Par contre, les paramètres de van der Waals ou les paramètres de torsion sont plus difficilement accessibles. Le schéma général de résolution est le suivant: on estime grossièrement les coefficients et on les applique à une série de molécules de référence dont les structures et les propriétés²³ sont bien connues expérimentalement. On vérifie ensuite la concordance entre les propriétés calculées à partir du modèle moléculaire et leurs valeurs expérimentales. On modifie ensuite les coefficients de départ en utilisant une stratégie de minimisation de l'erreur. Après répétition de cette boucle d'optimisation, la paramétrisation devient ainsi "consistante". Il est toutefois hasardeux d'extrapoler ces résultats à des molécules présentant des

²³ ces propriétés peuvent être aussi bien des propriétés moléculaires (géométriques ou spectroscopiques) que macroscopiques (énergies de sublimation des cristaux, coefficients de transport ou enthalpies de fixation ligand-enzyme). Les premières sont utilisées lors de la calibration des termes de "valence" du champ de forces, tandis que les dernières servent à affiner les paramètres des interactions intermoléculaires (voir ref:)

groupements chimiquement différents de ceux des composés ayant servi à étalonner le champ de forces.

Les calculs quantiques ne peuvent être appliqués quant à eux qu'à des molécules de petite taille et présentant peu de degrés de liberté. En fait, il y a un passage graduel entre les méthodes quantiques "ab initio" exactes et les méthodes de calcul empiriques. Ces méthodes "hybrides"²⁴ introduisent certaines approximations au niveau de l'Hamiltonien quantique, en compensant les erreurs ainsi engendrées en optimisant des paramètres ajustables. En principe, si on utilise une paramétrisation empirique appropriée, on peut modéliser n'importe quelles interactions, même s'il s'agit de réactions chimiques²⁵. Un champ de forces est en quelque sorte une interpolation de la réalité et ses fonctions n'ont pas nécessairement de signification physique. Sa seule justification est qu'il permet de prédire les propriétés moléculaires avec une précision acceptable. Les champs de force représentent donc un système d'équations non-linéaires et paramétrisées de telle façon qu'ils soient capable d'évaluer "correctement" l'énergie en fonction de la géométrie moléculaire. Ces équations de champ de force s'inspirent d'une manière plus ou moins évidente des connaissances intuitives que nous avons sur le comportement des atomes, mais c'est uniquement par calibration qu'ils arrivent à fournir un modèle quantitatif de la réalité. *Le succès d'un champ de force n'est pas dû au réalisme des équations constituant, mais à sa capacité de reproduire avec un nombre limité de paramètres certaines propriétés pouvant être comparées avec des données expérimentales.* On peut penser que les autres approches de ce style, comme les réseaux neuronnaires²⁶, qui eux aussi sont par essence des modèles non-linéaires avec un nombre géant des paramètres ajustables, mais complètement dépourvu de sens physique, peuvent faire office de "champ de forces".

N'importe quel modèle non-linéaire pose un nombre de questions techniques difficiles. Il s'agit surtout de l'interdépendance et de la transférabilité des paramètres intervenants, ainsi que du problème de la redondance de la description empirique. A l'heure actuelle, il n'y a pas de règles strictes concernant le choix des fonctions ou le nombre minimal et

²⁴ pour des références, voir MOPAC User Manual, QCPE, Univ. of Indiana, Bloomington, Indiana, USA.

²⁵ A. Warshel, S. Russel, *J. Am. Chem. Soc.* 108, 6569 (1986)

²⁶ D.E. Rummelhart, J.L. McClelland, "Parallel Distributed Processing Exploration in Microstructure of Cognition" The MIT press, Cambridge, MA (1986)

suffisant des paramètres à utiliser dans un champ de force. Les champs de forces déjà existants ont certes des éléments en commun, mais différent beaucoup entre eux. S'il y avait une base physique ferme pour la méthode des champs de force empiriques, un seul champ de force serait suffisant. Il s'agit donc de modèles non-linéaires aux paramètres ajustables, mais cependant certains de ces paramètres sont attribués à partir de données expérimentales. Le processus d'optimisation du pouvoir prédictif d'un champ de force génère *l'interdépendance* de tous les paramètres (ajustables ou expérimentaux) du champ de force. En conséquence, l'exportabilité des paramètres vers d'autres champs de force n'est généralement pas possible.

De plus, l'addition d'un nouveau terme dans un champ de force (par exemple un potentiel de solvation) nécessite en principe la reparamétrisation complète de toutes les interactions. Notamment, la calibration des certains champs de forces classiques, ne comprenant pas un terme de solvation, est forcément faite en utilisant des propriétés moléculaires mesurées en *solution*, simplement parce que les données expérimentales relatives aux structures des molécules à l'état gazeux sont moins abondantes. L'étape de calibration oblige donc les paramètres du champ de force²⁷ à s'ajuster de telle façon que le résultat d'un calcul dans le vide ressemble le plus possible aux données expérimentales utilisées²⁸. Autrement dit, les paramètres du "vide" de certains champs de force tentent d'exprimer implicitement les effets de solvant, ce qui en soi est intéressant même si généralement ceci n'offre pas une description quantitative de la solvation. Cependant, si on ajoute un terme de solvation (des molécules d'eau explicites ou un potentiel continu). Ces paramètres ne sont plus appropriés parce que leur façon implicite d'émuler le solvant est maintenant redondante. Tout ceci pose pas mal de problèmes méthodologiques et comporte en principe deux solutions "extrêmes":

4.a) le fait d'avoir ajouté un nouveau type de potentiel détruit complètement la consistance des paramètres antérieurs, et donc il faut reprendre tout le travail de détermination des anciens ainsi que des nouveaux paramètres afin d'aboutir à un champ de force consistant. Ceci est

²⁷ K. Rasmussen, in "AIP Conference Proceedings - ECCC1 Computational Chemistry, Nancy/France, May 1994" Ed: F. Bernardi & J.-L. Rivail, AIP Press, Woodbury, NY (1995)

²⁸ une autre "dilemme" du même type est illustrée par la contradiction du fait que les géométries d'énergie potentielle minimale doivent être adoptées par les molécules à 0 K, pendant qu'on les assimile aux géométries expérimentales déterminées à la température ambiante pendant l'étape de calibration de la surface de potentiel.

la solution idéale en accord avec la philosophie du champ de force consistant. Cependant, un tel travail peut facilement prendre des années.

4.b) on accepte l'ancien champ de force tel qu'il est, avec ses qualités et ses défauts, et on optimise uniquement les nouveaux paramètres du terme ajouté. En maximisant la qualité générale du modèle, la paramétrisation obtenue pour le nouveau terme sera de nature à compenser les erreurs des termes importés de l'ancien champ de forces. Ceci est une procédure moins élégante, mais plus expéditive, et donc plus appliquée. Elle ne permet pas de trouver l'optimum global du modèle en fonction de ses paramètres, ce qui est en principe possible, mais pratiquement très difficile à obtenir dans le cadre de l'approche 4.a.

5. Mécanique Moléculaire.

Comme nous venons de le voir, le potentiel moléculaire se décompose sous la forme d'une somme de contributions énergétiques correspondant à différents paramètres structuraux (coordonnées internes) de la molécule:

$$V = \sum_{\text{liaisons}} V_b + \sum_{\text{angles}} V_\phi + \sum_{\text{torsions}} V_\theta + \sum_{\text{atomes non-liés}} V_{vdW} + V_{Coul} \quad (3)$$

Comme les coordonnées internes sont des fonctions des coordonnées cartésiennes atomiques, le potentiel moléculaire est une fonction implicite de $3N$ positions atomiques. Parmi ces $3N$ degrés de liberté, 6 (trois translationnels et trois rotationnels) sont redondants car le potentiel moléculaire est invariant pour un changement de référentiel.

$$V = V(r_{ik}, \phi_l, \theta_m) = V(x_{ij}) ; i, k = 1 \dots N; j = 1 \dots 3; l = 1 \dots N_{ang}; m = 1 \dots N_{tors} \quad (4)$$

Le but d'un calcul de mécanique moléculaire est de trouver les conformations d'énergie minimale de la molécule, ceci en fonction du champ de forces utilisé: c'est à dire qu'il s'agit de minimiser la fonction multivariable V par rapport aux différentes positions atomiques x . Cette optimisation convergera toujours vers le minimum le plus proche de la géométrie de départ (ou malheureusement vers le *point stationnaire* le plus proche). La plupart des algorithmes modernes d'optimisation utilisent les dérivées premières et secondes. Le développement de Taylor du potentiel

moléculaire par rapport aux positions atomiques s'écrit (en notation bra-ket):

$$V(x) = V(x_0) + \langle \nabla V | x-x_0 \rangle + \frac{1}{2} \left\langle x-x_0 \left| \left[\frac{\partial^2 V}{\partial x^2} \right] \right| x-x_0 \right\rangle \quad (5)$$

En principe, toutes les méthodes d'optimisation différentielles (c'est à dire qui font appel aux dérivées du potentiel) ont la même structure algorithmique²⁹:

5.a) on détermine la direction de descente au point courant x_0 , c'est à dire un vecteur unitaire u de l'espace à $3N$ dimensions, ceci en fonction des dérivées. Les méthodes diffèrent entre elles dans la manière d'évaluer le vecteur u . Ainsi, on peut prendre par exemple une direction de descente qui est celle du gradient mais en sens opposé, ce qui donne toujours la pente maximale parmi les différentes tangentes à l'hypersurface énergétique (*steepest descent* ou "méthode de la plus grande pente"). La méthode des gradients conjugués (*conjugate gradients*) emploie également uniquement les dérivées premières, mais elle tient compte en plus du gradient précédent pour générer des directions de descente mutuellement conjuguées. Les méthodes dites *pseudo-Newton* approximent la matrice Hessienne des dérivées secondes en partant du gradient uniquement (Davidon-Fletcher-Powell³⁰). Enfin, les méthodes du type *Newton-Raphson* utilisent la matrice Hessienne.

$$u(x_0) = u \left(\left[\frac{\partial V}{\partial x} \right]_{x=x_0}, \left[\frac{\partial^2 V}{\partial x^2} \right]_{x=x_0} \right) \quad (6)$$

5.b) on recherche le pas optimal λ d'avancement dans la direction déterminée. En imposant une direction u , le problème d'optimisation est réduit à un problème monodimensionnel:

$$? \lambda \Rightarrow V(x) = V(x_0 + \lambda u) =! \min. \quad (7)$$

5.c) après avoir trouvé le nouveau point, il faut vérifier s'il est assez proche d'un minimum (ceci n'est pas si évident). Si le nouveau point ne vérifie pas les critères de tolérance imposés, on passe à l'itération suivante.

²⁹ W.H.Press et al., "Numerical recipes, The Art of Scientific Computing", Cambridge University Press, Cambridge (1986)

³⁰ R.Fletcher, "Practical Methods of Optimisation", vol.1, John Wiley & Sons, New York (1980)

Le résultat de l'optimisation est fortement dépendant de la géométrie de départ, car la méthode ne permet pas d'identifier le minimum absolu d'énergie potentielle, mais seulement le premier minimum local rencontré. On ne peut donc pas franchir de barrières d'énergie séparant deux minima. Cependant, la mécanique moléculaire peut devenir un outil d'investigation conformationnelle efficace en introduisant des fonctions de contraintes qui forcent certains paramètres structuraux à prendre des valeurs imposées. Les exemples les plus simples sont les fonctions harmoniques du type:

$$V_c = k_c (q - q_0)^2 \quad (7)$$

où q peut être n'importe quelle coordonnée interne (le plus souvent un angle de torsion ou une distance interatomique). Si la constante de force k est assez grande, la somme du potentiel moléculaire et de ce potentiel de contrainte va atteindre le minimum imposé avec q le plus proche possible de q_0 . Le reste de la géométrie moléculaire va s'ajuster afin d'obtenir la plus petite valeur de l'énergie potentielle compatible avec la contrainte $q \approx q_0$. Il est évident que si on choisit correctement les coordonnées sur lesquelles les contraintes vont s'appliquer et si on exécute plusieurs minimisations pour toute une série de valeurs q_0 imposées, on pourra explorer différentes régions de l'espace conformationnel moléculaire. Malheureusement, pour de grandes molécules, la région explorée reste extrêmement limitée. Si on se limite à générer par exemple toutes les conformations intercalées d'une chaîne hydrocarbonée contenant n liaisons C-C, cela revient à faire 6^n calculs d'optimisation si on impose 6 valeurs différentes pour chaque angle.

6 Dynamique Moléculaire (Molecular Dynamics MD)

Le but de la dynamique moléculaire est d'étudier le mouvement des atomes dans le champ de forces moléculaire. En considérant que la dérivée du potentiel par rapport aux coordonnées est une force, on peut écrire l'équation de Newton pour chaque degré de liberté x_i :

$$F_i = m_k \cdot \ddot{x}_i = - \frac{\partial V}{\partial x_i} \quad (8)$$

m_k est la masse de l'atome k associé au degré de liberté i . On obtient ainsi un système de $3N$ équations (8), qui doit être résolu numériquement à partir d'une géométrie de départ (par exemple une conformation d'énergie

minimale obtenue par mécanique moléculaire) et des vitesses atomiques initiales³¹. La physique statistique relie le paramètre macroscopique qui définit l'agitation thermique moyenne à la température. Il est alors envisageable d'attribuer à chaque atome une vitesse initiale aléatoire tirée dans une *distribution de Maxwell à la température désirée*. La méthode de dynamique moléculaire consistera à intégrer³² simultanément en fonction du temps le système d'équations (8). La trajectoire de dynamique moléculaire sera une succession de petits déplacements des atomes dans la direction de la vitesse et sous l'influence de la force totale, en utilisant un pas d'intégration constant (typiquement de l'ordre de $1-2 \times 10^{-15}$ seconde pour les systèmes biologiques). La méthode de Monte Carlo^{33,34} (MC) utilise des déplacements aléatoires pour explorer l'hypersurface d'énergie potentielle et ne fournit pas d'informations chronologiques sur l'évolution du système.

6.i. L'exploration de l'espace conformationnel par Dynamique Moléculaire: Les atomes ayant une énergie cinétique et une énergie potentielle, ils peuvent en principe franchir des barrières d'énergie. C'est une des applications importantes de la dynamique moléculaire: la MD peut être employée comme méthode d'exploration de l'espace conformationnel. Les limitations sont aussi évidentes: l'échelle de temps des phénomènes accessibles par dynamique moléculaire est actuellement de l'ordre de la nanoseconde. Tout changement conformationnel se produisant à l'échelle de la nanoseconde et plus (c'est à dire la quasi totalité des phénomènes de repliement des macromolécules biologiques ou repliement des protéines "*protein folding*") reste actuellement inaccessible par simulation de dynamique moléculaire. Pour des raisons de stabilité numérique, le pas d'intégration ne peut pas dépasser quelques femtosecondes et doit rester inférieur à l'échelle de temps des vibrations les plus rapides rencontrées dans le système. Une technique couramment utilisée dans le but d'accéder à des pas d'intégration plus longs est le gel par l'introduction de contraintes géométriques^{35,36} des degrés de libertés associés aux fréquences de vibrations les plus hautes (par ex. la vibration des liaisons)

³¹ Si la structure de départ n'est pas en équilibre (les dérivées de l'équation 8 ne sont pas toutes nulles), il n'est pas nécessaire d'attribuer des vitesses initiales aux atomes.

³² L. Verlet, *Phys. Rev.*, 159,98 (1967)

³³ W.L. Jorgensen, C. Ravimohan, *J. Chem. Phys.* 79,926 (1983)

³⁴ M.Mezei, *Mol.Phys.* 47, 1307 (1982)

³⁵ J.P. Ryckaert, G. Ciccotti and H.J.C. Berendsen, *J. Comput; Phys.*, 23,327 (1977)

³⁶ W.F. van Gunsteren and H.J.C. Berendsen, *Molecular Physics*, 34,1311 (1977)

Une méthode efficace d'exploration conformationnelle consiste à diriger le système vers des minima de plus basse énergie en diminuant progressivement l'énergie cinétique des atomes (*simulated annealing*³⁷). Au début d'une telle simulation, cette énergie est très importante (elle correspond typiquement à 1000 K), ce qui permet au système de traverser la plupart des barrières de potentiel. Le "refroidissement" simulé fait que la molécule va descendre progressivement sur la surface d'énergie potentielle comme pendant une minimisation d'énergie, mais avec une probabilité beaucoup plus faible de se faire "piéger" dans un minimum local parce qu'elle possède toujours de l'énergie cinétique et garde sa capacité de traverser des barrières.

Des contraintes dirigées (*umbrella sampling*³⁸) peuvent être employées pour forcer une trajectoire MD à traverser une barrière, en choisissant q de l'équation (7) comme la coordonnée de la transition. C'est à dire que l'état de départ correspond à une certaine valeur q_A tandis qu'on cherche à joindre l'état q_B au delà d'une barrière d'énergie difficilement franchissable à température ambiante. Les états intermédiaires décrivent un "chemin" d'énergie minimale reliant le minimum de départ de celui d'arrivée, ce chemin étant d'une certaine manière imposé par la contrainte appliquée pour forcer la transition. En variant q_0 - la contrainte imposée - entre q_A et q_B , la coordonnée q sera confinée autour de la valeur courante de q_0 grâce au potentiel de contrainte, pendant que la dynamique moléculaire va permettre aux autres degrés de liberté d'explorer tout l'espace de phase accessible et de trouver les zones d'énergie minimale compatibles avec $q \approx q_0$. On peut ainsi dériver le potentiel de force moyenne (PMF) caractérisant la transition entre les deux états et même évaluer la constante de vitesse de cette transition³⁹. L'idée de base du "umbrella sampling", c'est-à-dire de modifier la surface d'énergie, connaît aussi d'autres applications⁴⁰

³⁷ D.H.J. Mackay, A.J. Cross, A.T. Hagler, in "Prediction of Protein Structure and the Principles of Protein Conformation" Ed. G. Fasman, Plenum Press, N.Y., (1989)

³⁸ G.H. Snyder, R. Rowan, S. Karplus & B.D. Sykes, *Biochemistry* 14, 3765 (1975)

³⁹ S.H. Northrup, M.R. Pear, C.Y. Lee, J.A. McCammon & M. Karplus, *Proc. Natl. Acad. Sci. USA.* 79, 4035 (1982)

⁴⁰ J. Kostrowicki & H.A. Scheraga, *J Phys. Chem.* 96,7442 (1992)

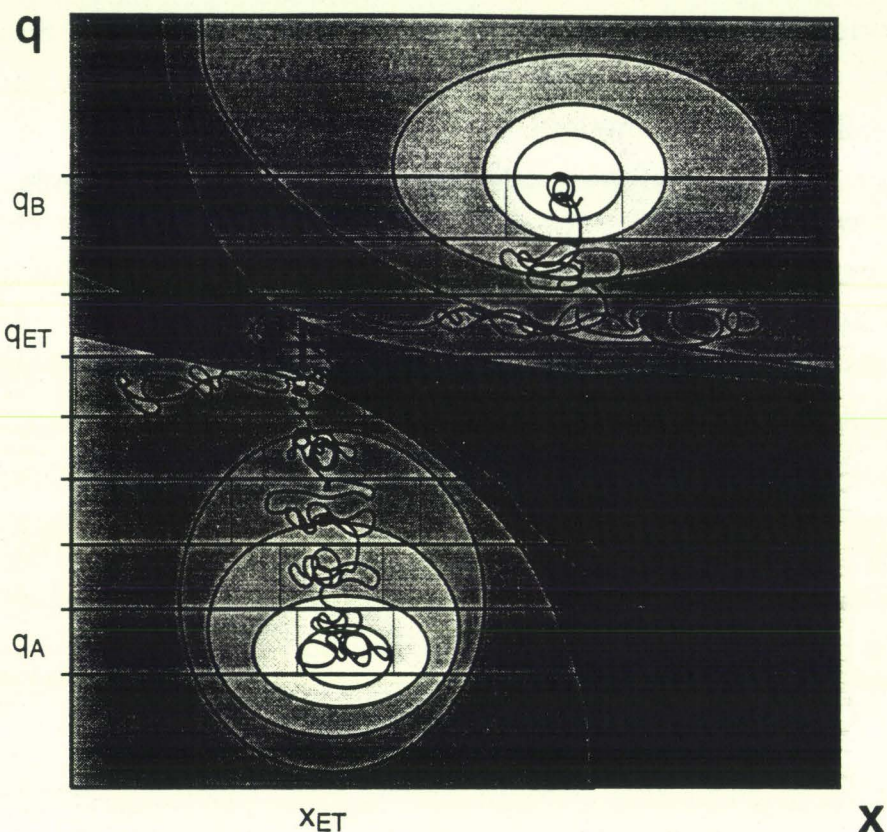


Fig. 2 - Un exemple de 'umbrella sampling' où la coordonnée de réaction q est modifiée de l'état initial q_A à l'état final q_B . X représente un autre degré de liberté qui est libre d'adopter des valeurs minimisant l'énergie potentielle à q donné. Le choix de la coordonnée q comme coordonnée réactionnelle est possible parce que l'état de transition ET (symbolisé par "+" sur le dessin) se trouve bien dans l'intervalle $q_A < q_{ET} < q_B$. Cependant, la voie réactionnelle définie par q n'est pas "orthogonale" à la barrière séparant A et B. Par conséquent, la zone autour de l'état de transition ne sera pas bien échantillonnée parce que d'autres domaines de l'espace conformationnel à $q \approx q_{ET}$ sont d'énergie plus basse et seront d'avantage visités.

La MD décrit un parcours aléatoire "random walk" dans l'espace conformationnel et peut revenir sur des points déjà visités. Ceci est en accord avec la réalité physique et signifie simplement que les conformations visitées plusieurs fois sont effectivement plus probables. Au contraire, l'utilisation d'un algorithme "self avoiding walk", qui évite de perdre du temps de calcul en rééchantillonnant des géométries déjà connues, serait plus judicieux pour l'exploration de l'espace conformationnel. De telles méthodes "à mémoire" qui tiennent compte des minima déjà visités sont

décrites dans la littérature comme étant beaucoup plus efficaces pour l'analyse conformationnelle des molécules^{41,42}.

6.ii. *Dynamique Moléculaire et Thermodynamique.* Un autre aspect très important de la dynamique moléculaire est de pouvoir relier par simulation les paramètres physiques moléculaires aux valeurs macroscopiques mesurables d'un système. Ces valeurs macroscopiques sont des moyennes calculées lors du balayage de l'espace conformationnel. On peut donc assimiler l'ensemble des géométries obtenues par une simulation MD à un ensemble statistique et utiliser les équations de la physique statistique pour obtenir les valeurs moyennes macroscopiques des paramètres recherchés. Le problème fondamental qui est très délicat concerne la convergence de la série de valeurs moléculaires individuelles vers la valeur macroscopique et surtout la signification concrète de cette valeur. Il n'est pas évident que l'ensemble des conformations adoptées par une molécule pendant la simulation MD soit une représentation correcte du comportement moyen d'une mole de molécules sur une échelle de temps à peu près 10^9 fois plus longue que la durée de la simulation.

Le température thermodynamique Θ , grandeur définie uniquement pour des systèmes à l'équilibre, est une mesure de l'énergie cinétique moyenne:

$$\overline{E_c} = \left\langle \sum_i^{N_f} \frac{p_i^2}{2.m} \right\rangle = \frac{N_f k \Theta}{2} \quad (9)$$

où N_f représente le nombre de degrés de liberté, k la constante de Boltzmann et p_i et m_i sont respectivement l'impulsion et la masse de l'atome i . La dynamique moléculaire nous offre au contraire une succession d'images instantanées du système et on peut donc définir par analogie une température momentanée de la simulation dont la moyenne doit égaler la température thermodynamique:

$$\frac{(3N - 6). kT}{2} = \sum_i^N \frac{m_i.v_i^2}{2} \quad (10)$$

tout en tenant compte que $N_f = 3N - 6$ parce qu'il y a conservation de l'impulsion et du moment angulaire total du système (= 6 degrés de liberté)⁴³.

⁴¹ G.M. Crippen & H.A. Scheraga, *Proc. Nat. Acad. Sci. USA*, 64,42 (1969)

⁴² T. Huber, A.E. Torda, W.F. van Gunsteren, *J. Comp.-Aided Mol. Design*, in press (1994)

Pendant l'intégration des équations (8), l'énergie totale du système est constante; on dit que l'ensemble est *microcanonique* (sans échange d'énergie ou de matière avec l'extérieur). Par contre, les fonctions thermodynamiques macroscopiques usuelles H et U (dont la dernière correspond à l'énergie potentielle estimée par le champ de forces à l'échelle moléculaire) peuvent être obtenues en partant d'un ensemble isotherme-isobare ou isotherme-isochoire. Dans ce cas on dit que l'ensemble est *canonique*.

Les équations de Newton seules ne permettent pas de contrôler la température: l'énergie totale étant constante, l'énergie cinétique E_C va fluctuer en fonction des variations de l'énergie potentielle E_P ($E=E_P+E_C=\text{constante}$). Pour que la simulation soit correcte, il faut répartir l'énergie cinétique de façon homogène au niveau des différents degrés de liberté, c'est pour cette raison que chaque simulation MD doit commencer par une étape d'équilibration. Le contrôle de la température est très souvent effectué par la méthode du couplage à un thermostat⁴⁴ (*thermal bath coupling*). Les équations de Newton sont corrigées à l'aide d'un terme de Langevin qui tient compte du transfert de chaleur entre l'extérieur et le système. On introduit la température momentanée T afin de quantifier la probabilité que les atomes reçoivent ou perdent de l'énergie cinétique au contact du thermostat. Si les atomes sont plus "froids" que le bain, ils vont recevoir de l'énergie, si ils sont plus "chauds" ils vont être "refroidis" et finalement si $T = T_0$ ils vont obéir aux équations de Newton.

$$F_i = m_k \cdot \ddot{x}_i = - \frac{\partial V}{\partial x_i} + m_k \gamma \dot{x}_i \left(\frac{T}{T_0} - 1 \right) \quad (11)$$

On peut utiliser des termes analogues pour maintenir la pression⁴⁵ constante. La constante de couplage γ donne l'intensité d'échange avec le bain de chaleur. Même si pratiquement les équations (11) sont beaucoup utilisées, il n'a pas été prouvé qu'elles génèrent véritablement un ensemble canonique, ce qui a par contre été démontré pour les méthodes à

⁴³ Ceci est valable dans le vide. Uniquement 3 degrés de liberté (de translation) doivent être soustraits lors de simulation utilisant les conditions périodiques aux bords, il n'y a plus conservation du moment angulaire total.

⁴⁴ H.J. Berendsen, J.P.M Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, *J.Chem.Phys.* 81, 3684 (1984)

⁴⁵ M.P. Allen & D.J. Tildesley, "*Computer Simulations of Liquids*", Clarendon Press, Oxford Science Publishers (1987)

température constante de Nosé⁴⁶ et des "collisions stochastiques" d'Andersen⁴⁷.

La fonction de partition Z , fonction de base de la physique statistique, s'écrit comme une somme sur tous les niveaux d'énergie quantiques k accessibles au système:

$$Z = \sum_n e^{-E_n / kT} \quad (12)$$

En mécanique classique, cette somme est remplacée par une intégrale sur tous les degrés de liberté. Si de plus l'énergie potentielle ne dépend pas des vitesses des particules, on peut découpler les degrés de liberté spatiaux et cinétiques et intégrer séparément sur les deux sous-espaces de l'espace de phase.

$$Z = \int e^{-\frac{V(x_1 \dots x_i) + K(p_1 \dots p_i)}{kT}} dx_1 \dots dp_i = \int e^{-\frac{V(x_1 \dots)}{kT}} dx_1 \dots \int e^{-\frac{K(p_1 \dots)}{kT}} dp_1 \dots = Z^* \cdot Z' \quad (13)$$

L'intégrale Z' peut se résoudre analytiquement, sachant que l'énergie cinétique est une somme des contributions $p^2/2m$ des énergies cinétiques par degré de liberté. La grandeur $\langle P \rangle$, pouvant être directement comparée à son équivalent macroscopique mesurable est simplement la moyenne de Boltzmann des valeurs microscopiques P_i associées aux états i :

$$\langle P \rangle = \frac{\sum_i P_i \cdot e^{-V_i / kT}}{\sum_i e^{-V_i / kT}} \quad (14)$$

ou encore dans l'équivalent classique:

$$\langle P \rangle = \frac{1}{Z} \int P(x_1 \dots p_i) \cdot e^{-H(x_1 \dots p_i) / kT} dx_1 \dots dp_i \quad (15)$$

Le résultat d'une simulation MD se concrétise par l'obtention d'une série de géométries de la molécule étudiée. Chaque géométrie j obtenue est complètement définie: l'énergie totale V_j , tous les composants de l'énergie potentielle sont connus, ainsi que les détails géométriques, les propriétés dipolaires... Pour une série de conformations moléculaires obtenues par

⁴⁶ Wm.G.Hoover, "Molecular dynamics", Springer Verlag, Berlin Heidelberg, (1986)

⁴⁷ H.C. Andersen, *J. Chem. Phys.*, 72,2384 (1980) ; *J. Comp. Physics*, 52,24 (1983)

dynamique, les probabilités d'échantillonnage reflètent implicitement les probabilités Boltzmanniennes des états. Les états de basse énergie potentielle seront représentés plus souvent parmi la série des V_j , et donc la valeur moyenne d'une propriété P se calcule comme la moyenne arithmétique des P_j obtenus à partir d'une telle simulation.

$$\langle P \rangle = \frac{1}{N} \sum_j^N P_j \quad (16)$$

Au contraire, lorsqu'on utilise une autre méthode d'échantillonnage (par ex. une recherche systématique des minima d'énergie par mécanique moléculaire) dont la probabilité de trouver un minimum j ne dépend plus de l'énergie potentielle V_j , l'équation (14) doit être utilisée, en comptant chaque niveau d'énergie une seule fois même s'il a été trouvé plusieurs fois par la méthode.

En réalité, il n'est pas possible d'obtenir un ensemble statistique décrivant de manière exhaustive une espèce macromoléculaire. La dynamique moléculaire est donc limitée à l'étude de certaines propriétés uniquement sur la base de certaines régions de l'espace conformationnel. En fait, l'espace conformationnel est divisé en différentes zones se situant autour des principaux minima et on se limite seulement à ces zones (chaque zone étant caractérisée d'une manière plus ou moins complète). Actuellement il est impossible (sauf pour des molécules très simples) de reproduire par calcul des valeurs caractéristiques d'un composé en tenant compte de tous les équilibres conformationnels possibles. On peut toutefois tirer certaines conclusions à partir d'études de conformères "individuels", tout en restant bien sûr conscient des limites du modèle.

Ceci nous amène à une brève discussion des aspects statistiques d'une simulation MD. L'équation (16) exprime le fait évident que les propriétés macroscopiques d'un système sont des moyennes pondérées des valeurs échantillonnées par dynamique moléculaire. La série des P_i est essentiellement aléatoire. L'estimateur⁴⁸ de la valeur moyenne d'une variable aléatoire est lui-même une variable aléatoire qui converge vers la moyenne exacte quand on augmente le nombre des points disponibles:

⁴⁸ B. Grais, "Méthodes statistiques", Dunod, Paris (1992)

$$\hat{x} = \frac{1}{n} \sum_{i=1}^n x_i \approx \bar{x} = \int_{-\infty}^{+\infty} x\rho(x)dx \quad (17)$$

On sait que la variance de l'estimateur de la moyenne est liée à la variance de la variable par la relation:

$$\sigma^2(\hat{x}) = \overline{(\hat{x} - \bar{x})^2} = \frac{\sigma^2(x)}{n} \quad (18)$$

C'est à dire qu'en partant de n valeurs connues pour la variable x , on peut estimer la valeur moyenne de cette variable avec une erreur qui est de l'ordre de grandeur de la dispersion σ de l'estimateur, qui à son tour est proportionnelle à la dispersion de la variable x et décroît comme $n^{-1/2}$. Concrètement, pour diminuer d'un facteur 10 l'erreur d'estimation sur une valeur moyenne, il faut utiliser 100 fois plus de points dans son calcul.

Des méthodes d'échantillonnage plus efficaces (*enhanced sampling*^{49,50}) introduisent des "copies" multiples de l'élément à analyser (ce qui peut être par exemple un ligand dans un site enzymatique ou une boucle bien localisée dans une protéine) par rapport à une seule représentation de l'environnement de cet élément (typiquement une grande protéine). Dans une simulation classique, on évalue à chaque pas les interactions entre l'élément "intéressant" et son environnement, mais aussi tous les interactions entre les atomes de l'environnement, ce dernier calcul étant beaucoup plus long. L'astuce consiste donc à introduire N copies multiples sans évaluer les interactions entre elles, de telle façon qu'ils ne perturbent pas réciproquement leur mouvement. Ainsi, chaque pas qui comporte un seul calcul sur tous les atomes de l'environnement, nous permet d'échantillonner N fois plus d'information concernant l'élément d'intérêt, ce qui serait équivalent à une réduction d'un facteur $1/N$ du temps de calcul. La limitation de l'approche est du au fait que l'environnement subis *l'influence moyennée de toutes les copies* et son comportement peut être modifié par rapport à celui en présence d'un copie unique.

En posant $P_i=E_i$, l'obtention de valeurs d'énergie (ou d'enthalpie) est assez aisée dans le cas où les simulations aboutissent à un bon échantillonnage de

⁴⁹ C. Simmerling & R. Elber, *J. Am. Chem. Soc.* 116, 2534 (1994)

⁵⁰ R. Elber & M. Karplus, *J. Am. Chem. Soc.* 112, 9161 (1990)

l'espace conformationnel. Néanmoins, ce calcul est connu pour sa faible convergence dans des systèmes complexes.

L'évaluation de l'entropie pose de grands problèmes de calcul. L'entropie d'une espèce moléculaire est une grandeur dépendant de l'hypervolume accessible dans l'espace de phase. C'est la raison pour laquelle son évaluation rigoureuse nécessite la connaissance de tous les équilibres conformationnels. On peut par contre (en principe) évaluer correctement l'enthalpie d'un composé à partir uniquement de quelques conformations stables. Les variations d'enthalpie pendant une réaction chimique sont en effet beaucoup plus importantes que les différences d'enthalpie conformationnelle. Ceci n'est plus vrai pour l'entropie: on peut évaluer l'entropie autour d'un minimum (l'hypervolume conformationnel autour du minimum) mais ceci ne dit rien sur la nature et le nombre des autres minima accessibles.

Karplus *et coll*⁵¹ ont proposé une méthode approximative pour évaluer l'entropie conformationnelle. En se plaçant dans une zone assez restreinte autour d'un minimum, on peut utiliser le développement de Taylor (5) du potentiel pour $x_0 = x_{\min}$. Le terme de premier ordre qui correspond au gradient en x_0 est nul (minimum d'énergie). L'expression du potentiel devient alors:

$$V(x) = V(x_0) + \frac{1}{2} \langle x-x_0 | H | x-x_0 \rangle \quad (19)$$

où H est la matrice Hessienne en x_0 . C'est une matrice carrée symétrique ayant un déterminant positif et qui peut être décomposée sous la forme:

$$H = \left[\frac{\partial^2 V}{\partial x^2} \right]_{x=x_0} = L^T \cdot L \quad (20)$$

où L est une matrice carrée à déterminer et L^T sa matrice transposée. En utilisant l'approximation continue de l'espace des phases, on peut écrire le terme dépendant uniquement des coordonnées spatiales de la fonction de partition Q^* de la manière suivante:

$$Q^* = \int e^{-V(x)/kT} dx \quad (21)$$

⁵¹ M.Karplus and J.N.Kushick, *Macromolecules* 14,325, (1981)

L'intégration doit se faire sur tout l'espace conformationnel. On rappelle que x est le vecteur position ($3N$ dimensions) et on symbolise par dx l'élément de volume d'espace conformationnel. La densité de probabilité de trouver la molécule en un point x de l'espace conformationnel devient alors:

$$P(x) = \frac{e^{-V(x)/kT}}{Q^*} \quad (22)$$

Enfin, l'expression de l'entropie en fonction de cette densité de probabilité est:

$$S = -k \int P(x) \cdot \ln P(x) dx \quad (23)$$

Pour un potentiel harmonique tel que (19), on peut calculer analytiquement la fonction de partition Q^* et l'entropie S en appliquant la décomposition (20) de la matrice Hessienne. Le résultat de ce calcul donne:

$$S = \frac{1}{2} N_f k + \frac{1}{2} k \cdot \ln[(2\pi)^{N_f} \cdot \det(\sigma_{ij})] \quad (24)$$

où:

$$\sigma_{ij} = \overline{(x_i - \bar{x}_i)(x_j - \bar{x}_j)} \quad (25)$$

représente la matrice de covariance du système (la moyenne est prise sur tout l'espace de phase et les indices i et j désignent les N_f degrés de liberté spatiaux). La matrice de covariance caractérise le degré de couplage existant entre les diverses coordonnées: lorsqu'elles sont indépendantes, S devient une matrice diagonale.

On peut utiliser la formule (24) pour évaluer l'entropie si on connaît les termes de la matrice de covariance. En supposant qu'une simulation de MD génère un ensemble statistiquement significatif de géométries, on peut appliquer (25) directement pour les coordonnées x_i ainsi obtenues.

Le calcul des fréquences propres de vibration⁵² ν_i d'une molécule se base aussi sur l'approximation harmonique (19). La connaissance de ces fréquences permet d'estimer l'énergie et l'entropie de vibration. Cette méthode diffère par rapport à la précédente par la façon d'analyser

⁵² E.B. Wilson, J.C. Decius, "Molecular Vibrations" Dover, N.Y. (1980)

l'hypersurface d'énergie autour du minimum considéré. Dans le premier cas, ceci se fait à l'aide de la dynamique moléculaire, dont la trajectoire peut en principe quitter la zone strictement harmonique autour d'un minimum et donc incorporer dans certaines limites des corrections dues à l'anharmonicité du potentiel. Au contraire, la méthode vibrationnelle fait uniquement usage des constantes de force (matrice Hessienne) évaluée au minimum d'énergie et n'explore pas directement les environs de celui-ci.

7. Méthodes de calcul des variations d'énergie libre.

En se basant sur la discussion précédente, qui nous a permis de relier les simulations MD aux grandeurs thermodynamiques, on va succinctement illustrer les techniques de simulation permettant d'estimer les variations d'énergie libre. Ces méthodes ont été discutées en détail dans la littérature⁵³. Considérons deux états d'un système, dont les énergies libres sont F_1 et F_2 respectivement. Ces états sont arbitraires comme par exemple le début et la fin d'une réaction chimique, la fixation d'un ligand dans un site, où même des états entre lesquels il n'y a pas de correspondance physique évidente. Nous devons uniquement trouver une coordonnée de transformation λ reliant les deux états et pouvant décrire aussi bien tous les états intermédiaires $F(\lambda)$. L'énergie libre étant une fonction d'état, peu importe si le chemin de cette transformation n'a pas de sens physique. Un exemple typique d'application^{54,55} d'une telle approche est le calcul des différences d'énergie libre de fixation entre deux ligands peu différents, en utilisant le cycle thermodynamique suivant:

⁵³ P. Kollman, *Chem. Rev.* 93, 2395 (1993)

⁵⁴ B.G. Rao, R.F. Tilton, U.C. Singh, *J. Am. Chem. Soc.* 114, 4447 (1992)

⁵⁵ S.B. Singh, Ajay, D.E. Wemmer, P.A. Kollman, *Proc. Natl. Acad. Sci. USA* 91, 7673 (1994)

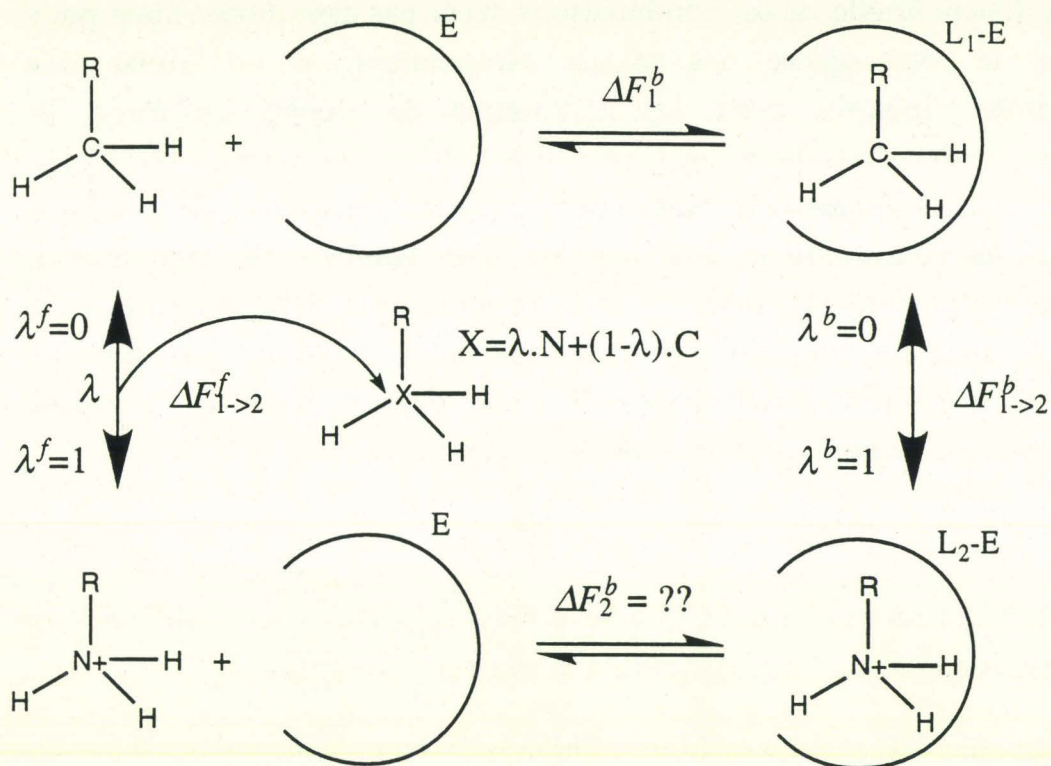


Fig. 3 - Le cycle thermodynamique illustrant le principe des calculs d'énergie libre de fixation d'un ligand par rapport à une molécule d'affinité connue.

ce qui permet de calculer l'énergie libre de fixation du ligand L₂ à partir de la valeur connue pour L₁ et les valeurs calculées pour les "transmutations" - processus complètement fictifs- de L₁ en L₂ en présence et en absence de l'enzyme:

$$\Delta F_2^b - \Delta F_1^b = \Delta F_{1 \rightarrow 2}^b - \Delta F_{1 \rightarrow 2}^f \quad (26)$$

Le paramètre λ quantifie cette "transmutation" entre un atome C vers N, ce qui veut dire qu'il contrôle les propriétés de l'atome "variable" X(λ) en les faisant varier d'une façon continue entre celles d'un carbone et celles d'un azote. Tous les paramètres ζ du champ de force décrivant les contributions énergétiques des éléments structuraux impliquant l'atome X (et en plus les masses atomiques) seront donc des combinaisons linéaires des valeurs bien définies des atomes C et N:

$$\zeta_{X(\lambda)} = \lambda.\zeta_N + (1-\lambda).\zeta_C \quad (27)$$

La forme fonctionnelle de ces combinaisons n'est pas essentielle, mais peut jouer sur la convergence des calculs. Notamment, si on utilise des combinaisons linéaires pour les paramètres de champ de force, le Hamiltonien des états intermédiaires utilisant ces paramètres dépendra d'une façon non-linéaire de λ . Néanmoins, pour les λ extrêmes de 0 et 1, on retrouvera les Hamiltoniens décrivant les deux ligands réels, ceci étant le seul aspect important. Toutefois, il faut s'assurer que dans le cadre d'un champ de force, le couplage (27) des coefficients n'aboutit pas sur des singularités mathématiques⁵⁶. Il est cependant toujours possible d'imposer le couplage directement au niveau des Hamiltoniens:

$$H(\lambda) = \lambda.H_1 + (1-\lambda)H_0 \quad (28)$$

Une fois la coordonnée λ ainsi définie, il existe plusieurs façons de calculer les variations d'énergie libre associées à la variation de celle ci:

7.i. *L'intégration thermodynamique (Thermodynamic Integration, TI)*. En accord avec la définition de la coordonnée λ , on peut écrire:

$$F_0 - F_1 = \int_0^1 \frac{\partial F}{\partial \lambda} . d\lambda \quad (29)$$

En sachant que:

$$F = -kT . \ln Z \quad (30)$$

l'équation (29) peut se reformuler:

$$F_0 - F_1 = -kT \int_0^1 \frac{1}{Z(\lambda)} \cdot \frac{\partial Z(\lambda)}{\partial \lambda} . d\lambda \quad (31)$$

L'équation (12) permet d'obtenir la dérivée de Z par rapport à λ , et l'équation (14) montre que le résultat obtenu n'est rien d'autre que la valeur moyenne de la dérivée du Hamiltonien par rapport à λ :

$$F_0 - F_1 = \int_0^1 \left\langle \frac{\partial H(p,x,\lambda)}{\partial \lambda} \right\rangle_{p,x} . d\lambda \quad (32)$$

⁵⁶ Discover User Manual, version 2.8, San Diego, Biosym Technologies, (1992)

La manière dont on a défini le couplage entre les états initial et final à l'aide du paramètre λ -équations (27) ou (28)- permet d'évaluer facilement cette dérivée. L'intégrale (32) se calcule numériquement en divisant l'intervalle $[0,1]$ en petits sous-intervalles $[\lambda_i, \lambda_{i+1}]$ et en faisant une simulation complète du système pour chacun de ces intervalles, à $\lambda = \lambda_i$ constant. Nous pouvons calculer la valeur moyenne à partir des valeurs échantillonnées pour la grandeur $\frac{\partial H}{\partial \lambda}$ pendant chaque simulation.

ii. *Méthode de perturbation (Perturbation Method, PM)*. Considérons deux états λ_1 et λ_2 de la transformation considérée. Il est évident que:

$$F_2 - F_1 = -kT \ln \frac{Z(\lambda_2)}{Z(\lambda_1)} = -kT \ln \int e^{-\frac{H(x,p,\lambda_2) - H(x,p,\lambda_1)}{kT}} \cdot \frac{e^{-\frac{H(x,p,\lambda_1)}{kT}}}{Z(\lambda_1)} dx \dots dp \quad (33)$$

ce que représente la valeur moyenne pour l'exponentielle de la différence d'énergie des deux états:

$$F_2 - F_1 = -kT \ln \left\langle e^{-\frac{H(x,p,\lambda_2) - H(x,p,\lambda_1)}{kT}} \right\rangle_{p,x,\lambda_1} \quad (34)$$

Il faut donc calculer cette valeur moyenne à partir d'une simulation utilisant l'Hamiltonien $H(\lambda_1)$, ce qui est très simple et ne demande pas d'évaluer des dérivées supplémentaires (32). Cependant, l'applicabilité de (34) est restreinte aux cas où les Hamiltoniens des deux états ne sont pas très différents. En effet, si les énergies $H(\lambda_2)$ évaluées à partir des points visités par la simulation "pilotee" par $H(\lambda_1)$ sont supérieurs à disons $H(\lambda_1) + 2kT$, le terme exponentiel à évaluer devient si petit qu'il disparaît dans le bruit de fond statistique. L'état 2 doit être assimilable à une petite "perturbation" de l'état 1, d'où le nom de la méthode.

iii. *Integration thermodynamique par la méthodes des différences finies (Finite Difference Thermodynamic Integration FDTI)*. En choisissant $\lambda_2 = \lambda_1 \pm \delta\lambda$ où $\delta\lambda$ est suffisamment petit pour pouvoir appliquer (34), on peut combiner les deux approches précédentes TI et PM en réécrivant (29):

$$F_0 - F_1 = \int_0^1 \frac{\delta F}{\delta \lambda} d\lambda \quad (35)$$

et en utilisant PM pour calculer δF à chaque pas d'intégration. Ceci peut être assimilé en quelque sorte à une approche TI utilisant des dérivées "numériques" - obtenues en réponse d'une petite variation effective de λ - tandis que l'équation (32) assume que l'influence du paramètre λ peut se traduire en terme des pentes calculées analytiquement. Un avantage clair de l'approche FDTI est une conséquence du fait que à l'aide d'une seule simulation à λ donné, on peut évaluer simultanément $F(\lambda+\delta\lambda)$ et $F(\lambda-\delta\lambda)$. Si $\delta\lambda$ est suffisamment petit, $\delta F/\delta\lambda$ doit être égal quel que soit le signe de la perturbation $\delta\lambda$; ceci peut servir comme test de convergence. FDTI semble converger plus vite par rapport aux PM ou TI⁵⁷. Comme pour n'importe quelle intégration numérique, différentes méthodes de quadrature sont applicable pour résoudre (35).

Il est intéressant de noter que les calculs d'énergie libre convergent généralement plus vite que les méthodes de calcul de l'enthalpie ou de l'entropie à partir de la fonction de partition (12). Par conséquent, il est envisageable d'obtenir ces valeurs à partir de deux calculs d'énergie libre à deux températures légèrement différentes⁵⁸, en sachant que:

$$\frac{\partial G}{\partial T} \approx \frac{G(T+\delta T) - G(T)}{\delta T} = -S \quad (36)$$

$$H = G + TS = G - T \frac{\partial G}{\partial T} \quad (37)$$

8. Les modèles mathématiques du "DOCKING" des substrats enzymatiques.

La reconnaissance d'un substrat par une enzyme et sa fixation ("*binding*") sont des processus qui reposent sur les interactions directes électrostatiques, van der Waals, transfert de charge entre les atomes du système enzyme-substrat-solvant, plus une série de contributions entropiques translationnelles, rotationnelles, vibrationnelles ou associées aux degrés de liberté des molécules de solvant. Une compréhension détaillée de tous ces effets⁵⁹, ou de leur importance relative, reste encore un but à atteindre - des études récentes viennent de mettre en évidence les rôles très importants des

⁵⁷ M. Mezei, D. Beveridge, *Ann. N.Y. Acad. Sci.*, 482,1 (1986)

⁵⁸ C.L. Brooks III, *J. Phys. Chem.*, 90,6680 (1986)

⁵⁹ J.Israelachvili, "*Intermolecular & Surface Forces*", Academic Press, (1992)

interactions entre cations et aromatiques⁶⁰ ou des aspects moins bien connus tels que l'effet "hydrophobe"⁶¹. Malgré les progrès^{62,63} récents dans le domaine de la modélisation moléculaire, il n'existe pas une approche généralisée pour traiter ce type de problème. Dû à la complexité des phénomènes, chaque cas en soi demande une analyse complète dans le but de choisir les approximations nécessaires dont l'applicabilité est strictement restreinte au système étudié. Néanmoins, de telles approches peuvent connaître des succès remarquables.

Le "binding" est quant à lui caractérisé par des valeurs thermodynamiques ΔH , ΔS , ΔG et par la constante d'équilibre correspondante (d'association ou de dissociation). En principe, le but des études de modélisation est d'obtenir ces valeurs par calcul, ce qui est loin d'être trivial. Par contre, on peut essayer d'expliquer l'affinité d'un substrat pour une enzyme en partant de paramètres accessibles par calcul. Ceci n'est *seulement* possible que dans des situations où pour toute une série de substrats du même enzyme, les effets non-évaluables sont constants. Ainsi, en théorie on devrait retrouver l'équation thermodynamique:

$$\Delta G = -RT \ln K \quad (38)$$

En pratique on est limité à trouver des équations linéaires empiriques:

$$\ln K = a_0 + a_1 \cdot X_1 + \dots + a_n \cdot X_n \quad (39)$$

où X_i sont des paramètres calculables et les coefficients a_i sont obtenus en étalonnant l'équation par régression linéaire. *Pour construire un modèle mathématique de "docking" enzyme-substrat il faut nécessairement disposer de valeurs de constantes d'affinité expérimentales pour toute une série de substrats analogues. Le but de ce modèle est de prédire les affinités pour d'autres analogues, en introduisant dans l'équation (39) ainsi établie les valeurs X obtenues par simulation.* Ces X calculables peuvent être en principe n'importe quels paramètres moléculaires *covariants* avec le ΔG d'association enzyme-substrat. Cette covariance peut avoir une base physique évidente (idéalement $X = \Delta G$) ou tout à fait obscure, mais ceci est un aspect secondaire si la relation est validée statistiquement et peut être

⁶⁰ D.A. Dougherty, *Science*, 271,163 (1995)

⁶¹ R.M. Jackson and M.J.E. Sternberg, *Prot. Eng.*, 7,371 (1994)

⁶² Ajay & M.A. Murcko, *J. Med. Chem.* 38, 4953 (1995)

⁶³ P.M. Dean, *"Molecular Foundations of Drug-Receptor Interaction"* Cambridge Univ. Press (1987)

utilisée pour des prédictions. Si l'approche possède un fondement physique, ceci peut aider à évaluer les coefficients de l'équation de régression sans pourtant pouvoir se soustraire à la règle d'une calibration préalable du modèle. Même si on prend X égal à l'enthalpie libre calculée à partir d'une méthode perturbative en conditions parfaites, il ne faut pas oublier que toute l'approche se base sur un champ de force *étalonné préalablement* et qui en plus nécessite très souvent⁶⁴ des réajustements des paramètres pour fournir des résultats en accord avec l'expérience. Finalement, des approches modernes faisant appel aux relations empiriques *non-linéaires* (réseaux neuronaux) sont en train de s'imposer progressivement⁶⁵.

On peut donc tenter de classier⁶⁶ les approches courantes par rapport à la nature des paramètres X utilisées pour expliquer l'affinité d'un ligand pour l'enzyme. Cependant, il n'y a pas de limite claire entre ces types de modèles:

7.i. Les modèles "physiques" essayent d'estimer les quantités thermodynamiques d'association enzyme-ligand ("docking") en faisant appel aux techniques présentes antérieurement. Il faut noter que les calculs de perturbation, en principe généralement applicables, sont beaucoup trop coûteuses en temps calcul et donc limitées à être utilisées uniquement dans le cadre des familles des ligands très similaires. Des évaluations moins fondamentales, mais plus rapides des paramètres d'affinité ont été proposées. Une large variété d'hypothèses simplificatrices sont introduites:

- l'utilisation des modèles continus de solvant pour décrire l'électrostatique et l'effet hydrophobe^{67,68}
- des termes de champ de force simplifiés ou ignorés ou l'utilisation des minima d'énergie obtenus par mécanique moléculaire^{69,70}
- le récepteur, le ligand ou les deux sont traités comme des corps rigides^{71,72}

⁶⁴ Y.Y. Shi, A.E. Mark, C.X. Wang, F. Huang, H. J. Berendsen, W.F. van Gunsteren, *Protein Eng.* 6,289 (1993)

⁶⁵ I.V. Tetko, A.I. Luik, G.I. Poda, *J. Med. Chem.* 36,811 (1993)

⁶⁶ T. Lengauer and M. Rarey, *Curr. Op. Struct. Biol.*, 6,402 (1996)

⁶⁷ B. Honig, K. Sharp, A.S. Yang, *J. Phys. Chem.* 97,1101 (1993)

⁶⁸ D. Sitkoff, K. Sharp, B. Honig, *J. Phys. Chem.* 98,1978 (1994)

⁶⁹ S.Krystek, T. Stouch, J. Novotny, *J. Mol. Biol.* 234,661 (1993)

⁷⁰ S. Vajda, Z. Weng, R. Rosenfeld, C. DeLisi, *Biochemistry.* 33,13977 (1994)

⁷¹ M.S. Searle, D.H. William, U. Gerhard, *J. Am. Chem. Soc.* 114, 10697 (1992)

⁷² J. Novotny, R.E. Bruccoleri, F.A. Saul, *Biochemistry.* 28,4735 (1989)

- de nombreux travaux considèrent les contributions de l'entropie de translation, rotation et vibration comme étant constantes.

Pour évaluer l'enthalpie du processus d'association, il faut calculer la différence:

$$E + S \rightleftharpoons ES \quad (40)$$
$$\Delta H_b = H_{ES} - H_S - H_E$$

où chaque terme peut être obtenu en partant d'une simulation MD (si les géométries les plus stables des molécules sont connues). Pour l'enzyme, cette géométrie doit évidemment reposer sur des données expérimentales⁷³. Selon la complexité du substrat, on peut se dispenser de données expérimentales. Le problème le plus délicat est posé par le complexe, où on rencontre 6 degrés de liberté supplémentaires: la position relative du substrat par rapport à l'enzyme. Même si on connaît la géométrie du ligand lié (et il faut souligner que celle ci peut être très différente de la géométrie préférentielle adoptée par le substrat libre), il n'est pas évident de placer correctement cette structure par rapport à l'enzyme. Pour un ligand flexible, il faudrait en principe essayer de placer toutes ses conformations possibles dans toutes les orientations relatives par rapport à l'enzyme et retenir les combinaisons d'énergie minimale. Une telle approche de "docking *ab initio*" est généralement très coûteuse en temps de calcul. Dans les cas simples, on peut deviner l'emplacement des substrats rigides dans les sites catalytiques enzymatiques, mais en principe il faut avoir au moins des informations expérimentales approximatives (par exemple, les structures d'enzymes homologues complexées à des substrats analogues.

Le calcul de l'entropie d'association pose encore plus des problèmes et se base en principe sur une estimation des probabilités de ce que les auteurs considèrent être les niveaux énergétiques caractérisant les états liés et libres.

Des nombreux travaux ont été dédiés à l'amélioration de l'échantillonnage des états conformationnels des ligands dans les sites enzymatiques. Une approche courante est de reconstruire des conformations des ligands dans le site enzymatique à partir d'une liste des fragments du ligand dont les conformations sont connues^{74/75}. En se basant sur l'hypothèse de l'additivité

⁷³ Protein Data Bank, Brookhaven National Laboratory

⁷⁴ W. Welch, J. Ruppert, A.N. Jain, *Chemistry & Biology*, 3,449 (1996)

⁷⁵ K. Gulukota, S. Vajda and C. Delisi, *J. Comp. Chem.*, 17, 418 (1996)

des contributions des fragments à l'affinité globale du ligand, les géométries des fragments maximisant l'interaction locale fragment-site seront préférées par rapport aux autres conformations. Cependant, il est évident qu'un fragment adoptera une géométrie complètement improbable si cela permet au reste du ligand de mieux « s'ancrer » au site de l'enzyme.

Une idée originale et intéressante est d'associer les degrés de liberté externes (translation et rotation par rapport à l'enzyme) et internes du ligand à des bains thermiques différents⁷⁶ -les premiers à haute température et les deuxièmes à température ambiante. Ainsi, le ligand peut parcourir le site à une vitesse très élevée, en évitant de se faire « piéger » par des optima locaux d'énergie d'interaction site-ligand. Cependant, les conformations peuplées du ligand correspondent à une température normale - le procédé profite donc d'une accélération de l'échantillonnage due aux températures élevées (1700 K), tout en évitant les distorsions géométriques des ligands dans de telles conditions.

Les algorithmes génétiques^{77,78,79}, reconnus comme des outils performants d'optimisation des fonctions multivariées, sont des approches prometteuses pour la recherche de la position et conformation optimales d'un ligand dans un site. Les coordonnées du système sont représentées sous une forme codée - le "chromosome". Ce chromosome (un vecteur C_i dont chaque "locus" i définit la valeur courante d'un degré de liberté correspondant i) contient toute l'information nécessaire pour caractériser complètement le système enzyme-ligand. De plus, cette information est mémorisée d'une manière séquentielle dans ce chromosome, exactement comme dans les systèmes biologiques. Il est donc possible d'engendrer un processus d'évolution "darwinienne" virtuelle à partir d'une collection de chromosomes aléatoires, en leur permettant de se "reproduire" (ce qui permet l'échange de l'information entre chromosomes par mécanisme *decrossing-over*) ou de subir des "mutations" ponctuelles (modifications aléatoires de l'information contenue dans un locus). L'étape de "sélection naturelle" consiste à favoriser les chromosomes décrivant des meilleures solutions d'ancrage du ligand dans le site et de détruire ceux qui codent des états dont l'énergie d'interaction site-ligand est trop élevée. En poursuivant

⁷⁶ A. Di Nola, D. Roccatano and H.J.C. Berendsen, *Proteins*, 19,174 (1994)

⁷⁷ C.M. Oshiro, I.D. Kuntz and J.S. Dixon, *J. Comp.-Aided Mol. Des.*, 9,113 (1995)

⁷⁸ A.R. Leach, *J. Mol. Biol.*, 235,345 (1994)

⁷⁹ G. Jones, P. Willett, R.C. Glen, *J. Mol. Biol.*, 245,43 (1995)

ce jeu évolutif pendant plusieurs générations, des "individus" plus adaptés (de meilleurs modes de fixation site-ligand) s'imposeront dans la population, en remplaçant les "concurrents" moins performants. "L'être parfaitement adapté à son milieu" (l'optimum global du problème) n'apparaîtra peut-être pas avant que les contraintes de temps calcul forceront l'arrêt de ce cycle évolutif virtuel. Cependant, la convergence⁸⁰ de l'approche est bien supérieure à celle des méthodes Monte-Carlo classiques.

7.ii. Les modèles de "pharmacophore" sont des approches très rapides essentiellement basées sur une description du type "clé-serrure" de l'interaction entre le ligand et son récepteur. Un "pharmacophore" caractérise l'orientation spatiale relative des groupes fonctionnels nécessaires pour qu'on puisse placer ce ligand dans le récepteur de telle façon que chacun de ces groupes se retrouve en position favorable pour participer aux interactions attractives avec des groupements chimiquement complémentaires⁸¹ du récepteur (donneur d'hydrogène - accepteur d'hydrogène => pont d'hydrogène; cation - anion => pont salin ...). L'affinité est donc mesurée en comptant ces interactions bien individualisées et pondérées en fonction de leur intensité relative. Pour définir un tel pharmacophore, il suffit de connaître un bon ligand d'un récepteur et d'adopter l'emplacement de ses groupements impliqués dans la fixation comme le "motif" nécessaire, cas où le score d'affinité s'évalue comme un degré de recouvrement entre les groupements actifs d'une molécule avec ceux du pharmacophore. L'activité d'une molécule est ainsi associée à la présence de ce motif dans sa structure. La simplicité des telles approches les rend applicables au "criblage virtuel"⁸² de très grandes banques de données moléculaires. Cependant, la recherche d'un pharmacophore dans une structure moléculaire très flexible peut devenir longue parce que chaque conformation de basse énergie doit être testée. Ces approches peuvent être raffinées pour inclure des pénalisations du score d'affinité en fonction de l'excès d'énergie intramoléculaire dans la conformation capable de mimer le pharmacophore. D'un autre côté, leurs limitations sont aussi évidentes:

- aucune prise en compte des effets entropiques.

⁸⁰ P. Willett, *Trends Biotechnol.*, 13,561 (1995)

⁸¹ R.S. Bohacek & Colin McMartin, *J. Am. Chem. Soc.* 116, 5560 (1994)

⁸² R.F. Burns, R.M.A. Simmons, J.J. Howbert, D.C. Waters, P.C. Threlkeld, B.D. Gitter, "Virtual Screening as a tool for evaluating chemical libraries", lecture at "Exploiting Molecular Diversity - Small Molecule Libraries For Drug Discovery", Jan. 23-25, (1995), La Jolla, California (USA).

- l'hypothèse implicite que deux molécules similaires se fixeront de la même façon sur un récepteur n'est pas toujours valable.

- la caractérisation "ponctuelle" des interactions site-ligand offre relativement peu d'information concernant le site de fixation: le modèle ne dit ni où *il ne faut pas* placer des groupements, ni où placer des *groupements supplémentaires* pour éventuellement augmenter l'affinité par rapport aux ligands connus. Ces problèmes sont adressés par des approches du type CoMFA^{83,84} ("*Comparative Molecular Field Analysis*") où l'analyse des positions relatives des groupements chimiques est remplacée par une comparaison des champs moléculaires entourants dans tous les points de l'espace.

7.iii. *Les modèles QSAR* ("*Quantitative Structure-Activity Relationships*"). Enfin, les propriétés physiques, chimiques ou biologiques des molécules peuvent être expliquées -voir équation (39)- en termes de différents descripteurs calculés à partir de la *structure moléculaire* où uniquement à partir de la *topologie moléculaire*. Les raisons physiques "cachées" par ces relations ne sont généralement pas facile à mettre en évidence et le choix des descripteurs à utiliser se fait d'une manière automatique, en laissant un programme d'analyse statistique détecter les grandeurs qui arrivent à expliquer les données expérimentales connues parmi une liste des indices de tout type. On va donc réserver le titre "QSAR" pour faire référence à ce type de modèles, malgré le fait que ce titre peut convenir à n'importe quelle autre approche de modélisation moléculaire dont le but est d'expliquer l'activité à partir de la structure. Typiquement, une telle approche implique les étapes suivantes:

1°) - on sélectionne un lot des molécules dont la propriété expérimentale à expliquer est connue.

2°) - on calcule autant de paramètres que possible à partir de la structure moléculaire. Il n'y a pas de restrictions quant aux descripteurs moléculaires utilisés dans des modèles QSAR, et en conséquence on peut utiliser aussi bien les indices topologiques calculés à partir du graphe moléculaire^{85,86,87},

⁸³ G. Klebe, U. Abraham, *J. Med. Chem.*, 36,70 (1993)

⁸⁴ ASP User's Guide, Issue 1, Oxford Molecular Ltd. (1993)

⁸⁵ A.T. Balaban, *Chem. Phys. Lett.*, 89,399 (1982)

⁸⁶ L.H. Hall & L.B. Kier, "*Reviews in Computational Chemistry*" Ed. Lipkowitz & Boyd (1992)

⁸⁷ M.J. Randic, *J. Am. Chem. Soc.*, 97, 6609 (1975)

les propriétés électrostatiques comme le moment dipolaire, les propriétés électroniques⁸⁸ ou stériques⁸⁹ des substituants où simplement la masse moléculaire ou la lipophilie.⁹⁰

3°) - on utilise un programme d'analyse statistique pour trouver les variables explicatives X_i à partir de cette liste et déterminer implicitement les coefficients a_i de l'équation (39) les reliant à la propriété expliquée⁹¹.

4°) - on calcule les mêmes descripteurs pour de nouvelles molécules et puis on applique cette équation pour prédire la propriété d'intérêt.

Notre publication "Trypanothione Reductase Inhibitors from a Novel Virtual Screening Approach" présente une adaptation originale de telles méthodes de calcul aux particularités de l'enzyme parasitaire Trypanothione Réductase, dans le but d'obtenir un modèle de "criblage virtuel" capable à reconnaître les inhibiteurs TR dans une base de données moléculaire.

A notre connaissance, il n'y a pas de méthodes utilisant la fonction de partition telle qu'on peut évaluer à partir d'une énumération des niveaux énergétiques obtenus par modélisation moléculaire. C'est sans doute parce que l'énumération *complète* de ces niveaux est un problème intraitable. L'idée centrale de notre travail est de calculer la fonction de partition non pas d'après les règles précises de la physique statistique, mais en utilisant une série des règles simples -certaines très courantes en modélisation et d'autres définies par nous-mêmes:

- un ensemble de règles d'échantillonnage conformationnel pour les ligands en maximisant la diversité⁹² des géométries échantillonnées

- le traitement rigide de l'enzyme et du ligand pendant le docking, en faisant néanmoins rentrer chacune des conformations plausibles du ligand dans le site. Par conséquent, les contributions de translation, rotation et vibration ne sont pas prises en compte dans le calcul de Z.

⁸⁸ H. Wiener, *J. Am. Chem. Soc.*, 69,2636 (1947)

⁸⁹ A. Verloop, W. Hoogenstraaten, J. Tipker, "*Drug Design*", Ed. E.J.Ariens, vol VII, 165 (1976)

⁹⁰ C. Hansch & A. Leo, "*Substituent Constants for Correlation Analysis in Chemistry and Biology*" John Wiley & Sons (1979)

⁹¹ Tsar User's Guide, Issue 3, Oxford Molecular Ltd. (1993)

⁹² P.M. Dean (editor) "*Molecular Similarity in Drug Design*", Blackie Academic & Professional, Glasgow (1995)

- l'utilisation de points de départ multiples pour l'optimisation de la position relative du ligand par rapport à l'enzyme.

- l'introduction de nouveaux termes de champ de force pour tenir compte des effets non pris en compte dans les champs de forces classiques et notamment l'adaptation d'un modèle de solvation continu très simple. Ces termes ont été choisis pour être compatible avec une approche utilisant un réseau tridimensionnel de points équidistants ("*grid-based docking approach*") pour décrire le potentiel produit par les atomes de l'enzyme dans le site actif⁹³.

- les niveaux d'énergie correspondent aux minima locaux et sont donc obtenus par de simples calculs de mécanique moléculaire.

Dans ce cadre, il devient facile d'énumérer les niveaux d'énergie, mais que signifie une "fonction de partition" ainsi calculée ? Deux possibilités sont envisageables:

1°) cette valeur n'a pas de signification, ni d'utilité. Ceci nous obligerait à modifier les hypothèses de départ.

2°) cette valeur n'est pas reliée à la valeur théorique de la fonction de partition. Néanmoins, en prenant la différence entre les "enthalpies libres" calculées à partir de ces valeurs à l'état libre et à l'état lié du ligand, on obtient une quantité qui *se corrèle* avec les affinités mesurées des ligands, tout en étant différente de l'enthalpie libre de fixation. Ceci implique donc une étape de *calibration* consistant à trouver la relation de régression optimale reliant les descripteurs calculés à la variable expliquée (l'affinité), suivi par une étape de *validation statistique*. Si cette corrélation est statistiquement validée, l'approche peut être appliquée pour prédire les affinités des nouveaux ligands, même si le sens physique de ce qu'on appelle la "fonction de partition" dans notre modèle n'est pas évident. Un tel cas apparaîtra si les phénomènes ignorés dans notre modèle sont soit *constants pour toute la série des ligands*, soit *covariants* avec les autres termes pris en compte et peuvent donc être représentés de manière implicite.

⁹³ B.A. Luty, Z.R. Wasserman, P. F.W. Stouten, M. Zacharias and J.A. McCammon, *J. Comp. Chem.*, 16,454 (1995)

On peut donc définir notre approche comme "un modèle 3D-QSAR basé sur des descripteurs s'inspirant de la physique statistique du docking". Une telle définition est néanmoins applicable à n'importe quel modèle moléculaire, où l'empirisme et l'application stricte des lois de la physique se rencontrent toujours. Dans notre publication, on utilise des terms tels que "descripteur enthalpique et entropique" pour souligner le fait que leur définitions sont empruntées à la physique statistique *sans pourtant remplir les conditions strictes de l'applicabilité (l'existence des ensembles canoniques)* de ces équations dans le contexte local. Il est cependant intéressant de savoir si ces descripteurs sont apparentés aux enthalpies et entropies réelles malgré les simplifications de base de leur définition, ou si au contraire, ce sont des descripteurs dont la combinaison linéaire explique l'affinité des ligands sans pourtant pouvoir les mettre séparément en relation avec les grandeurs thermodynamiques dont leur définition fait référence. En absence de valeurs expérimentales pour ces grandeurs, une réponse définitive à cette question n'a pas pu être donnée, mais nous avons mis en évidence certains indices montrant que les descripteurs calculés peuvent être en effet des estimateurs pour ceux-ci.

Grâce aux hypothèses simplificatrices choisies, il s'est avéré possible d'utiliser une approche de docking "ab initio" pour une étude de criblage virtuel, technique traditionnellement basée sur des évaluations très rapides, mais peu réalistes, des "scores" d'affinité. Il est clair que notre approche est beaucoup moins rapide que les méthodes classiques du criblage virtuel. Elle peut s'appliquer à des bases de milliers de molécules, mais non pas à des centaines de milliers. On peut affirmer qu'elle est en compensation plus réaliste et plus fiable.

A Virtual Screening Approach Applied to the Search of Trypanothione

Reductase Inhibitors

Dragos Horvath^{a,b,c}

*a. Institut Pasteur de Lille/SCBM, URA CNRS 1309, 1, rue Calmette, 59019
Lille Cedex, France. Tel: (33)03.20.87.73.65; Fax: (33)03.20.87.73.77; E_mail:
dragos@calmette.pasteur-lille.fr*

*b. Unité de Conformation des Macromolécules Biologiques, Université Libre
de Bruxelles, CP 160/16, P2, Avenue P. Heger, B-1050 Bruxelles, Belgium*

*c. University « Babes-Bolyai », Cluj-Napoca, Dept. of Organic Chemistry, str.
Arany Janos nr. 11, 3400 Cluj-Napoca, Romania*

ABSTRACT:

A "virtual screening" of a data base of 2500 2D molecular sketches by an original prediction algorithm of the inhibitory potency of Trypanothione Reductase (TR), the flavoprotein replacing Glutathione Reductase (GR) in the metabolism of Trypanosomatidae, has detected several structures of putative TR ligands. Their inhibitory potency has been experimentally tested, with a high rate of success. The fully automated prediction algorithm converts the 2D molecular sketches into 3D ligand structures, explores the conformational space of the latter and finally performs a grid-based, rigid-body docking of the resulting family of ligand conformations into the TR site, calculating enthalpic and entropic binding indexes. The values of the fittable parameters that occur in the equations used to predict the inhibitory potencies have been calibrated on a learning set of TR inhibitors of known inhibition constants. Moreover, the docking model has been used to obtain hints about the binding modes of TR ligands and the nature of site-ligand interactions.

Keywords: Trypanothione Reductase inhibitors, site-ligand interactions, automated docking, conformational sampling, free energy calculations, affinity predictions, QSAR, continuum solvent models.

I.Introduction:

The enzyme Trypanothione Reductase (TR) appears to be one of the most promising targets for trypanocidal drugs^{1,2}. TR, the parasitic homologue of human Glutathione Reductase (GR), is involved in the regulation of oxidative stress in the parasite cells. Both TR and GR are homodimeric FAD-dependent reductases that catalyze the reduction of a disulfide bridge -S-S- to its dithiolic form (-SH)₂ at the cost of one NADPH/H⁺ cofactor molecule³. It is generally considered^{4,5,6,7,8,9,10} that the mutual selectivity of these homologous enzymes towards their own substrates (oxidized trypanothione TS₂ and respectively oxidized glutathione GSSG) is based on differences in both electrostatic and hydrophobic properties of the binding sites. Several basic residues in the GR site are replaced by acidic or hydrophobic residues in the TR¹⁰.

These differences have been exploited to design specific inhibitors for the parasitic TR, not interfering with the metabolism of the host. Known TR inhibitors can be mimics of the natural substrate¹¹ or completely different molecules, most of which show a common pattern of a bulky aromatic or hydrophobic moiety linked to a (poly)amino chain by means of a hydrophobic spacer¹².

Molecular modeling is a valuable tool in drug design, its applications ranging from offering a qualitative understanding of the ligand binding mechanisms, to semiquantitative or quantitative predictions of binding free energies.

The fast pharmacophore approaches may not offer very precise affinity predictions, but their low computational cost makes them suitable for "virtual screening" of molecular data bases^{13,14,15}. In the pharmacophore philosophy, good ligands are molecules that have appropriately oriented functional groups such as to bind to a maximum of "anchoring points" like hydrogen bond (HB) donors, HB acceptors, hydrophobic groups, to the enzyme wall. The binding score is actually a count of the number of groups of the ligand that match this relative disposal, weighted by the assumed importance of each interaction. De novo ligand design techniques search for fragments that might be used to interconnect some functional groups placed in the vicinity of some key spots of the site^{16,17,18,19}. Quantitative Structure-Activity Relationships (QSAR)^{20,21,22,23,24,25,26} and refs. therein relating the activities of a series of known ligands to some calculable molecular properties can be used to circumvent the absence of structural data for the binding site.

At the opposite, binding can be studied by simulating the complete system of the enzyme + ligand + solvent + counterions using an appropriate Molecular Mechanics (MM) force field at a given temperature and for a sufficiently long time period in order to make sure that the simulated trajectory the system has visited all the major energy minima^{27,28}. When this is the case, statistical mechanics²⁹ can be used to calculate any measurable thermodynamic property of the system. Free energy calculations by thermodynamic integration^{30,31} or perturbation techniques^{30,32} have been used to determine the relative binding energy differences upon a slight change in the structure of the ligand or a point mutation in the active site. Since these approaches can be extremely time-consuming, their main goal is

not the large-scale prediction of binding affinities, but the deeper understanding of the binding processes.

An alternative approach makes use of a *grid description*^{33,34} of the different potentials within the enzymatic site. These potentials are evaluated only in the points of a grid spanned over the region containing the site of the enzyme. For any other point of space, they are obtained by an interpolation starting from the exact values in the nearest neighboring grid points.

In this work, we report an original computational approach to predict the TR inhibitory potency of molecules in order to detect new putative trypanocidal drugs by virtual screening of molecular data bases. While low-cost pharmacophore models are the method of choice employed for such purposes, the TR system presents a certain number of peculiarities that require a more elaborate approach.

The current "virtual docking" approach consists in defining a set of simplified "virtual physical laws" supposed to govern both the behavior of the free and of the bound ligands, so that a "virtual enthalpy" and "virtual entropy" of binding can be evaluated. A compromise must be reached between the *computational effort* and the *degree of realism of this set of rules* in order to make the method suitable for screening a large number of putative ligands. The calculated binding indexes should nevertheless provide information about the binding of the ligand, even if they do not necessarily correspond to the real enthalpy and entropy. In this way, they could be used as explaining variables of the real affinity as in a typical QSAR

approach. The "degree of realism" of the simulation can be optimized with respect to some fittable parameters.

The flexibility of most of the TR inhibitors (see previously cited refs. on TR) makes use of conformational analysis techniques mandatory. For the same reason, the entropic effects at binding are expected to be important.

The known crystal structures of both complexes of the TR¹¹ and GR^{7,8} active sites show a lot of weak and water-mediated site-substrate interactions, but no outstandingly strong anchoring points that concentrate most of the binding affinity. The fact that there are few changes³⁵ in the positions of the side chains of the free¹⁰ and complexed¹¹ TR sites suggests that a rigid site model might be a valid working hypothesis.

Accordingly, a rigid-body docking procedure based on a grid description of the site-substrate interactions has been preferred to the other molecular modeling techniques. We use electrostatic, van der Waals, hydrophobic and desolvation terms to describe the interactions between the rigid site and the ligands.

We account for the ligand flexibility without losing the simplicity of the rigid-body docking approach, by subjecting a *relevant family of rigid conformations* of the ligand to the docking procedure³⁶. The speed achieved by decoupling the internal and intermolecular degrees of freedom is essential for the current approach.

The adjustable parameters appearing in the expressions of the docking energy were evaluated by fitting the calculated to the experimental affinities

for a set of ligands of known affinities^{12,35,37,38,39,40,41,42,43}. As a first test, the model has been used to predict the binding affinities of a second series of such ligands (these molecules are shown in Fig. 1 and Table 1). Finally, it has been applied to search for new possible inhibitors among various molecular structures from different data bases. Molecules with a favorable predicted binding free energy or enthalpy have been subjected to affinity testing and their inhibitory potency was measured.

II. Methods:

1. Automated generation of 3D structures of ligands.

i. General overview of existing methods: The prediction of the three dimensional structures of small molecules has found a variety of solutions, concretized in a series of algorithms of different complexities^{44,45}.

The detection of all the minimum energy conformations of a molecule is a problem growing exponentially with respect to the internal torsional degrees of freedom. Knowledge-based rules can be used to restrain the considered values of a torsional angle in a specific chemical context⁴⁶. A rigid-rotor approach ignores the relaxation of the strain energy on behalf of the other intramolecular degrees of freedom. The current torsion angle driving procedure available in commercial molecular modeling packages⁴⁷ add torsional constraint terms to the molecular Hamiltonian. Molecular Dynamics (MD) or Monte Carlo (MC) simulations can be used to sample the conformational space, without offering the guarantee that all relevant minima have been visited. Biased MM « poling » techniques, which enhance crossing the energy barriers show an improved sampling efficiency⁴⁸.

ii. The basic working hypotheses of our conformational search ("Sampling Axioms"). The conformational search strategy reported here is based on the well known fact that the energy minimization of any molecular geometry

will converge towards the nearest local minimum on the potential energy surface. A very large number of crude geometries is generated by systematic rotations around the torsional axes of the molecule, out of which only a subset of representative geometries, each converging towards a different minimum need to be filtered out. While it is difficult to predict which of the input geometries are redundant, we have based our selection on the following three empirical rules, which will be furtheron referred as the "sampling axioms":

- a). Geometries with overlapping nonbonded atoms will be discarded.
- b). The more similar two starting geometries are, the higher the probability that they will converge towards the same energy minimum. We look for a set of starting geometries of *maximal diversity*.
- c). Between two different starting geometries, the one with the largest minimal distance between two nonbonded atoms will be preferred. This is applied in order to counteract the tendency of the following minimization step in vacuum to lead to "compact" structures that do not necessarily represent the populated conformations in solution.

The main advantage of this approach is that no time-consuming energy calculations are done at the step of "combinatorial explosion" when all possible rotamers of a molecule are generated. This huge number of conformations are merely subjected to filtering according to the three conditions, leading to the selection of an imposed number of typically 300-500 most diverse starting conformations out of several million possibilities

in times of minutes. According to the sampling axioms, a single enantiomer will be stored for each chiral conformation. Since enantiomeric conformations are no longer equivalent in the chiral site environment, the discarded enantiomer will be recalculated by the docking algorithm by applying the corresponding symmetry operation on the coordinates of the stored geometries. A detailed description of the building and conformational search algorithm is given in Appendix A.

iv. Energy minimization of the selected conformations. The geometries selected in this way are subjected to energy minimization using the Discover/Biosym program and the CVFF^{49,50} force field with a distance-dependent dielectric constant of $2r$. A steepest descent followed by pseudo-Newton minimization⁵¹ is interrupted when the maximum derivative falls below 0.1 kcal/Å. Minimized conformations for which both the total potential energies and Coulomb energies differ by less than 0.05 kcal are considered to be identical and are kept only once.

The remaining geometries are then sorted in increasing order of the energy values and those having more than 5 kcal/mol of excess strain energy with respect to the best found minimum are discarded.

This window of allowed intramolecular strain with respect to the best minimum must be large enough to encompass conformations that are not populated in the free state, but might nevertheless be favored upon binding. For reasons of computational expense, only the 30 most stable conformations are considered for docking if the number of sampled geometries within 5 kcal/mol from the best minimum exceeds this number.

iv Modeling of the protolytic equilibria in ligands: The effective charge distribution of polyamines, the class to which many of the known TR inhibitors belong, depends on the protolytic equilibria⁵² in which these are involved. The classical molecular modeling tools require a clear-cut choice of the protonation state to be made prior to the simulations. Ideally, all possible protonation states of a ligand should be modeled explicitly, including the ones that are less probable to exist in solution, but might nevertheless be the main binding species. This is certainly too time-consuming. Therefore, we have designed a way to account for an averaged effect of the protolytic equilibria on the docking energies.

In an empirical way, we can express the fact that an ammonium group is in equilibrium with the uncharged amine by weighting down its charge $Q=+1$ by the probability of finding it in a protonated state at $pH=7$: $Q'=p.Q$. We define the free energy governing the protolytic equilibria in function of the probabilities of protonation p_i :

$$G_{prot} = -\sum_i h_i p_i + K \sum_{i \neq j} \frac{p_i p_j}{d_{ij}} \quad (1)$$

where h_i is the specific protonation enthalpy of the group i , d_{ij} is the distance between the groups and the effective charge of the group i is $Q.p_i$. The first term ($-h_i p_i$) in equation (1) expresses the fact that isolated basic amino groups tend to accept a proton due to a favorable proton binding enthalpy $-h_i$. Accordingly, the minimum of G should be at $p_i=1$. If the proton binding affinity were the only driving force of proton binding, all amino groups should be completely protonated in solution. In reality, the

mixing entropy terms will oppose this tendency, always ensuring an equilibrium between protonated and non-protonated groups ($p_i < 1$). However, isolated aliphatic amino groups at pH=7 are *almost* completely protonated, so that we will simply ignore the contributions of the mixing entropy terms. Mono- and diaromatic amines were considered to appear exclusively under unprotonated form at pH=7.

Furthermore, if there are several amino groups in the same molecule, the protonation of one will impede on the proton affinity of the second, since the simultaneous presence of both positive charges at a distance d gives rise to an unfavorable Coulomb interaction of magnitude $+e^2/d$ (where e is the electronic charge unit). The fraction of molecules with both groups i and j in a protonated state is $p_i p_j$, and accordingly, the second term in the sum of equation (1) stands for the overall contribution of the intramolecular proton-proton Coulomb repulsion in polyprotonated species.

Practically, we chose to introduce a common value for all the h_i , without accounting for the chemical environment, so that $\Pi = h/K$ becomes the overall *fittable* parameter that controls the protonation state of the ligands. Imposing a minimal G_{prot} with respect to these p_i :

$$\frac{1}{K} \frac{\partial G_{prot}}{\partial p_k} = -\Pi + \sum_{i \neq k} \frac{p_i}{d_{ik}} = 0 \quad (2)$$

we obtain system of equations which is solved to obtain the p_i . The main advantage of equations (2), e.g. their linear character, constitutes on the other hand the source of artifacts occurring in the extreme cases of either

very high or very low protonation probabilities, when the ignored mixing entropy terms in $\log p_i$ and $\log(1-p_i)$ would actually control the free energy of the system. Therefore, calculated p values that may fall outside the range of (0,1) need to be interpreted as either total deprotonation or total protonation of the amino group. Considering a piperazine molecule that is mainly monoprotonated and to some extent diprotonated at pH 7, with $p_n > 0.5$ and $d = 3 \text{ \AA}$ between the two nitrogen atoms we estimate $\Pi > 0.166 \text{ \AA}^{-1}$.

In the Biosym/CVFF force field, the +1 charge of ammonium groups is spread over the N atom and its 4 neighbors: all these atomic charges will be multiplied by the corresponding p factors.

2. Definition and mapping of the potentials in the TR site:

i. The enzymatic site and the ligands: Unpolar CH_n groups in the substrates were considered as "united atoms" in order to speed up the docking computation. The charges carried by the deleted hydrogens were summed up with that of the carbon atom, which produced most of the time electroneutral atoms. Since there are no such united atoms in the current CVFF force field, they were assigned the van der Waals parameters of the unpolar carbon potential type.

For reasons of homogeneity, the same treatment has been applied to the CH_n groups of the site during the calculation of the potentials at the grid points. The experimentally determined X-ray structures^{10,11} of TR from *C. fasciculata* have been used for this purpose.

a). *X-ray geometries*: A first option we have considered here is to calculate the potential grids using the *X-ray geometries* of either the free or the complexed TR sites. The protonation state of ionizable groups of the protein has been set by the Biopolymer/Biosym hydrogen setup routine at pH=7.

b). *Average geometries considering the movement of the side chains*. An alternative approach was to subject the free enzymatic site to 100 picoseconds (ps) of MD simulation in vacuum at 300K, while fixing the backbone atoms and allowing only the sidechains of the catalytic site to move. Geometries were sampled every ps and their corresponding potential values at every grid point have been averaged.

This approach is not meant to account for the *flexibility* of the site during the docking process, but has been performed in order to *smooth out* the very steep variations of the potentials near the site atoms, which may cause optimization artifacts when placing the ligands in the site. In other words, we have replaced the rigid X-ray geometries by an average "fuzzy" geometry of the site, which is considered as *rigid* during the docking process.

ii. Definition of the grid: A parallelepipedic box of 30x40x35 Å³ has been defined around the active site of the TR enzyme, which is available in the Brookhaven Protein Data Bank⁵³. The grid spacing was chosen to be of 0.5 Å, shown to be reasonably accurate, except for the immediate vicinity of the site wall atoms, which might need an explicit treatment.³³ We have used the grid description throughout the site, in the perspective of low cost in computer time.

iii. The expressions of the mapped potentials: The grid method can only be used for energy terms that can be written as a product between a scalar potential V and a corresponding property of the ligand atom that quantifies the influence exercised on it by this potential. These different potentials are evaluated at each point P of the grid, in function of its distances d_{iP} to the site atoms i .

a). *The Coulomb and van der Waals potentials* were parametrized according to the CVFF force field:

$$V_{coul}(P) = \sum_{i=1}^{siteatoms} \frac{Q_i}{4\pi\epsilon_0 \cdot (2d_{iP}) \cdot d_{iP}} \quad (3)$$

$$V_{vdW}^{rep}(P) = \sum_{i=1}^{siteatoms} \frac{A_i}{d_{iP}^{12}} \quad (4)$$

$$V_{vdW}^{att}(P) = - \sum_{i=1}^{siteatoms} \frac{B_i}{d_{iP}^6} \quad (5)$$

b). *The desolvation term* proposed by Gilson and Honig⁵⁴ is one of the few continuum solvent models that can be used in a grid calculation (also see the very similar term³²), in contrast to more elaborated expressions of the solvation potentials⁵⁵ which do however not have this property.

The desolvation can be understood as the effect of changing the dielectric properties of the space traversed by the electric field lines of the charge of atom i when approaching a low-dielectric (ϵ_{int}) atom j of volume v_j which displaces the high-dielectric (ϵ_{sol}) solvent, assuming that no distortions of the electric field lines occur at the boundary between the two dielectric

media. Of course, atom i plays at the same time the role of displacing atom with respect to the charge Q_j of atom j . Two desolvation potentials expressing the displacement of the solvent around the ligand atoms by the site atoms and reversely, need to be defined:

$$V_{desolv}^{sit \rightarrow lig}(P) = \frac{1}{8\pi^2 \epsilon_0} \left(\frac{1}{\epsilon_{int}} - \frac{1}{\epsilon_{sol}} \right) \sum_{i=1}^{site_atoms} \frac{v_i}{d_{iP}^4} \quad (6)$$

$$V_{desolv}^{lig \rightarrow sit}(P) = \frac{1}{8\pi^2 \epsilon_0} \left(\frac{1}{\epsilon_{int}} - \frac{1}{\epsilon_{sol}} \right) \sum_{i=1}^{site_atoms} \frac{Q_i^2}{d_{iP}^4} \quad (7)$$

These terms are always positive and from the point of view of a pharmacophore approach they can be interpreted as a penalty against placing charged ligand atoms near the site wall due to the loss of their solvation shell.

c). The "hydrophobic effect" is usually included in calculations as a term proportional to the hydrophobic surface that becomes solvent-inaccessible during the binding⁵⁶. In order to make this approach compatible with our grid technique, we ignore the multiple overlaps between atomic spheres and use the fact that the area cut out by the sphere of the atom j , with a radius r_j , from an overlapping atom i of radius r_i and with an interdistance center $d_{ij} < r_i + r_j$ can be written as:

$$b_{j \rightarrow i} = \pi r_i (r_i + r_j - d_{ij}) \left(1 + \frac{r_j - r_i}{d_{ij}} \right) \quad (8)$$

Here, the spheres correspond to "united atom" unpolar CH_n groups, considered to have equal radii $r=2.5 \text{ \AA}$, so that equation (8) becomes linear in

d. We define a "hydrophobic potential" that is zero when there is no contact ($d_{ij} > 2r$) and otherwise is proportional to the linear expression of the overlapping hydrophobic areas $b_{j \rightarrow i} = \pi r(2r - d_{ip})$:

$$V_{hphob}(P) = -\pi r \sum_{i=1}^{siteatoms} \delta(i) \cdot [2r - d_{iP}] \quad (9)$$

where $\delta(i)$ is 1 if i is a nonpolar united atom and 0 otherwise.

3. The energy functions of the model and their fittable parameters:

i. Molecular Hamiltonian or 3D QSAR descriptors? In contrast with other QSAR models based on computed 2D or 3D descriptors in which no straightforward physical explanation can be attached to the obtained equations, the terms used in this work maintain a physical foundation. Therefore, the *fittable* weighting coefficients that will be introduced furtheron can be considered either as physical properties for which the values cannot be measured or are uncertain in the current simplified context, or as typical QSAR parameters, meant to account for different relevance of different energy contributions to the biological activity, as suggested in literature.²⁵

a). *Electrostatic term*: The Coulomb term using a distance-dependent dielectric constant of 2.0 to account for the presence of solvent and the Debye-Hueckel atmosphere of counterions²⁹ may perform better than the classical one⁵⁷, but it is still not a realistic treatment of charge-charge

interactions in dielectrically heterogeneous media⁵⁸. We introduce a fittable weighting coefficient χ to allow the choice of the effective dielectric constant that fits best with the experiment.

b). Desolvation: Many of the parameters occurring in the Gilson-Honig desolvation term, such as the values of dielectric constants, the atomic volumes and radii, have been introduced by applying a macroscopic theory at the molecular level. Since the variance of the individual atomic volumes v_i in equation (6) is not relevant⁵⁹, these have been replaced with an average volume $\langle v \rangle$ that can be taken out of the sum. In equation (14), we set $v_i = \langle v \rangle = 1$ and let λ scale up this average volume to its effective value.

c). Hydrophobicity: The hydrophobic potential must be expressed in kcal/mol by multiplication with an appropriate coefficient η .

d). Nonbonded interactions: The van der Waals terms have been borrowed from the Biosym-CVFF force field, only accounting for the site-ligand interaction in vacuum. The effective change in attractive van der Waals energy may be much smaller than this, since the ligand-solvent attractive van der Waals term that is lost upon the desolvation at binding is not accounted for in the present approach.

Rigorously, we need to evaluate the *potential of mean force* of the nonbonded terms exercised by the moving solvent on each ligand. Accurate dynamic simulations of these interactions⁶⁰ are however beyond the purposes of this work.

Other authors³³ have included explicit site-ligand van der Waals attractive terms in their docking models, while enzyme-solvent and ligand-solvent dispersive interactions are implicitly accounted for by the use of a continuum solvent model. The van der Waals terms are an expression of the site-ligand *shape complementarity*. Therefore, we expect them to be important descriptors of the docking process. We will take over the vacuum repulsive van der Waals term without any corrective factor, assuming that the role of this short-ranged term is to discard the geometries with site-ligand clashes, without significantly contributing to the energy of a correctly placed ligand. In contrast, a weighting factor ω will be used to scale the van der Waals dispersion term.

d). Protonation model: The coefficient Π describing the protolytic equilibria is meant to fill in the missing information concerning the pK_a values of the successive protonation steps of the free ligands and their possible shifts upon binding to the enzyme wall.

Considering the previous discussions, the expression of the site-ligand interaction energy becomes:

$$E_{S-L} = E_{S-L}^{rep} + \omega E_{S-L}^{att} + \chi E_{S-L}^{coul} + \lambda E_{S-L}^{des} + \eta E_{S-L}^{hphob} \quad (10)$$

Each of the energy terms in equation (10) can be expressed in function of the previously defined potentials:

$$E_{S-L}^{rep} = \sum_{i=1}^{ligand_atoms} A_i V_{vdw}^{rep}(i) \quad (11)$$

$$E_{S-L}^{alt} = \sum_{i=1}^{ligand_atoms} B_i V_{vdw}^{alt}(i) \quad (12)$$

$$E_{S-L}^{coul} = \sum_i^{ligand_atoms} Q_i V_{coul}(i) \quad (13)$$

$$E_{S-L}^{des} = \sum_i^{ligand_atoms} Q_i^2 V_{des}^{sit \rightarrow lig}(i) + v_i \cdot V_{des}^{lig \rightarrow sit}(i) \quad (14)$$

$$E_{S-L}^{hphob} = \sum_i^{ligand_atoms} \delta(i) V_{hphob}(i) \quad (15)$$

Here, both Coulomb and desolvation terms are implicit functions of the parameter Π , since they are calculated using the weighted atomic charges as derived from the protonation model.

ii. Energy levels of the free ligands: An *a posteriori* solvation correction must be added to the energies of the ligand geometries that were minimized in vacuum. The Gilson-Honig continuum model has been shown not to be able to reproduce accurately vacuum-to-water transfer energies of small molecules.⁵⁹ Equations (6) and (7) may be reasonably accurate at large interatomic distances, but cannot be applied for bonded atoms. The *differences* of the solvation energies between different conformations can however be obtained, since the relative position of the bonded and the geminal neighbors is practically conformation-independent and only the contributions of the nonbonded atoms are important. Denoting the sum of all intramolecular force field terms *except* the Coulomb contribution (E^c_i) by E^{nc}_i , the intramolecular desolvation contribution by E^d_i and introducing the

weighting parameters in the expression of the intramolecular total energy E_i^{free} , we have for each conformation i from the set $C_i(M)$:

$$E_i^{\text{free}} = E_i^{\text{nc}} + \chi E_i^{\text{c}} + \lambda E_i^{\text{a}} \quad (16)$$

iii. Energy levels of the bound ligands. Docking energy. The site-ligand interaction energy as defined in equation (10) is a function of the geometry of the docked conformer i , the relative position $\mathbf{r}=(r_x, r_y, r_z)$ of the mass centers of the two objects and the 3 angles $\alpha=(\alpha_x, \alpha_y, \alpha_z)$ defining the rotation of the ligand with respect to the reference system of the site. In function of the *current conformation* and of the *initial placement* of this conformer in the site, the *minimization* of E_{S-L} with respect to these 6 degrees of freedom will converge towards the closest local minimum, defining an energy level of the docked ligand.

a). *Initial placement of the ligands in the site.* This optimization is quite problematic, since it may get trapped in a lot of local minima and depends on the initial orientation of the ligand in the site. Therefore, a mapping of the TR site was performed in order to design different starting points from which the ligand will be let to evolve during optimization. These points (see Fig. 2) should:

- 1). *cover the whole accessible site* such as to ensure that the ligand will find the optimal binding pocket when starting from at least one of these different initial positions

2). *be equally spread across the site* or else their distribution will bias the relative populations of the minima, e.g. the ligand will converge more often towards binding locations found in zones that were more densely covered with starting points.

In order to fulfill these requirements, a grid of 2Å spacing has been generated around the active site of the enzyme. The grid points for which the nearest site atom was found at a distance between 4 and 7Å have been considered as potential starting locations for the centers of ligands. A Monte-Carlo algorithm which will not be detailed here has been used to pick out a number of $N_s=7$ starting points out of these candidates, in maximizing the *distance between the two closest points* of the selected distribution.

b). The docking algorithm: The docking algorithm begins by translating the current conformation with its center of mass in one of the N_s starting points. A complete systematic rotation of the ligand with a coarse step of $\pi/2$ is performed in order to find a bump-free initial orientation. A combination of optimization methods (systematic rotations, translations, iterative searches by moving a single coordinate at once and eventually a conjugate gradients minimization) is used instead of an unique minimization algorithm, to reduce the risk of falling into false minima. This is repeated with all the other starting points. Eventually, the whole procedure is repeated with the *mirror image* of the current conformation in order to retrieve the enantiomeric geometries that were discarded at the 3D-building step. If a chiral center is present in the ligand molecule, this approach ensures the docking of the racemic mixture.

For each conformation $i=1\dots N_{\text{free}}$ of the free ligand, the different N_s optimizations carried out from each starting point $k=1\dots N_s$ will lead to a series of local energy minima E_n^{dock} , where $n=k+(i-1) \cdot N_s$. The further N_s minimizations that are performed for each *mirror-image* geometry i' will provide other $n' = N_s \cdot N_{\text{free}} + k + (i-1) \cdot N_s$ energy levels to this series. All these energy levels E_n^{dock} , $n=1\dots N_{\text{dock}} = 2 \cdot N_s \cdot N_{\text{free}}$ will be used to characterize a bound ligand that had N_{free} sampled conformations.

$$E_n^{\text{dock}} = E_i^{\text{free}} + \min_{r,\alpha} E_{S-L}(i/i'; k; r, \alpha) \quad (17)$$

4. Evaluating the overall binding scores on the base of the conformational binding energies:

i. Binding "enthalpy" and "entropy": While the so defined free and bound energy levels do not form canonical ensembles, the equations of statistical mechanics are nevertheless a plausible choice for a *set of rules to derive some global docking scores* from these conformational score values.

a). *The energy difference between the best minima of the bound and free states:* In a first approximation, we can consider that the energy difference between the most stable bound state and the most stable free state may be an index of the binding affinity of a ligand:

$$\Delta H^* = \min_n (E_n^{\text{dock}}) - \min_i (E_i^{\text{free}}) \quad (18)$$

b). *Definition of the "binding enthalpy" index*: Adopting the statistical mechanics formalism as the *definition* of the binding enthalpy index and considering a Boltzmann factor β we obtain:

$$\Delta H = \frac{\sum_n E_n^{dock} \cdot e^{-\beta E_n^{dock}}}{\sum_n e^{-\beta E_n^{dock}}} - \frac{\sum_i E_i^{free} \cdot e^{-\beta E_i^{free}}}{\sum_i e^{-\beta E_i^{free}}} \quad (19)$$

c). *Statistical Mechanics definition of the entropy index*: The binding process is function of the *global* entropy difference of the *families of populated conformations* in the free and bound states. This would require extensive calculations⁶¹ based on the free energy perturbation method (FEP) which are not compatible with the constraints on computer time imposed by the aim of this work. Since here we make use of well-defined conformational families of the ligands, we may obtain the TΔS index between the free and bound states:

$$T\Delta S = \Delta H - \Delta G = \Delta H + \frac{1}{\beta} \ln \left(\frac{\sum_i e^{-\beta E_i^{dock}}}{\sum_j e^{-\beta E_j^{free}}} \right) \quad (20)$$

In our rigid-body docking approach, we do not consider the flexibility of the conformers around their local energy minimum. Our model can only describe the entropy in terms of the *number* of existing populated states and their *relative importance* (see Fig. 3).

We also assume that the loss of disorder due to merging two independently moving particles into a complex, not reflected by the quantity $T\Delta S$, is constant for all the ligands, at least as far as these are about the same size. To put it in another way, it is well-known that most of the QSAR models work because the effects that were not explicitly accounted for are either almost constant throughout the set of considered species or are covariant with one of the considered descriptors. Analogously, it is still possible that under *equal* conditions of conformational sampling and docking, the differences between flexible and less flexible ligands may be evidenced by equation (20) despite a possible incompleteness of the sampling.

d). An empirical estimation of entropy: Fig. 3 suggests an alternative and completely empirical manner in which entropic effects may be accounted for. If many energy levels are distributed close to the lowest level of a system, many states of the system have high probabilities to be populated and therefore its entropy will be large. These energy levels have important Boltzmann weights and will therefore significantly contribute to the *increment* of the enthalpy H of that ensemble with respect to its minimal energy level H^* .

This lead us to address the question whether the agglomeration of energy levels around the best minimum will have as consequence the increase of $(H-H^*)$. It is important to point out that H and H^* are strongly intercorrelated variables and therefore, *in principle* they need not be used simultaneously in a multilinear regression model. However, in this

particular case, if TS and $(H-H^*)$ were to some extent covariant, this would prove that the latter magnitude acts as an alternative empirical descriptor of the entropy *within the framework defined by the sampling axioms*. Consequently, the *difference* $(\Delta H - \Delta H^*)$ may be an alternative indicator of the binding entropy $T\Delta S$. There are however two main concerns with respect to the use of this descriptor:

- is it able to contribute to the explanation of the variance of the affinities of the molecules and, if this is the case, does this alternative recipe to estimate entropy perform better or worse than the $T\Delta S$ entropic index?
- does it contain information that is complementary to the one furnished by the previously defined $T\Delta S$ entropic index, in order to be used as an additional, non-redundant explaining variable of our model ?

5. Calibration of the docking model:

In the previous sections, we have defined three indexes - the energy difference between the lowest levels in equation (18), binding enthalpy in equation (19) and binding entropy in equation (20). These are non-linear functions of the parameters $(\omega, \chi, \lambda, \eta, \Pi)$. The Boltzmann factor β appearing in the expressions of ΔH and $T\Delta S$ has been set equal to $1/RT = 1.67 \text{ mol/kcal}$ corresponding to $T=300\text{K}$.

i. The relation between the affinity constant and the calculated binding indexes: These indexes are not the exact thermodynamic binding values in terms of which the inhibition constant K (the inverse of the affinity constant) can be expressed as $\ln K = \Delta H/RT - \Delta S/R$. Therefore, we wish to establish a linear correlation between the logs of the affinity constants of the ligands and the previously defined indexes, as done in a typical QSAR approach:

$$\ln K_i = a.\Delta H_i + b.\Delta H_i^* + c.T\Delta S_i + d \quad (21)$$

If such an equation can be found and statistically validated, it can be used furtheron for the *prediction* of affinities.

Please note that, in order to avoid confusion, we will denote the inhibition constant by K, not by K_i as accustomed in enzymology. The latter symbol will design the inhibition constant of *a specified compound i*.

ii. Fitting procedure: We have used the measured values of the inhibition constants of a series of TR inhibitors from the literature (see Fig. 1 for the generic structures and Table 1 for the $\ln K$ values and refs.) to calibrate the fittable parameters ($\omega, \chi, \lambda, \eta, \Pi, a, b, c, d$) of our model, e.g. to minimize the discrepancies between calculated and measured K values.

a). The two classes of parameters: The parameters ($\omega, \chi, \lambda, \eta, \Pi$) and (a, b, c, d) form two distinct classes of variables. The former are input in the docking algorithm, while the latter define the linear relation between the herein

calculated binding indexes and the affinity. We will refer to them as the "docking" and the "affinity" parameters respectively.

The optimal docking parameters are found by means of a Monte Carlo (MC) procedure, while the values of the affinity parameters are derived by means of a multilinear regression in function of the binding indexes that were calculated for the current configuration of the docking parameters. The details of the optimization procedure with respect to both classes of parameters will be outlined in Appendix B.

iii. Acceptable linear relations: While in QSAR there is generally no concern about the coefficients of a linear relation as far as it has been proven to have predictive power, the fact that we have defined our indexes starting from some basic energetic arguments entitles us to impose some restrictions concerning the relations that seem acceptable:

- A relation in ΔH and $T\Delta S$ should have a positive coefficient a and a negative c . Ideally, $a = -c = 1.67$ kcal/mol.

- We may also accept any linear combination of ΔH , ΔH^* and $T\Delta S$ according to our supposition that $(\Delta H - \Delta H^*)$ may be an alternative descriptor of the binding entropy. This is more difficult to interpret and is closer to a QSAR-like equation than to a physically significant relation. However, we still expect a positive coefficient for the binding enthalpy ΔH .

Any configuration of docking parameters that correspond to a local maximum of r^2 , leading to affinity coefficients that do not respect these conditions, will be considered as fitting artifact and discarded.

6. Testing and using the model.

Other inhibitors (see Table 1) of known inhibitory potency have been used to test whether the previously calibrated model is able to predict their binding affinities. Then, the method was applied as such on a library of 2500 molecules (part of which are compounds available in ACD⁶²) but had never been tested for TR inhibitory potency. We selected the molecules of molecular mass between 200 and 500 g/mol, with low predicted $\ln K < 3$ or with calculated "binding enthalpies" that are strongly negative ($\Delta H < -20$ kcal/mol) to be screened for TR inhibition potency.

The percentages of inhibition are given as $(r_0 - r_{inh})/r_0 \times 100\%$, where r_0 and r_{inh} are the rates of the enzymatic reduction of trypanothione disulfide by TR, measured in absence and in presence of the inhibitor, added in decreasing concentrations to the reaction environment under the experimental conditions described in⁶³. An error of about 10% may affect the measured percentages of inhibition.

II. Results and discussions:

1. An estimation of the accuracy of the desolvation term:

We have advanced the hypothesis that the Gilson-Honig solvation model is able to estimate the *conformational differences* of solvation energies in spite of its failure to predict their absolute values. This affirmation has been verified by comparing the Gilson-Honig desolvation terms of the conformations of *free* ligands with the corresponding energies obtained from a more realistic potential of mean force of solvation, making use of a modified⁶⁴ boundary element method (BEM) to solve the Poisson Equation.

Ideally, the Gilson-Honig term should be linearly related to the exact solvation potential, with the same slope λ appearing in the expressions (10) and (16) of the docking energies. Table 2 summarizes the relations established for the sets of docked conformations of several compounds, considering both a free slope λ' and the λ value found after the calibration of the model.

For most of the molecules, fair to excellent correlations between the two solvation terms were found. In the two situations having $r^2 < 0.5$, the failure to obtain a fair linear correlation can be explained by the fact that the solvation energy showed little change for the considered conformations. Indeed, the corresponding RMS deviations between the «exact» BEM-based solvation energies and those than are obtained as a linear function of the Gilson-Honig term are much smaller than those for molecules that display

considerably larger correlation coefficients. The Gilson-Honig model is unable to account for such fine fluctuations of the solvation energy, but the corresponding errors are accordingly small. However, the slopes of these regression lines are unreliable and need to be discarded from our discussion.

Considering only the 6 compounds with $r^2 > 0.7$, we obtain an average λ of 2.07 ± 0.56 . This is in a very good agreement with the value of 2.30 independently found at the calibration of the model and without making any reference to the exact solvation energies.

At $\lambda = 2.30$, there is no obvious decrease in the quality of the correlations. If we interpret this λ as an average atomic volume, we obtain an average atomic radius of 0.8 Å, which is a reasonable value when considering the excluded volume effect due to multiple overlapping of atomic spheres. All this encourages us to affirm that the use of the Gilson-Honig term offers a basically correct description of the conformational variations of the solvation energies.

2. The efficiency of the conformational sampling procedure:

We have compared our sampling approach with a classical conformational analysis based on a 100 ps (10^5 equilibration steps of 1 fs, with saving of the geometry at every 100-th step) of a high-temperature (1000 K) Molecular Dynamics simulation in vacuum, performed with the Discover-Biosym program⁴⁷. For this purpose, the minimization and selection procedures developed here have been alternatively run with the *diversity-biased input set* of initial geometries and respectively an *equal number of frames* chosen

from the MD trajectory of the compound (sorted by increasing potential energy).

The computer times needed to carry out the MD simulations were however longer (typically 15-20 minutes) compared to the torsional angle driving that completes in 1..2 minutes on an Iris Indigo R4000.

In Table 3, we show some typical examples of the energy differences between the best minima obtained by our sampling approach with respect to the ones generated by the MD run, as well as the final number of accepted conformations. Our sampling procedure has been able to find significantly lower energy minima for most of the ligands.

The bias in favour to more « extended » (and therefore hopefully better solvated) conformations according to samplig axiom c (see §1.ii, Methods) appeared to be successful for compound **8**, for which the MD geometry set led to a better vacuum minimum, while the diversity-biased set found conformations that are more stable in solution according to the considered solvation correction. In contrast, it did not work for ligands **2** and **7**.

The final number of accepted minima is also significantly higher when a maximum diversity is imposed to the set of initial geometries. The probability that two different geometries converge towards the same minimum is higher within the set of MD-sampled geometries.

Our sampling procedure guarantees that all the transitions around the 10 main torsion angles in the molecule are effectively explored, while even high-temperature MD may fail to cross certain barriers. This fact, combined

to the diversity biased selection of the visited conformations has been shown to lead to sets of geometries that are definitely less intercorrelated than the ones sampled by MD at 1000K.

3. Calibration results

i. Averaging of the side chain geometries of the TR site. A first important finding of our various calibration trials is that only the grid obtained by *averaging over the MD trajectory* of the side chain positions in the active site of the enzyme lead to good correlations of calculated vs. experimental binding energies. Averaging ensures the smoothing of the steep potential variations near the site atoms. Furthermore, a functional group of the site wall that is moving during the MD simulation can be "seen" in the final grid as occupying the whole region of space visited. Averaging would not have been a good strategy if the movement of site residues would have been strongly restrained upon complexation with a ligand. Therefore, our finding corroborates with the little differences observed between the side chain geometries of free and complexed TR site in the previously cited X-ray structures.

ii. Optimal parameters: Several different Monte Carlo optimization runs have lead to the sets of coefficients reported here.

An analysis of the configurations of docking parameters that have been explored at intermediate stages of the MC search showed that in some of these cases, the corresponding affinity parameters did not respect the

constraints imposed in §5.iii of the Methods section. The optimization procedures never converged towards such points, but lead to minima that agree well with these expectations.

The values reported in Table 4 ("docking" parameters) and Table 5 ("affinity" parameters) come from different runs with different learning sets. The calibration set A was restricted to smaller ligands, while sets B and C progressively include larger and more flexible compounds.

a.) *The "docking" parameters:* Encouragingly, the docking parameters do not radically differ in function of the chosen learning set. The largest discrepancies are observed between sets A and B, e.g. upon *increasing the molecular diversity* of this set.

Van der Waals term: The weighting factor ω , which was fixed to 1 in the run A, did not considerably drift away after fitting (sets B and C), although the Monte-Carlo search was carried out in the range $\omega \in [0,1]$.

The van der Waals term therefore appears to be an important descriptor of the docking process. However, this does not mean that the van der Waals effects due to the solvent can be neglected, but merely that they are not proportional to the vacuum van der Waals energies and therefore cannot be accounted for by weighting this term.

Desolvation term: As we have already shown, the desolvation weighting factor λ is in good agreement with the values that need to be applied in

order to bring the approximate Gilson-Honig term in agreement with more elaborate estimations of the solvation energies.

Hydrophobic term: Considering the pairwise intersections of the spheres without accounting for the "excluded volume" effect, lead to an overestimated buried area. Accordingly, the order of magnitude obtained for the hydrophobic coefficient is somewhat lower, but falls in the usual value range found in the literature.^{55,59}

Coulomb term: The weighting factor of the Coulomb term has been also found close to 1 (the range explored by the Monte-Carlo search was [0.5,1.5]), suggesting that the initial choice of the distance-dependent dielectric constant $\epsilon=2.d$ was reasonable.

The "protonation constant" Π : The obtained values for the protonation constant Π predict a fraction of the diprotonated piperazine ring of about 20% (case A), 60% (case B) and 30% (case C). While the former and the latter values are in good agreement with the expected population of such species at pH 7, case B predicts a quite high rate of diprotonation.

One would expect that all other docking parameters are strongly coupled to Π , since the latter controls the charge distribution that defines both the intra and intermolecular Coulomb and desolvation terms. This is not observed. The weighting factors χ, λ, ω and η adopt practically the same value in the sets B and C. Only polybasic ligands are affected by the parameter Π , while the monoamino derivatives in the learning set ensured that the other

weighting factors do not drift away in order to better accommodate the changes of the total charge of the former.

Actually, these about 30% difference in the total charge has a moderate influence on the calculated binding indexes and predicted binding constants of most of the *piperazine-containing* ligands. There are two notable exceptions for which this influence is extremely strong: compounds 28 and 30, both containing a fluorine atom at the same position of an aromatic ring.

The quality of the model in function of the parameter Π shows a very broad maximum for Π within the range between the two extreme values it takes in parameter sets A and C. In other words, the model is quite tolerant with respect to an average error of about 0.3 charge units on a piperazine moiety of a ligand. This is not astonishing, since some compensation between the favorable Coulomb interactions and the defavorable desolvation term is expected to occur as the global charge assigned to a ligand increases. Nevertheless, the correction for protolytic effects we have introduced cannot be dismissed as irrelevant, since otherwise the error committed with respect to the total charge of ligands would have been at least of one unit.

b). *The "affinity" parameters.* In contrast to the relative stability of the docking parameters, Table 5 shows that the relations between the log of the inhibition constant and the binding indexes significantly differ from one parameter set to the other. The best relations have been found by stepwise cross-validated regression and reflect the optimal trade-off between the number of included variables and the predictive power. For parameter set A,

this relation is in ΔH and $T\Delta S$ only, while in the other two cases, ΔH^* is also involved. The table also displays the regressions in ΔH and $T\Delta S$ only, the variables one would expect to correlate with the inhibition constant. These latter equations have, in contrast, quite stable coefficients. Again, the result obtained with set A differs the most from those obtained with sets B and C, as previously seen for the docking parameters.

It is essential to point out that the stepwise and cross-validated regression technique we have used (see Appendix B) ensures that the simultaneous presence of ΔH and ΔH^* leads to a better predictive power (cross-validated r^2) than the one obtainable by considering only one of these intercorrelated variables. While the correlation coefficient between ΔH and ΔH^* is as high as 0.995, the magnitude $(\Delta H - \Delta H^*)$ is completely independent with respect to both ΔH and ΔH^* (correlation coefficients of about 0.2). $(\Delta H - \Delta H^*)$ does however not represent the « noise » due to the imperfect proportionality between ΔH and ΔH^* , but may be an alternative way to express the binding entropy difference.

Writing the $T\Delta S$ values of the ligands as a linear combination of ΔH and ΔH^* leads to the linear correlations of fair quality listed in Table 6. It is found that $T\Delta S \approx 1.2 \dots 1.4(\Delta H - \Delta H^*)$, irrespectively of the set of docking parameters. However, the correlation coefficients of 0.5...0.6 ensure that these are not redundant variables. $T\Delta S$ and $(\Delta H - \Delta H^*)$ are nevertheless clearly covariant (see Fig. 4).

Interestingly, ΔH^* is admitted to enter in the model *upon the introduction of larger and more flexible ligands in the learning sets*. The entropic term $T\Delta S$ that worked reasonably well with set A of relatively small ligands has apparently been replaced by a more elaborate descriptor of entropy that incorporates both the original $T\Delta S$ and the alternative $(\Delta H - \Delta H^*)$. There are several observations that support this interpretation:

- if ΔH^* is not taken into account in the regressions, we obtain linear regressions for the parametrization schemes B and C. In this case, we can affirm that basically the *same* linear relation is obtained for both situations. We can say according to these relations that the overall "entropic" contribution to $\ln K$ is about $-2.T\Delta S$, which is again a *stable* result and surprisingly close to the theoretical $-1.6.T\Delta S$.

The large discrepancies in the coefficient of ΔH observed in Table 5 are mainly due to the fact that the $(\Delta H - \Delta H^*)$ *entropic* descriptor displays very different contributions in these regression equations. The overall "enthalpic" term amounts about $0.1\Delta H$, with a positive coefficient as expected, but much lower than the theoretical $1.6\Delta H$.

- Table 7 compares the computed ΔH and $T\Delta S$ terms for all the considered ligands in function of the 3 different parametrizations. It can be seen that the ΔH values obtained with the different parameter sets are very well

intercorrelated. It is the entropy index $T\Delta S$ that displays the stronger dependence on the used parametrization

c.) *QSAR model or physical binding model?* The fact that the previously discussed "docking" coefficients adjust to values that can be backed by various theoretical or computational arguments lead us to address the question whether the obtained models are indeed able to quantitatively predict the thermodynamic binding parameters or whether we have obtained a QSAR-like correlation, with a reasonable predictive power (see the discussion later on) but without a well-defined physical background.

Interestingly, the low coefficient found for the participation of the term ΔH would suggest that this variable is rather a QSAR index than a measure of the thermodynamic binding enthalpy. In spite of the fact that its weighting coefficient equals only 10% of its *theoretical* value $1/RT$, the enthalpy index ΔH accounts for roughly 25% of the explained variance, while $T\Delta S$ covers the rest of 75% of the explained variance for parameter set A.

The entropy indexes are strongly affected by the sampling of ligand conformers and are dependent on the current parametrization of the model. Nevertheless, the contribution of about $-2.T\Delta S$ in the "forced" regression equations suggests that while the $T\Delta S$ term for any particular ligand is affected by random errors, in general the calculated entropies may reflect quite well (up to a constant offset) the real binding entropies.

While the enthalpy index primarily depends on the *separation* between the bound and the free energy levels, the entropy index is only a function of the conformational energy differences within each of the ensembles of the bound and free states. Therefore, if the cause for the low coefficient of ΔH is a *conformation-independent* error in the substrate-ligand interaction energy, the entropy values are *not* affected by this problem.

We can outline two alternative explanations for the low coefficients of the enthalpic term:

- The first implies the *enthalpic* effects that were *not* included in the present approach, e.g. the ligand-solvent van der Waals interactions, cavity contributions or Debye-Hueckel effects. These terms depend on the overall molecular shape, exposed molecular surface and on the overall disposition of the charged groups respectively. They are not strongly conformation-dependent and corroborate well with the previous discussion of the presumed error affecting the binding enthalpy term.

- Alternatively, an *enthalpy-entropy* compensation as outlined in⁶⁵ could be responsible for the artificially small enthalpy coefficient. The rotational and translational entropy loss has been considered as a constant $T\Delta S^0$ for all the ligands in the series. However, the less tightly bound ligands (less negative enthalpy ΔH) still maintain a certain mobility in the site and effectively lose only a fraction $T\Delta S^{\text{eff}}$ of this assumed translational-rotational entropic contribution, and it appears that $T\Delta S^{\text{eff}} = \alpha \cdot \Delta H$. Since this $T\Delta S^{\text{eff}}$ does not explicitly appear in our model, it can be implicitly accounted for by setting

the weighting coefficient of the enthalpy equal to $(1-\alpha)$ since $\Delta H - T\Delta S^{\text{eff}} = (1-\alpha)\Delta H$. Situations in which α is as large as 0.8-0.9 are outlined in⁶⁵ and therefore, the 10-fold reduced weighting coefficient of ΔH can be well explained within this hypothesis.

While thermodynamic studies of the binding of TR inhibitors are not known to us, microcalorimetric measures of the binding enthalpies for glutathione and analogs to GR⁴ show that these are as low as -25 kcal/mol and need to be compensated by a very unfavorable binding entropy contribution in order to reach the ≈ -4 kcal/mol of binding free enthalpy corresponding to the affinity constant of about 80 $\mu\text{mol/l}$. The predicted ligand-TR binding enthalpies in the present study are as well within this order of magnitude.

The GR binding enthalpies may not be very instructive for the TR, so that we cannot infer whether this order of magnitude of the binding enthalpies applies for the latter enzymatic system. We can nevertheless forward the hypothesis that the same compensation between a large negative enthalpy and a large negative entropy yielding a small binding free enthalpy will also occur in the case of TR, given both the strong homology of the active sites and the important flexibility within both classes of ligands.

In conclusion, while there are hints that the calculated binding indexes may have at least the correct order of magnitude, some of the features of the model strongly suggest an empirical QSAR, as for example the increase of the predictive power upon the introduction of the $(\Delta H - \Delta H^*)$ entropy

descriptor and the low coefficient of ΔH in the expressions of $\ln K$. These indexes offer valuable information about the binding of the TR ligands in spite of the simplicity of the interaction models they are based on. On the other hand, it cannot be ruled out that the calculated entropy indexes may relate quite well to the actual binding entropies, while the enthalpy index is affected by an error that could be explained on behalf of ignored energetic or entropic contributions.

4. Predictive power of the model

i. The general applicability of the docking model. In contrast to a large part of the QSAR models that are studying the effect of different substituents on a common molecular scaffold, our model may in principle be applied for any organic compound of the size of 200-500 g/mol. While calibrated on the base of polyamino ligands, the model was fully capable to predict very well the affinity of a different species such as crystal violet **44**, which has been input as a single geometry obtained by MOPAC-AM1⁶⁶ minimization since its delocalized electron system makes it unsuitable for empirical MM calculations.

A limitation of the applicability of the model in its current form may arise from the modeling of protolytic equilibria in ligands containing both positive *and* negative groups. This is not a fundamental problem since the protonation model can be extended on the base of an analogous reasoning to include zwitterionic species as well.

ligands in «strong» vs. «weak» TR inhibitors, the two classes would differ by less than 3 order of magnitude (roughly 6 lnK units or 4 kcal/mol). The previously cited model, with a reported root mean square error of 1.14 units in the prediction of log K, would be hardly able to discriminate between them. Our approach has been proven able to do so, but on the other hand there is no guarantee that the regression equations (which obviously show good agreement with the experimental points within the affinity range within which they have been calibrated) would still be able to correctly account for the affinities of say nano- or picomolar TR inhibitors. In contrast, the high correlation coefficients reported for the model⁶⁷ prove that the predictions issued by this approach are valid over the whole affinity scale. To our knowledge, there are no TR inhibitors of affinities well below 1 μM and therefore, a further validation of our model would necessarily imply its application to other enzymatic systems. However, we do not believe that it can be generalized as such for any arbitrary enzyme-ligand systems, due to the very particular hypotheses it is based on, for example that the site geometry is not affected by the binding. New terms, notably some explicit contributions for the entropy loss due to side chain rotation hindering⁶⁹ upon binding, should be added in order to properly describe other enzymatic systems. The generalization of the model to include GR/GSSG-competitive inhibitors may be straightforward due to the important homology of TR and GR. However, this important homology also causes the complete overlap of the affinity ranges of the inhibitors of the two enzymes (the most potent GSSG-competitive GR inhibitors are 1 μM ⁷⁰), which does not solve our problem.

A key problem is to ensure that the model that has been calibrated on the subset of the "structural space" encompassed by the molecules in the learning set continues to be valid outside the explored region. Clearly, model A has been calibrated within a too narrow structural subspace and did not work for compounds of higher flexibility, mainly because the binding entropy descriptor failed to apply in these cases. The extension of the learning set led to the much better model B, with only 2 mispredicted compounds. The slight improvement of the overall quality scores of model C with respect to model B is probably not due to adding of 3 other molecules to the learning set, but simply due to the extended Monte Carlo and systematic search optimizations that lead to these results. Notably, compound 34 which was an outlier of model B could not be better predicted in spite that it has been allowed to enter the learning set of model C. A prolonged search after adding the other outliers of C to the learning set failed to find a better parametrization. The ligand 43, which is still larger than any one of the latter entries in the learning set, has been well predicted by both models B and C.

On the other hand, one may argue that the model A does not well predict the larger ligands because their conformations were not adequately sampled by the building algorithm. From this point of view, model A may be considered as being closer to a physically consistent docking model, while the further fitting of the large, "undersampled" ligands forced the model to perform a "repair" of the erroneous entropy term by admitting the alternative descriptor ($\Delta H - \Delta H^*$) to enter the regressions.

The number of 5 "docking"+4 "affinity" fittable parameters (out of which only 7 enter in model A) is reasonable compared to the 20...30 molecules considered for calibration. The size of this calibration set is definitely smaller than the ones used by other authors (a recent example⁶⁷ uses 51 complexes to fit the 12 weighting factors of the considered docking descriptors). However, the other 15...20 quite diverse compounds have been purposely omitted from the learning set in order to serve as test cases for the parameters found by fitting. Such tests are in our opinion the only unbiased proofs for the consistency of the model - using all the available ligands in the fitting procedure leads to the impossibility to affirm that the obtained RMS scores are not a consequence of overfitting.

5. Structural information from the docking computations

In general, accurate predictions of the binding geometries of ligands require extensive MD simulations considering all the possible degrees of freedom and explicit solvent. Our model does not account for the mutually induced conformational changes that normally occur upon binding. Close to the site wall, the grid description of the potentials becomes less accurate, while certain aspects as the loss of entropy due to constraining the free movement of the side chains, or short-ranged electronic interactions cannot be reproduced at all. Therefore, we expect that the predicted docking conformations are basically correct in terms of the general orientation of the ligand. Details of binding will however not be correctly represented by this approach.

conformations have been found. The fact that more of the docking trajectories explored by the mepacrine ligand are "attracted" towards binding mode 1, suggests an *entropic* advantage of the first binding mode with respect to the second. Such an explanation should nevertheless be considered with extreme caution, due to the extreme simplifications on which our model is based.

6. New TR inhibitors from virtual screening of molecular data bases:

A molecular data base containing about 2500 molecular sketches of compounds that were available from different sources has been screened by the here outlined algorithm in order to discover new TR inhibitors. The screened selection of molecules were highly functionalized, an estimated 50% of them being compounds bearing amino groups and having net positive charges. The likelihood of finding TR inhibitors in such a data base is much higher compared to a random molecular library.

The generation of the conformer families of these compounds was performed on two Silicon Graphics workstations (a 100 MHz, R4000 Indigo and a 33 MHz, R3000 4D35). This has been the longest step of the procedure and took about 2 weeks.

After a first calibration run led to the docking parameter set A from Table 4 and the corresponding affinity coefficients from Table 5, these have been applied to predict the inhibitory potency of these compounds. A "hit list" of 13 molecules has been selected and submitted to the testing⁶³. The virtual docking of the 2500 molecules took 96 hours on a 175 MHz R4400 Indy

workstation. It is interesting to note that the use of faster docking approaches in «reciprocal space»^{72,73} could significantly accelerate this last step. However, since the bottleneck of the procedure is clearly the conformational sampling step, in this particular case, the overall advantage of applying such a technique instead of the classical docking in Cartesian space would be minimal.

The results obtained on the 13 tested compounds (Fig. 7) are shown in Table 9. It can be seen that 9 out of the 13 molecules caused a measurable inhibition of the reduction of trypanothione disulfide by TR in presence of NADPH when they were present in the reaction environment at concentrations of 57 $\mu\text{mol/l}$ and lower. These are inhibitors of potency that is comparable to other molecules that were synthesized and published.

While it is perhaps too early to report a statistically significant success rate of the virtual screening algorithm, the obtained results are very encouraging. The success rate should be furtheron increased when using the newly derived parameter sets (B, C).

Fig. 7 shows that most of the retrieved chemical structures show the typical features of TR inhibitors, e.g. the presence of aromatic groups and of (poly)amino chains. Somehow different due to its heterocyclic aromatic moiety and containing a morpholine group, compound XI. matches nevertheless this typical TR-inhibitor pattern. The virtual screening procedure is able to "recognize" this overall pattern in a data base of highly diverse structures.

This is a positive result that validates our calibration procedure. The automated recognition of the (aromatic-hydrophobic spacer-positive group) motive in molecular data bases may provide interesting leads. However, other, much faster techniques like neural networks⁷⁴ may perform this pattern recognition task on the basis of the molecular topology only. Our procedure implying a quite time-consuming conformational analysis should be able to:

- distinguish between actives and non-actives *within* the subset of molecules that match the considered pattern. This is confirmed both by the calculated affinities of the ligands used in the calibration and test phase and by the high success rate of its predictions.

- discover *new* structural motives that provide TR-inhibiting properties.

Molecules **II** and **VIII**, are indeed qualitatively different, their novelty consisting in the absence of a flexible side chain carrying the ammonium group, which is included here in a condensed and rigid ring system. Molecule **II** may well represent a promising lead of a new series of related compounds. Quite analogous compounds have been reported in literature⁷⁵ to have trypanocidal activity, without however evidencing their action mechanism. Interestingly, molecule **VIII**, which has been predicted to be only slightly less active than **II**, had no measurable inhibitory potency under the given conditions. While we do not have yet an explanation for this, it is likely that the presence of the methylene bridge between the tricyclic system and the phenyl ring in **A** allows the latter to orient such as to minimize the sterical crowding around the ammonium group.

IV. Conclusions

The virtual screening approach here has been successful in finding new TR inhibitors from molecular data bases. We have introduced various approximations and hypotheses in order to quantify the binding properties of ligands by means of calculated indexes. These indexes correspond to the "virtual" binding enthalpy and entropy of the ligand according to the energy functions and sampling criteria that were defined by us. They have been shown to be useful as explicative variables with respect to the logs of the inhibition constants.

Fittable parameters weighting the importance of the different interactions in the model were calibrated with respect to measured affinities of known ligands. While the optimal values of the parameters governing the site-ligand interactions have been found to be in agreement with expectations, those of the affinity parameters weighting the enthalpic and entropic contributions to the affinity constant rather behave like empirical QSAR coefficients. No clearcut explanation about why the coefficient of the enthalpy fails to adopt its theoretical value could be forwarded, but several reasonable hypotheses which may account for this observation do exist.

The algorithm was able to fairly explain the affinities of 44 chemically different ligands and to detect new inhibitors. An interesting polycyclic molecule (II, Fig. 7) could be related to a series of trypanocidal compounds that were not yet tested with respect to TR-inhibiting properties.

Appendix A: Flowchart of the sampling algorithm

The program reads in the data (MDL-mol files) describing the molecular graph. Any hydrogens -if present- are deleted. Then, the molecule is broken up in cyclic and acyclic fragments as found by a ring detection routine. Linear and branched chains are build up according to standard covalent radii and valence angles, in fully staggered conformations. We do not impose any control on the resulting configurations of chiral atoms.

a) Retrieving the ring geometries: The cyclic fragments are matched up in a data base containing typical conformations of rings.

The ring recognition routine must decide whether the graph R of the current fragment is isomorphous with any of the graphs R' of rings in the database, e.g. if for each vertex $i \in R$ there exists an equivalent $j \in R'$. A topologically rigorous approach⁷⁶ to the problem of graph isomorphisms is beyond the purposes of the current work. We will outline a simple procedure conceived to work well for the recognition of cyclic fragments that usually occur in organic molecules and which do not have any special symmetry properties.

There are $n!$ ways to form pairs (i,j) with the atoms of two graphs of size n . This number can be limited if restricting the considered j to a subset R'_i obtained after discarding the atoms of R' that are obviously *not* equivalent to i . Therefore, the following steps are undertaken by the ring recognition routine:

- *Atom count*: if the two graphs do not have the same number of atoms, then obviously they cannot be isomorphous. Stop.

- *Evaluation of vertex descriptors*: the topological distance matrices $td(i,j)$ and $td'(i,j)$ and two sets of topological indexes^{24,77} T of the vertices are evaluated for both graphs R and R' . These indexes quantify the binding patterns around each point by means of the characteristic sums of the distribution of distances around it:

$$T_V(i) = \sum_{j \neq i} \frac{nv_i \cdot nv_j}{td(i,j)^3} \quad (1A)$$

$$T_N(i) = \sum_{j \neq i} \frac{nn_i \cdot nn_j}{td(i,j)^3} \quad (2A)$$

They differ by the fact that T_V takes into account the total number of valences nv_i fulfilled by an atom i with its neighbors, while T_N only counts the number of neighbors nn_i bound to this atom. Therefore, a first trial to establish an exact match is done using T_V , which is sensitive to the presence of unsaturations in a fragment.

T_V or T_N can be used to define for each vertex $i \in R$ a subset of possible correspondents in R' after discarding the j that are nonequivalent to i . Such subsets $R'_i = \{j \in R' \mid T_V(j) = T_V(i)\}$ may however contain unequivalent atoms for which the equality $T_V(j) = T_V(i)$ is fortuitous.

R and R' must decompose in equal numbers of subsets of corresponding sizes, e.g. the same "spectrum" of values must appear in both T_V and T_V'

vectors for the graphs to be isomorphous. Sorting these vectors must lead to exactly the same list of values or else the routine stops.

- *Vertex-to-vertex assignment*: The former condition being fulfilled, the last remaining step is to make an unambiguous assignment of the pairs of corresponding atoms ($i, j \in R'_i$).

In general, the graphs contain degenerate vertices and any one of the atoms $i_1, i_2, \dots, i_k \in R$ with $T_v(i_1) = T_v(i_2) = \dots = T_v(i_k)$ may be put in relation to any one of the k vertices $j \in R'_i$ that have the same T_v value. Therefore, any one of these j can be picked up at random in order to establish the first pair - for example (i_1, j_1) . Once such an assignment has been done, the degeneracy of vertices can be partially or totally lifted.

- *Backtracking routine*: The unambiguous assignment of atoms is performed by a backtracking routine. The algorithm is exemplified in Fig. A-1. It makes a guess for an assignment of the atom i within the subset R'_i . This assignment may be in *conflict* with the previous ones:

If a pair of image atoms from R' is found to be separated by a topological distance that is different from the one between the corresponding atoms in R , the routine will undo the assignment, mark it as impossible and explore the still open alternatives.

If, on the contrary, this assignment is correct, the routine will temporarily accept it and go on to find the correspondence of atom $i+1$.

If no valid assignment could be found for i given the previous assignments of $1, 2, \dots, i-1$, the routine has reached a local dead end and therefore will take a step back, undoing the current assignment of $i-1$ and trying another, not yet explored possibility for the latter. If none of the possible assignments for $i-1$ ensure a future successful assignment for i , it is the attribution of $i-2$ that needs to be revised, and so on.

If the routine has explored all the branches of possible assignments, without finding a conflict-free graph-to-graph mapping, the graphs are not isomorphous despite the equal T_V spectra.

- *Relaxing the criteria of equivalence if a match failed:* If no matching graph has been found in the existing data base using the T_V criterion, a search based on the T_N values will be redone. Ignoring the bond multiplicities might return a distorted, but not unreasonable geometry, that can easily be repaired by the following energy minimization step. This default solution allows important savings in the size of the ring data base.

After an isomorphism ($i, j \in R'_i$) has been established, the coordinates of the atoms j from the ring in the data base will be assigned to the corresponding vertices i of the ring to build. Multiple geometries may be specified for a ring system and all the corresponding conformers will be generated.

b). *Joining the fragments.* When attaching a fragment to an atom of a ring system having two free valences, there are certain stereochemical aspects to be taken into account. This program considers that if the size of the ring system is larger than 5, these two binding directions are sterically

unequivalent. In the case of a 3, 4 or 5 membered ring, the free valences of an atom are labeled as unequivalent whenever the ring carries more than one substituents, in order to generate all the possible cis/trans diastereomers. A list of possible junction modes is established and all the NJ corresponding geometries are generated, where NJ is the product of the number of stereochemically unequivalent free valences of each fragment times the number of retrieved ring geometries of each cyclic fragment.

c) Torsion angle driving: All exocyclic non-terminal bonds are possible torsional axes. If their number exceeds an imposed threshold (of 10), only the most central axes, which are the most relevant to the global molecular geometry, will be considered for the driving procedure. The driving of double bonds and amide bonds may be toggled on or off (they are set into trans conformations by default). The driving step is taken 120° for torsions of single bonds and 180° for double or partially double bonds. Each one of the NJ geometries obtained after joining the fragments are subjected to the torsion angle driving, so that the maximal number of explored conformations equals $N = 3^{10} \cdot NJ \approx 59000 \cdot NJ$.

d) Selection of the conformations. If the minimal distance between nonbonded heavy atoms d_{\min} is lower than 2 \AA , the conformation is discarded. Conformations that successfully pass the bump check will be compared to the previously accepted ones to assess whether they are different enough to be kept. The set of accepted conformations is binned in function of d_{\min} . The members of a bin b are geometries for which $b \cdot \epsilon \leq d_{\min} < (b+1) \cdot \epsilon$ where ϵ is typically chosen 0.1 \AA . Only conformations within the same

bin need to be compared to each other, which allows an important reduction of the otherwise huge number of N^2 tests. Two criteria: the maximal distance between two nonbonded atoms d_{\max} and a topographic descriptor T_E ⁷⁸ are used to verify the interconformational similarity. In equation (3A), e represent the Pauling electronegativity values of the atoms i and j separated by the distance $d(i,j)$.

$$T_E = \sum_{i \neq j} \frac{e_i \cdot e_j}{d(i,j)^2} \quad (3A)$$

If a previously accepted conformation has both d_{\max} and T_E values close to the tested one, the latter will be discarded. Consequently, only one of the two enantiomers of a chiral conformation will be present in the final set.

Eventually, an imposed number of accepted conformations will be written on disk after having appended the hydrogen atoms, priority being given to the most "extended" geometries with larger d_{\min} . A set of chemical context-dependent rules ensures the correct appending of the hydrogen atoms on ionizable groups according to an input pH value.

Appendix B: Fitting of the parameters of the model.

The whole optimization cycle is outlined in Fig. B-1. Given the unavailability of analytical derivatives and the multitude of local minima that correspond to fitting artifacts, simple but robust minimization procedures have been used.

A Monte Carlo procedure chooses points of the (5-dimensional) docking parameter space. Starting from the a reasonable guess for the values of the docking parameters, it performs a random step. This step length is chosen at the beginning in order to allow the exploration of the whole docking parameter space. Once a new point for a quintuplet of docking parameter values has been chosen, these values are now used to dock all the ligands in the learning set, as previously described, leading to a current list of ΔH , ΔH^* and T ΔS indexes for the whole series of ligands.

On the basis of this list, the optimal values of the affinity coefficients can be straightforwardly found for the current configuration of the docking parameters, by means of a *cross-validated stepwise regression* relating the experimental $\ln K$ values to the corresponding descriptive variables ΔH , ΔH^* and T ΔS .

Stepwise regression: In order to find out which one of the binding indexes (and the free intercept) are relevant to our model, we consider all the $2^4 - 1$ possible combinations of including or letting out the explaining variables. For each one of these combinations, the *cross-validated* RMS and r^2 is

evaluated in function of the included variables. A combination will be retained only if it improves the cross-validated correlation coefficient by at least 0.05 with respect to the best possible combination involving one variable less.

Cross-validation: The classical "jack-knife" procedure widely used in QSAR software⁷⁹ has been used. The cross-validated RMS and r^2 values reported here were obtained by leaving *one* point out at once. The *cross-validated r^2* value of the *optimal* combination of included and ignored explaining variables, an implicit function of both docking and affinity parameters, is the overall quality score of our docking model, the target function to be maximized by the fitting algorithm.

After this target function was evaluated and found to be an improvement with respect to its previous value, the current quintuplet of docking parameters is accepted, otherwise it is rejected. The Monte Carlo procedure continues with the next step.

If within the next 20 steps, no higher r^2 configuration was found, the step length is randomly either increased or decreased, with a rising probability of *decreasing* it. In this way, at late stages of the optimization, the MC procedure will gradually focus on the vicinity of the most interesting optimum that was found during the overall search of the whole space, until the step length becomes inferior to a threshold value for which the optimization stops.

Acknowledgments:

D.H. acknowledges financial support from the Regional Council of the Nord-Pas de Calais District of France. We are indebted to Glaxo/Wellcome France for having prepared the selection of ACD molecules and granted the permit to use the structures of Glaxo-Wellcome compounds in this study. Valérie Lucas is acknowledged for having performed the inhibition tests of the selected compounds. Dr. Luise Krauth-Siegel (EMBL Heidelberg) is acknowledged for having provided the X-ray crystallographic results on the binding of mepacrine prior to their publication. Many thanks to André Tartar (Head of the SCBM Lab., Institut Pasteur de Lille), Christian Sergheraert, Elisabeth Davioud (enzymology team of SCBM), Daniel van Belle (UCMB Lab., ULB Bruxelles), Guy Lippens (NMR Lab, Institut Pasteur de Lille), Eric Buisine (Molecular Modeling, Institut Pasteur de Lille), Mircea Diudea (Chemistry Dept., University of Cluj/Romania) for encouragements and helpful discussions.

References:

- ¹ De Castro, S.L. The challenge of Chagas' disease chemotherapy: an update of drugs assayed against *Trypanosoma Cruzi* *Acta Trop.* **1993**,53, 83-98
- ² Etah, E.A.O; Smith, K.; Fairlamb, A.H. "Trypanothione detoxication systems in trypanosomatids", Spring Meeting of the British Society for Parasitology, London **1993**
- ³ Cotgreave, I.A.; Moldeus, P.; Orrenius, S. Host biochemical defense mechanisms against oxidants. *Ann. Rev. Pharmacol. Toxicol.* , **1988**, 28, 189-212
- ⁴ Janes, W.; Schulz, G.E. Role of the charged groups of glutathione reductase in the catalysis of glutathione reductase: crystallographic and kinetic studies with syntethic analogues. *Biochemistry*, **1990**, 29, 4022-4030
- ⁵ Jockers-Scheruebl, M.C.; Schirmer, R.H.; Krauth-Siegel, R.L. Trypanothione Reductase from *Trypanosoma Cruzi*, catalytic properties of the enzyme and inhibition studies with trypanocidal compounds. *Eur. J. Biochem.*, **1989**, 180, 267-272
- ⁶ Pai, E.F.; Schulz, G.E. The catalytic mechanism of glutathione reductase as derived from X-ray diffraction analyses of reaction intermediates. *J. Biol. Chem.*, **1983**, 258, 1752-1757

⁷ Karplus, P.A.; Schulz, G.E. Refined structure of glutathione reductase at 1.54 Å resolution. *J.Mol.Biol.*, **1987**, *195*, 701-729

⁸ Karplus, P.A.; Schulz, G.E. Substrate binding and catalysis by glutathione reductase as derived from refined enzyme-substrate crystal structures at 2 Å resolution. *J.Mol.Biol.*, **1989**, *210*, 163-180

⁹ Kuriyan, J.; Kong, X.-P.; Krishna, T.S.R.; Sweet, R.M.; Murgolo, N.J.; Field, H.; Cerami, A.; Henderson, G.B. X-ray structure of trypanothione reductase from *Crithidia fasciculata* at 2.4 Å resolution. *Proc. Natl. Acad.Sci.USA* **1991**, *88*, 8764-8768

¹⁰ Hunter, W.N.; Bailey, S.; Habash, J.; Harrop, S.J.; Helliwell, J.R.; Aboagye-Kwarteng, T.; Smith, K.; Fairlamb, A.H. Active site of trypanothione reductase, a target for rational drug design *J.Mol.Biol.*, **1992**, *227*, 322-333

¹¹ Bailey, S.; Smith, K.; Fairlamb, A.H.; Hunter, W.N. Substrate interactions between trypanothione reductase and N¹-glutathionylspermidine disulphide at 0.28 nm resolution *Eur. J. Biochem.*, **1993**, *213*, 67-75

¹² Benson, T.J.; McKie, J.H.; Garforth, J.; Borges, A.; Fairlamb, A.H.; Douglas, K.T. Rationally designed selective inhibitors of trypanothione reductase; phenothiazines and related tricyclics as lead structures. *Biochem. J.*, **1992**, *286*, 9-11

¹³ Burns, R.F.; Simmons, R.M.A.; Howbert, J.J.; Waters, D.C.; Threlkeld, P.C.; Gitter, B.D. "Virtual screening as a tool for evaluating chemical libraries",

lecture at "Exploring Molecular Diversity - Small Molecule Libraries For Drug Discovery", Jan. 23-25, 1995, La Jolla, California, USA.

¹⁴ Smelie, A.S.; Crippen, G.M.; Richards, W.G.; Fast drug-receptor mapping by site-directed distances. A novel method of predicting new pharmacological leads. *J. Chem. Inf. Comp. Sci.* **1991**, *31*, 386-392

¹⁵ Martin, Y.C. 3D database searching in drug design. *J. Med. Chem.* **1992**, *35*, 2145-2154

¹⁶ Eisen, M.B.; Wiley, D.C.; Karplus, M.; Hubbard, R.E. HOOK: a program for finding new molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins*, **1994**, *19*, 199-221

¹⁷ Bohacek, R.S.; McMartin, C. Multiple highly diverse structures complementary to enzyme binding sites: results of extensive application of a *de novo* design method incorporating combinatorial growth. *J. Am. Chem. Soc.* **1994**, *116*, 5560-5571

¹⁸ Boehm, H.J. The computer program LUDI: a new simple method for the *de novo* ligand design. *J. Comp-Aided. Mol. Design*, **1992**, *6*, 61-78

¹⁹ Gillet, V.; Johnson, A.P.; Mata, P.; Sike, S.; Williams, P.; SPROUT: A program for structure generation. *J. Comp. Aided Mol. Design*, **1993**, *7*, 123-157

²⁰ Oprea, T.I.; Waller, C.L.; Marshall, G.R. Three-dimensional quantitative structure-activity relationship of human immunodeficiency virus(I)

protease inhibitors. 2. Predictive power using limited exploration of alternate binding modes. *J. Med. Chem.* **1994**, *37*, 2206-2215

²¹ Sternberg, M.J.; King, R.D.; Lewis, R.A.; Muggelton, S. Application of machine learning to structural molecular biology. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **1994**, *344*, 365-371

²² Benigni, R.; Cotta-Ramusion, M.; Giorgi, F.; Gallo, C. Molecular similarity matrices and quantitative structure-activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629-635

²³ Hocart, S.J.; Reddy, W.; Murphy, W.A.; Coy, D.H. Three-dimensional quantitative structure-activity relationships of somatostatin analogues. 1. Comparative molecular field analysis of growth hormone release-inhibiting potencies. *J. Med. Chem.* **1995**, *38*, 1974-1989

²⁴ Diudea, M.V.; Ivanciuc, O. "Molecular Topology" (in Romanian), Complex Editions **1995**; Cluj-Napoca, Romania

²⁵ Ortiz, A.R. ; Pisabarro, M.T.; Gago, F.; Wade, R.C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38*, 2681-2691

²⁶ Hansch, C. « Drug design », Edited by Ariens, E.J., Academic Press, NY, **1971**, *16*, 271

²⁷ van Gunsteren, W.F. "Methods for calculation of free energies and binding constants. Successes and problems" in 'Computer Simulations of

Biomolecular Systems, W.F. van Gunsteren & S. Wiener, ESCOM, 1989, 27-59

²⁸ Struthers, R.S.; Rivier, J.; Hagler, A.T. "Design of peptide analogs: Theoretical simulation of conformation, energetics and dynamics" in 'Conformationally Directed Drug Design; Peptides and Nucleic Acids as Templates or Targets', J.A. Vida & M. Gordon, ACS, Washington, 1984, 239-261

²⁹ McQuarrie, D.A. "Statistical Mechanics", Harper Collins Publishers, 1976

³⁰ Mezei, M.; Beveridge, D. Free energy simulations *Ann. NY Acad. of Sci.*, 1986, 482, 1-23

³¹ Mezei, M. The finite difference thermodynamic integration, tested on calculating the hydration free energy difference between acetone and dimethylamine in water. *J.Chem.Phys.*, 1987, 86, 7084-7088

³² Hodel, A.; Rice, L.M.; Simonson, T.; Fox, R.O.; Brunger, A.T. Proline cis-trans isomerization in staphylococcal nuclease: multi-substrate free energy perturbation calculations. *Protein Sci.* 1995, 4, 636-654

³³ Stouten, P.F.W.; Froemmel, C.; Nakamura, H.; Sander, C. An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. Simul.*, 1993, 10, 97-120

³⁴ Luty, B.A.; Zacharias, M.; Wasserman, Z.R.; Stouten, P.F.W.; Hodge, C.N.; McCammon, J.A. A molecular mechanics/grid method for evaluation of

³⁵ Baillet, S.; Buisine, E.; Horvath, D.; Maes, L.; Bonnet, B.; Sergheraert, C. 2-Aminodiphenylsulphides as inhibitors of trypanothione reductase. *Bioorg. Med. Chem.*, **1996**, *4*, 891-899

³⁶ Vajda, S.; Weng, Z.; Rosenfeld, R.; DeLisi, C. Effect of Conformational Flexibility and Solvation on Receptor-Ligand Binding Free Energies. *Biochemistry*, **1994**, *33*, 13977-13988

³⁷ Moutiez, M.; Lucas, V.; Davioud, E.; Tartar, A.; Sergheraert, C. "Indolylmaleinimide derivatives: synthesis and biological activities as new potential Trypanothione Reductase inhibitors" poster at COST-ACRIVAL, Grenada, Spain 1995

³⁸ Moutiez, M.; *Ph.D. Thesis*, University of Lille II, France 1995

³⁹ Krauth-Siegel, R.L.; Lohrer, H.; Buecheler, U.S.; Schirmer, R.H. in "Biochemical Protozoology" (G.H. Coombe & M.J. North, eds.), Taylor and Francis, London 1991

⁴⁰ Gomez, R.F.; Moutiez, M.; Aumercier, M.; Bethegnies, G.; Luyckx, M.; Ouaisi, A.; Tartar, A.; Sergheraert, C. 2-Aminodiphenylsulphides as new inhibitors of trypanothione reductase. *Int. J. Antimicrob. Agents*, **1995**, *6*, 111-118

⁴¹ Ponasik, J.A.; Strickland, C.; Faerman, C.; Savvides, S.; Karplus, P.A.; Ganem, B.; Kukoamine A and other hydrophobic acylpolyamines: potent

and selective inhibitors of *Crithidia fasciculata* trypanothione reductase.

Biochem. J., **1995**, *311*, 371-375

⁴² Girault, S.; Baillet, S.; Lucas, V.; Davioud, E.; Tartar A.; Sergheraert, C. 'Bis - Aminodiphenylsulphides as potent inhibitors of the TR from *T. Cruzi*' - poster at the 5-th European COST Conference on Antiparasitic Chemotherapy, 23-24 May **1996** (Heidelberg)

⁴³ Moreno, S.N.J.; Carnieri, E.G.S.; Docampo, R. Inhibition of *Trypanosoma cruzi* trypanothione reductase by crystal violet. *Mol. Biochem. Parasitol.* **1994**, *67*, 313-320

⁴⁴ Howard, A.E.; Kollman, P.A. An analysis of current methodologies for conformational search of complex molecules, *J. Med. Chem.* **1988**, *31*, 1669-1675

⁴⁵ Leach, A.R. An algorithm to identify a molecule's « most different » conformations. *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 661-670

⁴⁶ Leach, A.R.; Prout, K.; Dolata, D.P. The application of artificial intelligence to the conformational analysis of strained molecules. *J. Comp. Chem.* **1990**, *11*, 680-693

⁴⁷ Discover 2.9.0/3.1.0 User Guide, Jan. **1993**, Biosym Technologies, San Diego, CA.

⁴⁸ Smellie, A.; Teig, S.L.; Towbin, P. Poling: promoting conformational variation. *J. Comp. Chem.*, **1995**, *16*, 171-187

-
- ⁴⁹ Ermer, O. Calculation of molecular properties using force fields. Applications in organic chemistry. *Structure and Bonding* **1976**, *27*, 161-211
- ⁵⁰ Hagler, A.T.; Lifson, S.; Dauber, P.; Consistent force field studies of intermolecular forces in hydrogen bonded crystals. II. A benchmark for the objective comparison of alternative force fields. *J. Am. Chem. Soc.* **1979**, *101*, 5122-5130
- ⁵¹ Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T. "Numerical Recipes: The Art of Scientific Computing" Cambridge University Press, Cambridge **1986**
- ⁵² Albert, A.; Serjeant, E.P. "Ionization Constants of Acids & Bases", Butler & Tanner, Frome and London **1962**
- ⁵³ Protein Data Bank; Chemistry Department, Brookhaven National Laboratory, Upton NY 11973 USA.
- ⁵⁴ Gilson, M.K.; Honig, B. The inclusion of electrostatic hydration energies in molecular mechanics calculations. *J. Comp-Aided. Mol. Design* **1991**, *5*, 5-20
- ⁵⁵ Still, W.C.; Tempczyk, A.; Hawley, R.C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127-6129
- ⁵⁶ Eisenberg, D.; McLahan, A.D. Solvation energy in protein folding and binding *Nature*, **1986**, *319*, 199-203

⁵⁷ Mehler, E.L.; Eichele, G. Electrostatic effects in water-accessible regions of proteins. *Biochemistry*, **1984**, *23*, 3887-3891

⁵⁸ Jackson, J.D. "*Classical Electrodynamics*", J.Wiley & Sons, New York 1975

⁵⁹ Horvath, D.; van Belle, D.; Lippens, G.; Developement and parametrization of continuum solvent models. II. An unified approach to the solvation problem. *J.Chem.Phys. in press* **1996**

⁶⁰ Brooks, C.L.III; Karplus, M.; Montgomery-Pettitt, B., "*Proteins: A theoretical Perspective of Dynamics, Structure and Thermodynamics*", Advances in Chemical Physics vol. LXXI, John Wiley & Sons **1988**

⁶¹ Brooks, C.L. III, Thermodynamics of ionic solvation: Monte Carlo simulation of aqueous chloride and bromide ions. *J. Phys. Chem.* **1986**, *90*, 6680-6684

⁶² ACD - Available Chemicals Directory, Copyright **1995** - Molecular Design Limited

⁶³ Aumercier, M.; Meziane-Cherif, D.; Moutiez, M.; Tartar, A.; Segheraert, C. A microplate assay to screen trypanothione reductase inhibitors. *Analytical Biochemistry* **1994**, *223*, 161-164

⁶⁴ Horvath, D.; van Belle, D.; Lippens, G.; Wodak, S.J. Development and parametrization of continuum solvent models. I. Models based on the boundary element method. *J.Chem.Phys*) **1996**, *104*, 6679-6695

⁶⁵ Searle, M.S.; Williams, D.H.; Gerhard, U.; Partitioning of Free Energy Contributions in the Estimation of Binding Constants: Residual Motions and Consequences for Amide-Amine Hydrogen Bond Strengths. *J. Am. Chem. Soc.*, **1992**, *114*, 10697-10704

⁶⁶ Stewart, J.J. MOPAC: a semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1-105

⁶⁷ Head, R.D.; Smythe, M.L.; Oprea, T.I.; Waller, C.L.; Green, S.M.; Marshall, G.R. VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. *J. Am. Chem. Soc.*, **1996**, *118*, 3959-3969

⁶⁸ Boehm, H.J.; The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Design*, **1994**, *8*, 243-256

⁶⁹ Pickett, S.E.; Sterberg, M.E.; Empirical scale of conformational entropy in protein folding. *J. Mol. Biol.*, **1993**, *231*, 2674-2684

⁷⁰ Zoellner, H.; « *Handbook of Enzyme Inhibitors, Part A.* », VCH Verlag GmbH, Weinheim **1993**

⁷¹ Jacoby, E.M.; Schlichting, I.; Lantwin, C.B.; Kabsch, W.; Krauth-Siegel, R.L. Crystal structure of the Trypanosoma Cruzi trypanothione reductase - mepacrine complex. *Proteins*, **1996**, *24*, 73-80

⁷² Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A.A.; Aflalo, C.; Vakser, I.A.; Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA.*, **1992**, *89*, 2195-2199

⁷³ Harrison, R.W.; Kourinov, I.V.; Andrews, L.C.; The Fourier-Green's function and the rapid evaluation of molecular potentials. *Protein Eng.*, **1994**, *7*, 359-369

⁷⁴ So, S.S.; Karplus, M.; Evolutionary optimization in quantitative structure-activity relationship: an application of neural networks. *J. Med. Chem.* **1996**, *39*, 1521-1530

⁷⁵ Cavin, J.C.; Krassner, S.M.; Rodriguez, E. Plant-derived alkaloids active against *Trypanoma Cruzi*. *J. of Ethnopharmacol.* **1987**, *19*, 89

⁷⁶ Ivanciuc, O.; Balaban, T.S.; Balaban, A.T. Chemical graphs with degenerate topological indices based on information of distance. *J. Math. Chem.*, **1993**, *12*, 21-31

⁷⁷ Diudea, M.V.; Horvath, D.; Bonchev, D. MOLORD algorithm and real number subgraph invariants. *Croat. Chem. Acta* **1995**, *68*, 131-148

⁷⁸ Diudea, M.V.; Horvath, D.; Graovac, A.; 3D distance matrices and related topological indices. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 129-135

⁷⁹ *TSAR User's Guide*, Issue 3, Oxford Molecular Ltd. **1993**

Table Legends

Table 1: Structures used for the calibration and the test of the docking model: Nr. - current numbering of the molecules; Ref - molecular structure from in Fig. 1; R1-R5 - substituents in the corresponding positions, if the case; n - length of the hydrophobic spacer chain, if the case; ACT- natural log of the value (in $\mu\text{mol/l}$) of the inhibition constant according to the corresponding bibliographic reference ; Set. - presence of the molecule in the learning sets of different calibration trials (see Table 4).

Table 2: Correlations between the solvation energies from Boundary Element calculations E^{BEM} and the Gilson-Honig term E^{GH} for the sets of docked conformations of different compounds. Columns 2-4 refer to the unconstrained regression equations $E^{\text{BEM}} = y'_0 + \lambda' \cdot E^{\text{GH}}$, where both the intercept y_0 and the RMS are in kcal/mol. Columns 5 and 6 display the correlation parameters at an imposed λ equal to the value obtained at the calibration step of the docking model.

Table 3: The efficiency of our conformational sampling algorithm compared to a 100 ps Molecular Dynamics run at 1000 K. The energy differences are in kcal/mol. ΔE_{min} is the difference between the vacuum energies of the best minimum found by our method and the one obtained by optimization of MD-sampled geometries. $\Delta(E_{\text{min}} + E_s)$ is the corresponding difference between the solvent-corrected energies. N_{init} represents the number of initial geometries that were subjected to minimization, while N and N_{MD} are the number of distinct minima found by our method and respectively by the MD simulation.

Table 4: The "docking" parameters obtained from 3 different calibrations denoted by A,B and C, using different learning sets. The hydrophobicity parameter η has been expressed in cal/A² to make its value comparable with the hydrophobic "surface tension" coefficients reported in the literature. All other parameters are dimensionless weighting factors. *) - has been kept fixed during this calibration run.

Table 5: Established relations between the (natural) log of the inhibition constant and the calculated binding indexes are shown for each of the parametrization schemes from Table 4. The BEST relations are the ones found by stepwise regression, retaining only the variables that are relevant for the model. Alternative equations only in ΔH and $T\Delta S$ are also shown. The *cross-validated* RMS values refer to the calculated vs. experimental values of $\ln K$.

Table 6: Checking whether the binding entropy index as evaluated by equation (20) can be written as a linear combination of ΔH and ΔH^* . A,B and C refer to the different parametrization schemes listed in Table 4. The established linear relationships and the cross-validated correlation coefficients are shown.

Table 7: The correlation coefficients r^2 between the sets of binding indexes ΔH and $T\Delta S$ obtained with the three different parametrizations listed in Table 4. These r^2 result from linear regression calculations with fixed *unitary* slope and *null* intercept ($y=x$).

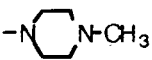
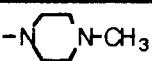
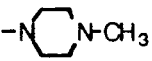
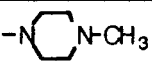
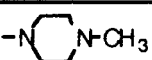
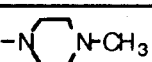
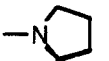
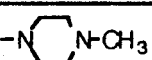
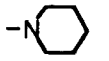
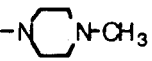
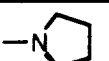
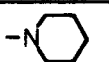
Table 8: The predictive power of the obtained models. "All ligands" refers to the full set of molecules shown in Table 1. Separate statistics are shown for the subsets of well predicted compounds (for which the predicted affinity constants were of the correct order of magnitude, e.g. $\Delta \ln K < 2$) and for the learning set

molecules. Nr - number of compounds in that subset; RMS - root mean squared error between predicted and experimental $\ln K$ values; r^2 - the corresponding correlation coefficients according to equation (22).

* Entry D. represents a further trial to refine the parameters by adding the outliers of model C to the learning set. Despite extensive MC search, no better parameter set has been found.

Table 9: The predicted inhibition constants and the measured percentages of inhibition for the 13 « virtual hits » obtained with the parameter set A of Table 4. Structure refers to the labels of these compounds (Fig. 7). The percentages of inhibition are given as $(r_0 - r_{inh})/r_0 \times 100\%$, where r_0 and r_{inh} are the rates of the reduction of trypanothione disulfide by TR, in absence and in presence of inhibitor at the given concentrations. In parallel, the current working reagents and solutions are tested against the reference inhibitor clomipramine (Fig. 1 A, $K_{ref} = 7 \mu\text{mol/l}$) - see entry "Ref".

Table 1

Nr	Ref	R ₁	R ₂	R ₃	R ₄	R ₅	n	ACT.	Set.
1	C.	-	-	-	-	-	-	0.91 ³⁷	all
2	G.	(CH ₂) ₂ NH ₂	(CH ₂) ₂ NH ₂	-	-	-	3	1.50 ³⁵	all
3	A.	-	-	-	-	-	-	1.95 ¹²	all
4	E.		Cl	H	H	H	3	2.71 ³⁸	all
5	D.		Cl	H	COMe	H	3	2.94 ³⁸	all
6	E.	NMe ₂	Cl	H	H	H	3	3.00 ³⁸	all
7	G.	(CH ₂) ₂ NH ₂	H	H	H	H	3	3.00 ³⁵	all
8	B.	-	-	-	-	-	-	3.22 ³⁹	all
9	D.		Cl	H	H	H	3	3.30 ⁴⁰	none
10	D.		Cl	H	Me	H	3	3.33 ³⁸	all
11	D.		Cl	Cl	H	Cl	3	3.40 ³⁸	all
12	D.		Cl	H	H	H	2	3.68 ⁴⁰	all
13	D.	NEt ₂	Cl	H	H	H	3	3.74 ⁴⁰	all
14	D.		Cl	H	H	H	3	3.78 ⁴⁰	none
15	D.		Cl	H	Cl	H	2	3.85 ³⁸	all
16	D.	NMe ₂	Cl	H	H	H	2	3.91 ⁴⁰	all
17	D.		Cl	H	H	H	3	3.95 ⁴⁰	none
18	D.		Cl	H	Cl	H	3	4.04 ³⁸	all
19	D.		Cl	H	Cl	H	3	4.06 ³⁸	none
20	D.		Cl	H	Cl	H	3	4.09 ³⁸	D.

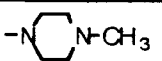
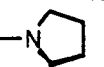
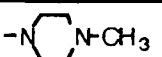
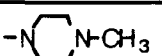
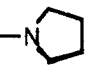
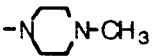
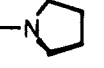
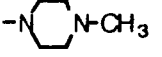
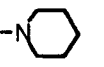
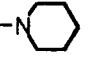
21	D.		H	H	Me	H	3	4.17 ³⁸	all
22	D.		Cl	H	H	H	2	4.34 ³⁸	all
23	D.	NMe ₂	Cl	H	H	H	2	4.38 ³⁸	none
24	D.	NMe ₂	H	H	Cl	H	2	4.38 ³⁸	none
25	D.		H	H	Cl	H	2	4.38 ³⁸	none
26	D.		Me	H	Me	H	3	4.38 ³⁸	all
27	D.		H	H	H	H	3	4.73 ³⁸	all
28	D.		Cl	H	F	H	3	4.74 ³⁸	D.
29	D.		H	H	Cl	H	2	5.08 ³⁸	all
30	D.		Me	H	F	H	3	5.35 ³⁸	none
31	D.	NEt ₂	Cl	H	H	H	2	6.21 ³⁸	all
32	D.		Cl	H	H	H	2	6.55 ³⁸	all
33	D.		Cl	H	Cl	H	2	6.90 ³⁸	all
34	I.	-	-	-	-	-	-	0.59 ⁴¹	C,D.
35	J.	-	-	-	-	-	-	2.01 ⁴¹	B,C,D
36	K.	-	-	-	-	-	-	2.77 ⁴¹	C,D.
37	L.	-	-	-	-	-	-	4.44 ⁴¹	none
38	M.	-	-	-	-	-	-	4.68 ⁴¹	none
39	N.	-	-	-	-	-	-	5.01 ⁴¹	B,C,D
40	O.	-	-	-	-	-	-	5.01 ⁴¹	none
41	P.	-	-	-	-	-	-	5.60 ⁴¹	none
42	Q.	-	-	-	-	-	-	5.82 ⁴¹	B,C,D
43	H.	-	-	-	-	-	-	1.00 ⁴²	none
44	F	-	-	-	-	-	-	1.60 ⁴³	none

Table 2

Comp.	$y'o$	λ'	RMS	r^2	RMS(λ)	$r^2(\lambda)$
16	-33.9	1.19	0.53	0.64	0.84	-
14	-30.2	0.69	0.48	0.39	1.02	-
9	-112.3	2.07	1.11	0.62	1.12	0.61
27	-33.0	1.06	0.57	0.49	0.88	-
8	-38.3	1.88	0.53	0.68	0.57	0.64
7	-101.1	1.88	1.05	0.65	1.10	0.61
2	-204.6	1.88	0.99	0.85	1.14	0.79
1	-113.8	1.51	1.04	0.79	1.55	0.53
6	-38.3	2.32	0.41	0.70	0.41	0.70
23	-110.7	2.07	0.46	0.91	0.50	0.90
4	-122.9	3.14	0.48	0.96	0.80	0.89
3	-35.64	1.51	0.24	0.86	0.31	0.75
13	-33.63	1.48	0.41	0.55	0.52	0.27

Table 3

Comp.	ΔE_{\min}	$\Delta(E_{\min}+E_s)$	N_{init}	N	N_{MD}
1	-1.10	-1.20	326	33	8
2	-4.77	-2.26	304	19	13
7	-1.13	+0.49	426	39	14
4	+0.12	0.00	36	5	6
8	+2.71	-2.15	300	87	56

Table 4

<i>Model</i>	χ	η	λ	ω	Π
A	1.228	3.75	2.66	1.000*	0.208
B	1.149	3.94	2.29	0.956	0.270
C	1.150	3.62	2.31	0.961	0.218

Table 5

	<i>Best relationship</i>	<i>RMS</i>	<i>Relationship in ΔH & $T\Delta S$</i>	<i>RMS</i>
A	$\ln K = 7.07 + 0.18\Delta H - 2.42T\Delta S$	0.78	$\ln K = 7.07 + 0.18\Delta H - 2.42T\Delta S$	0.78
B	$\ln K = 4.47 + 3.26\Delta H - 3.19\Delta H^* - 3.33T\Delta S$	1.01	$\ln K = 5.54 + 0.10\Delta H - 2.04T\Delta S$	1.12
C	$\ln K = 4.31 + 4.76\Delta H - 4.68\Delta H^* - 4.18T\Delta S$	0.95	$\ln K = 5.19 + 0.10\Delta H - 2.35T\Delta S$	1.30

Table 6

<i>Model</i>	<i>Established relationship</i> $T\Delta S = a.\Delta H - a^* \Delta H^*$	r^2 (CV)
A	$T\Delta S = 1.39 \Delta H - 1.37 \Delta H^*$	0.516
B	$T\Delta S = 1.25 \Delta H - 1.24 \Delta H^*$	0.553
C	$T\Delta S = 1.41 \Delta H - 1.40 \Delta H^*$	0.582

Table 7

<i>Index</i>	ΔH		$T\Delta S$	
SET	B	C	B	C
A	0.52	0.92	0.37	0.31
B	1.00	0.78	1.00	0.27

Table 8

Model	All ligands		Ligands with well predicted affinities			Ligands in learning set			Outliers
	RMS	r^2	Nr.	RMS	r^2	Nr.	RMS	r^2	
A	1.29	0.18	38	0.81	0.58	23	0.67	0.79	30,34,39,41, 42,43
B	1.17	0.32	42	0.99	0.46	26	0.90	0.62	30,34
C	1.12	0.38	41	0.85	0.62	29	0.87	0.69	20,28,34
D*	1.12	0.38	41	0.85	0.62	31	1.17	0.40	20,28,34

Table 9

Structure	Inhibition percentage at:			Predicted $K_i(\mu M)$
	57 $\mu mol/l$	28.5 $\mu mol/l$	5.7 $\mu mol/l$	
I.	36	9	0	14
II.	27	13	0	7
III.	7	0	0	14
IV.	31	0	0	12
V.	25	0	0	5
VI.	0	0	0	15
VII.	18	9	0	14
VIII.	0	0	0	2
IX.	0	0	0	5
X.	0	0	0	5
XI.	12	16	18	14
XII.	23	15	10	85
XIII.	40	30	0	3
Ref.	70...80	50..60	-	7(exp.)

Figure Legends

Fig. 1: Generic structures of the molecules used to calibrate and test the docking model. The substituents R_i and the number of carbons of the hydrophobic spacer are given in Table 1.

Fig. 2: The active site of TR from *C.fasciculata*. The centers of the spheres represent the different starting points at which the center of a ligand conformer is placed

Fig. 3: Binding of the ligand B is entropically favored over the ligand A, since more of its bound energy levels are populated.

Fig. 4: The covariance of the two different binding entropy indexes applied in the present study: $T\Delta S$ on the x-axis vs. $(\Delta H - \Delta H^*)$ on the y-axis (in kcal/mol) as resulting from parametrization scheme C (see Table 4).

Fig. 5 A,B,C: The plots of predicted vs experimental logs of inhibition constants for the molecules in Fig. 1. For each set of parameters A, B and C as listed in Table 4, the molecules used in the learning set are plotted with filled squares while the other are shown as triangles. The lines $y=x-2$ and $y=x+2$ on both sides of the diagonal $y=x$ are delimiting the "mispredicted" from the "well predicted" compounds for which the experimental and calculated inhibition constants are of the same order of magnitude.

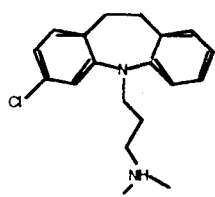
Fig. 6: Predicted binding mode of mepacrine (compound 8) in the TR site according to the docking results obtained with the "docking" parametrization C from Table 4. The conformations within 2 kcal/mol with respect to the best minima (drawn in bold lines) are shown. More of these conformations adopt binding mode 1. CPK spheres have been drawn around the atoms of the site that

are within 5 Å from the inhibitor molecule in its lowest energy conformation of binding mode 1. Specific TR residues are shown in "stick" representation.

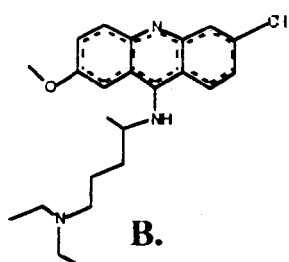
Fig. 7: The structures of the "hits" found while screening a molecular data base for putative TR inhibitors. The predicted $\ln K$ values and the measured percentages of inhibition are listed in Table 9.

Fig. A-1. Finding the graph isomorphisms by a backtracking procedure

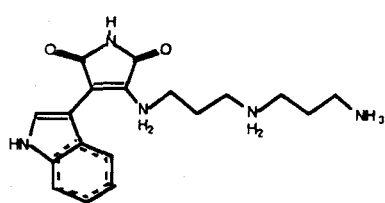
Fig. B-1. The fitting procedure of the parameters of the affinity model.



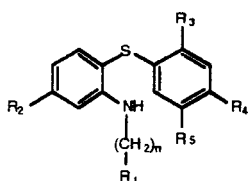
A.



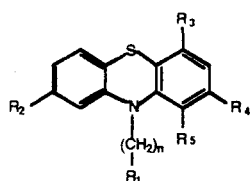
B.



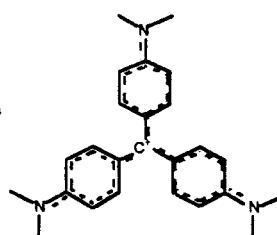
C.



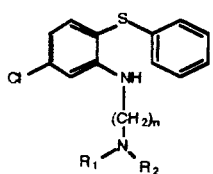
D.



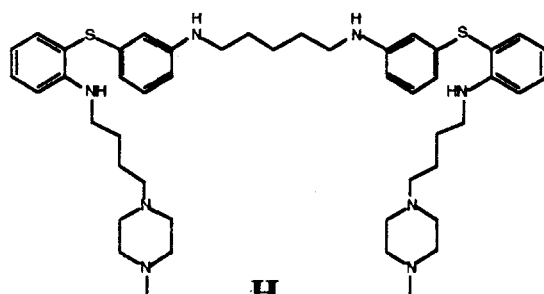
E.



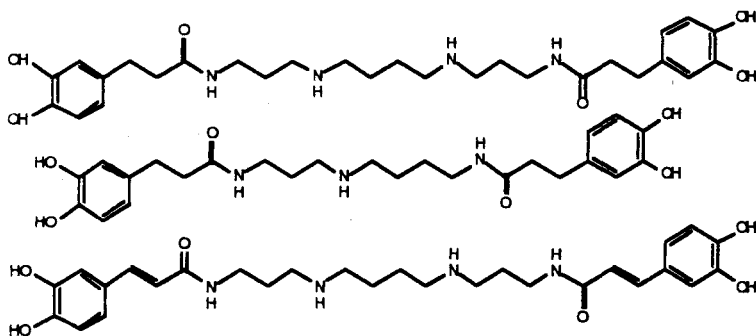
F.



G.



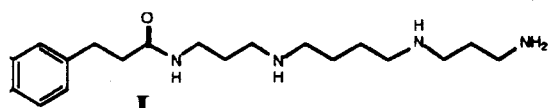
H.



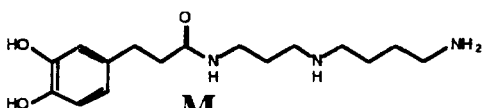
I.

J.

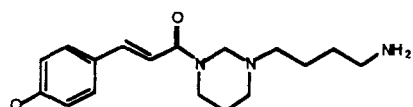
K.



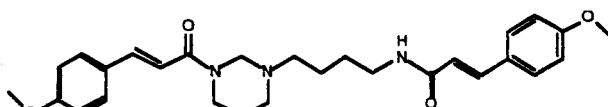
L.



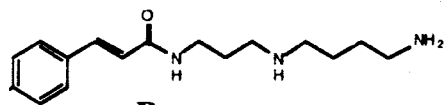
M.



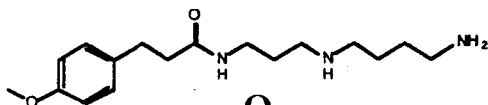
N.



O.



P.



Q.

Fig. 1

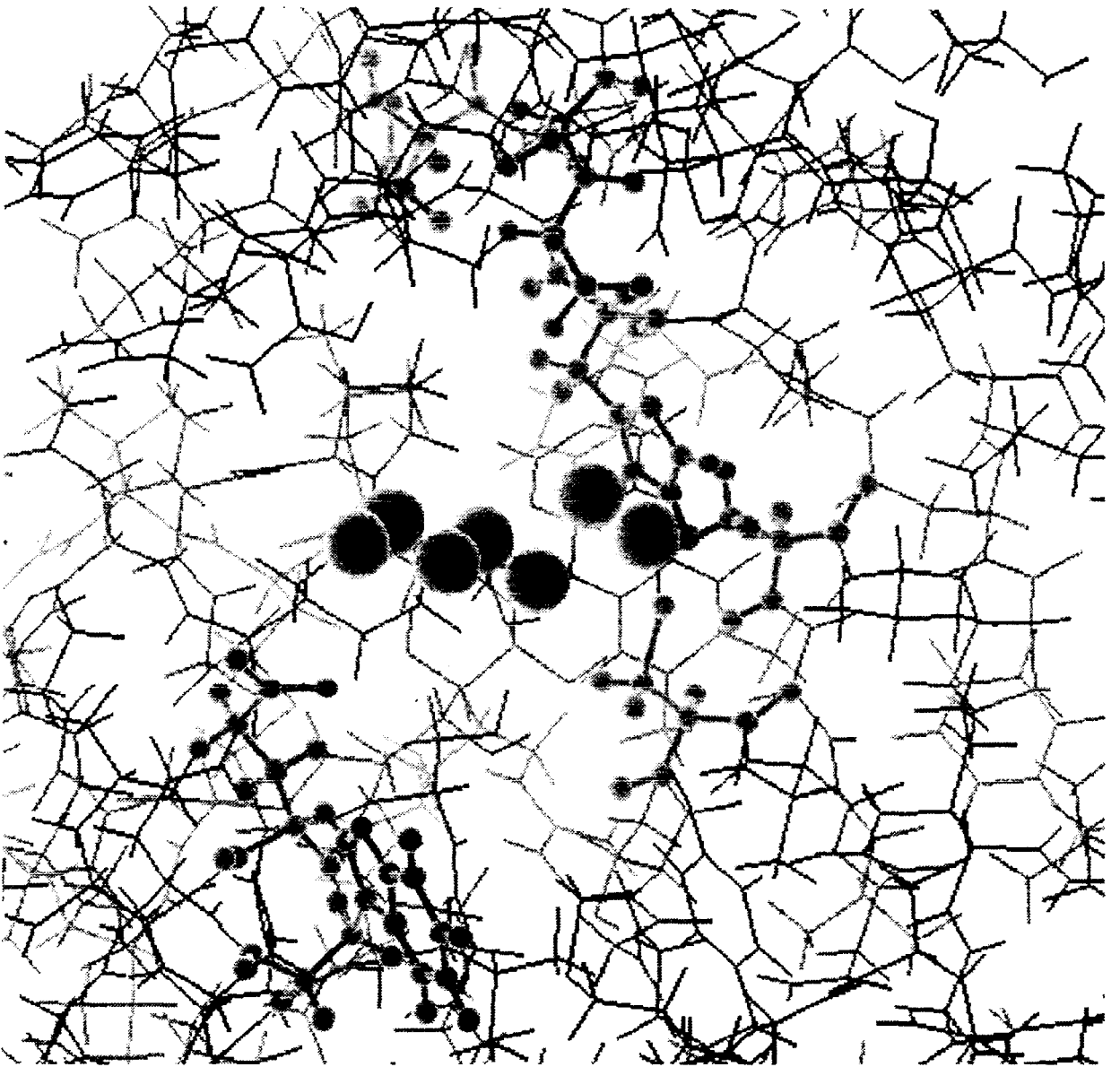


Fig. 2

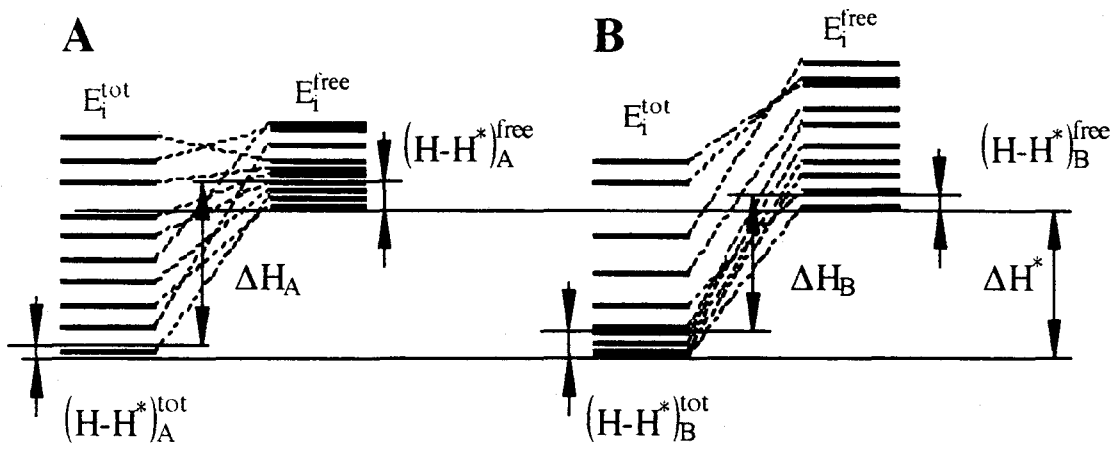


Fig. 3

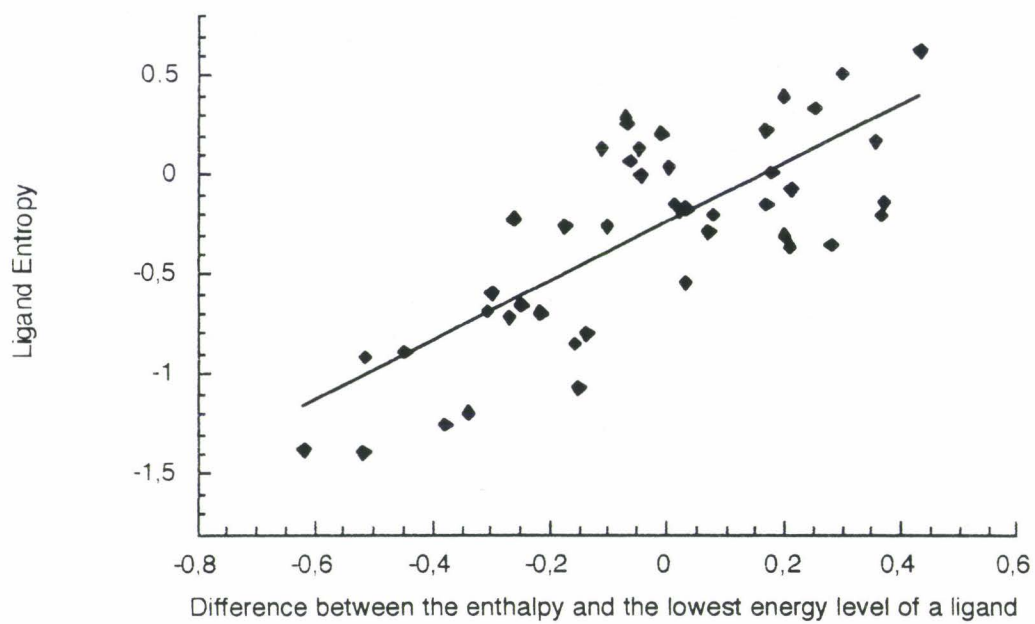


Fig. 4

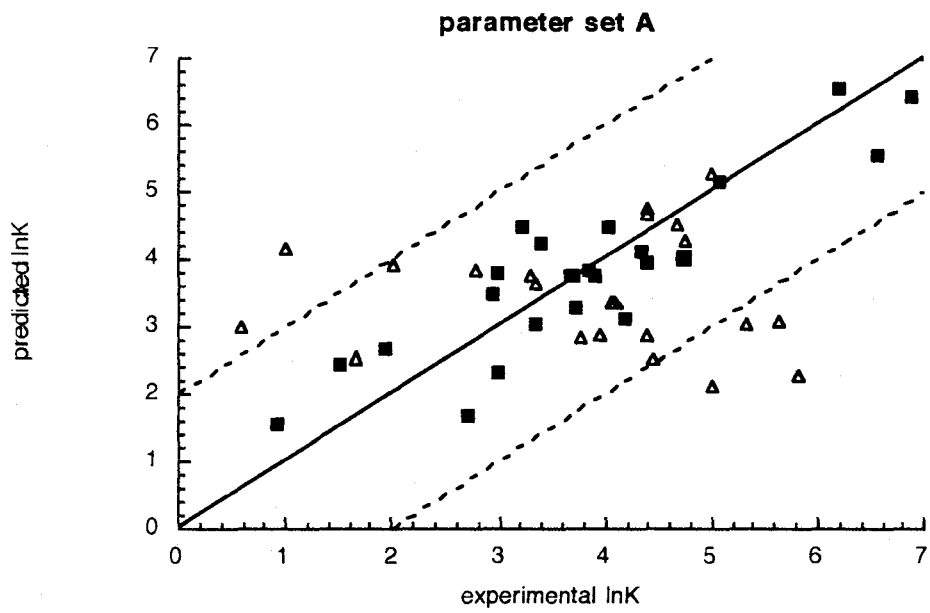


Fig. 5A

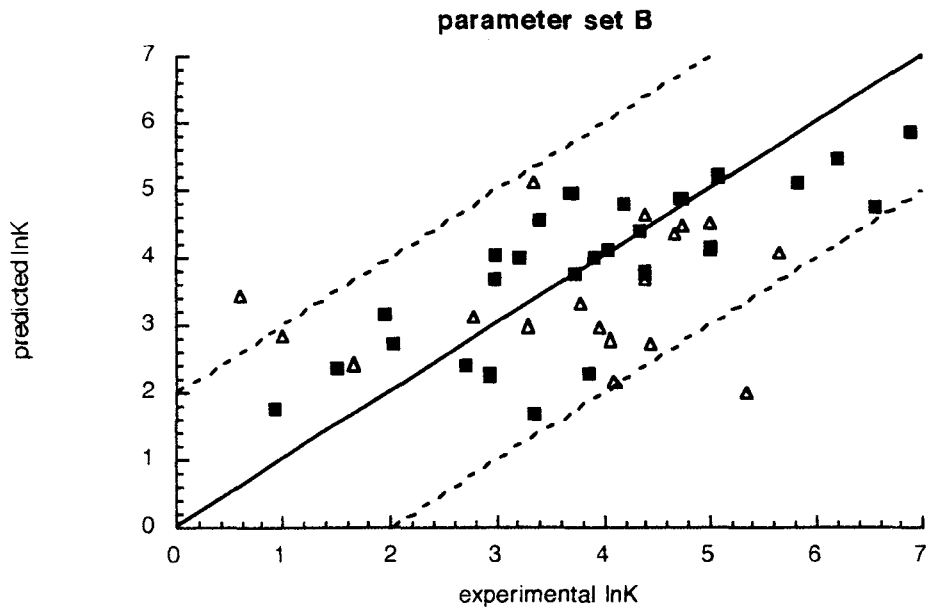


Fig. 5B

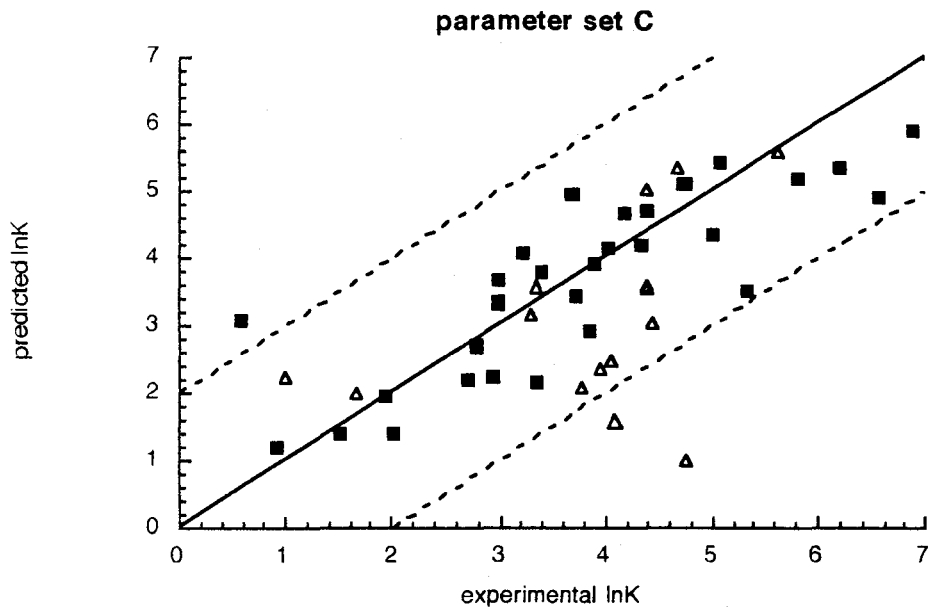


Fig. 5C

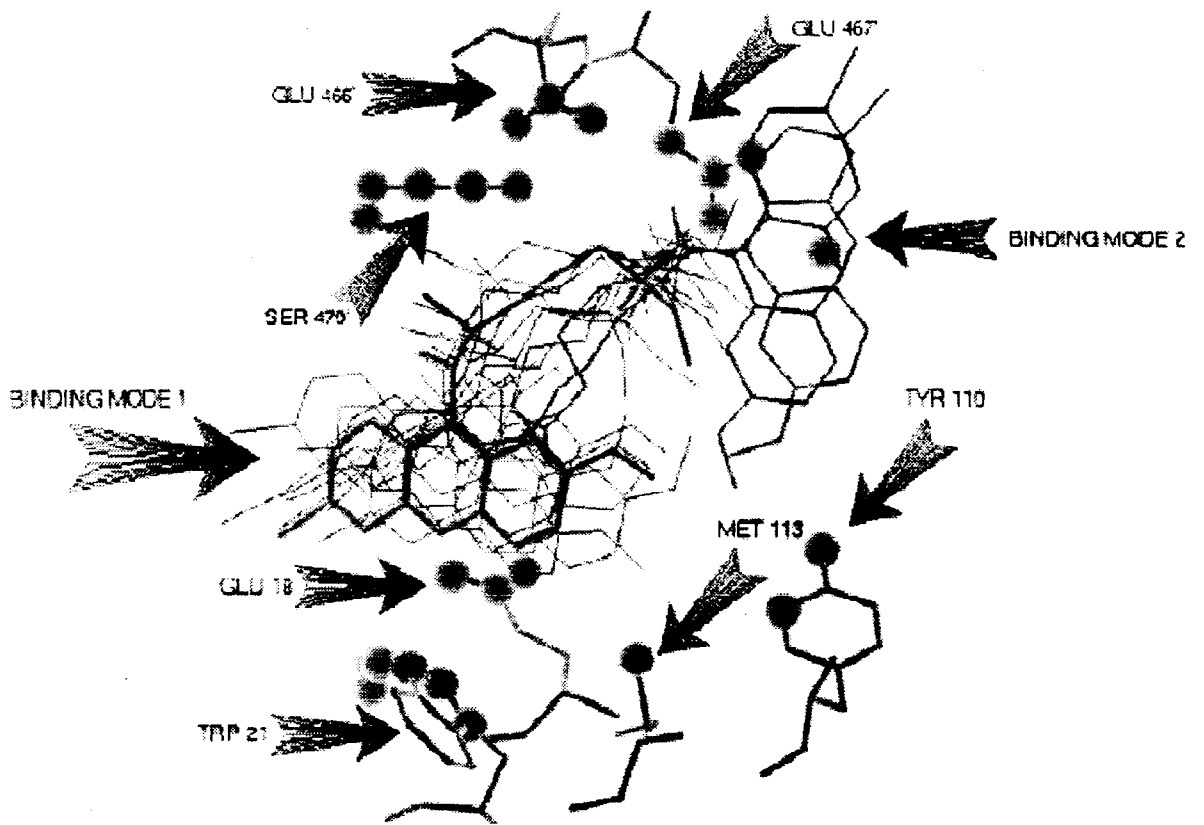


Fig. 6

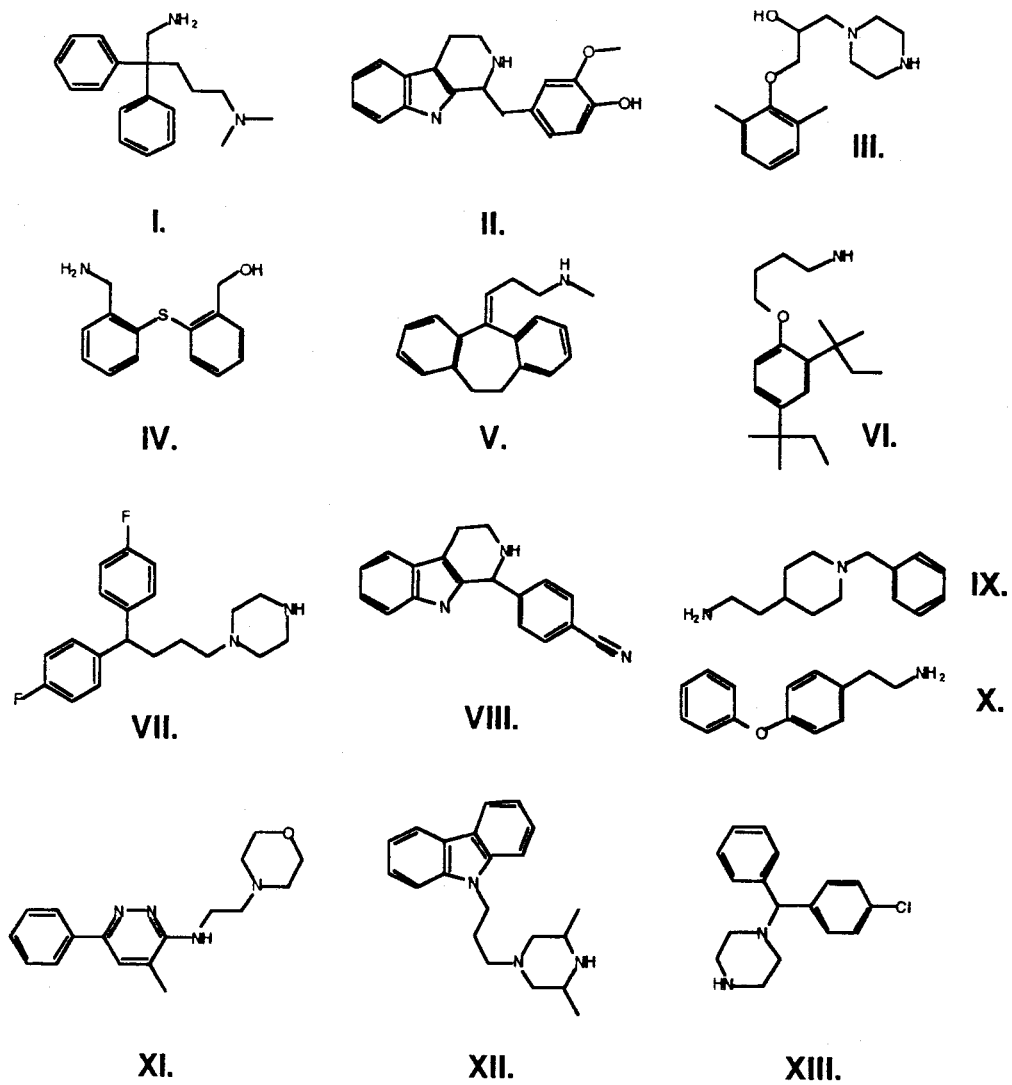


Fig. 7

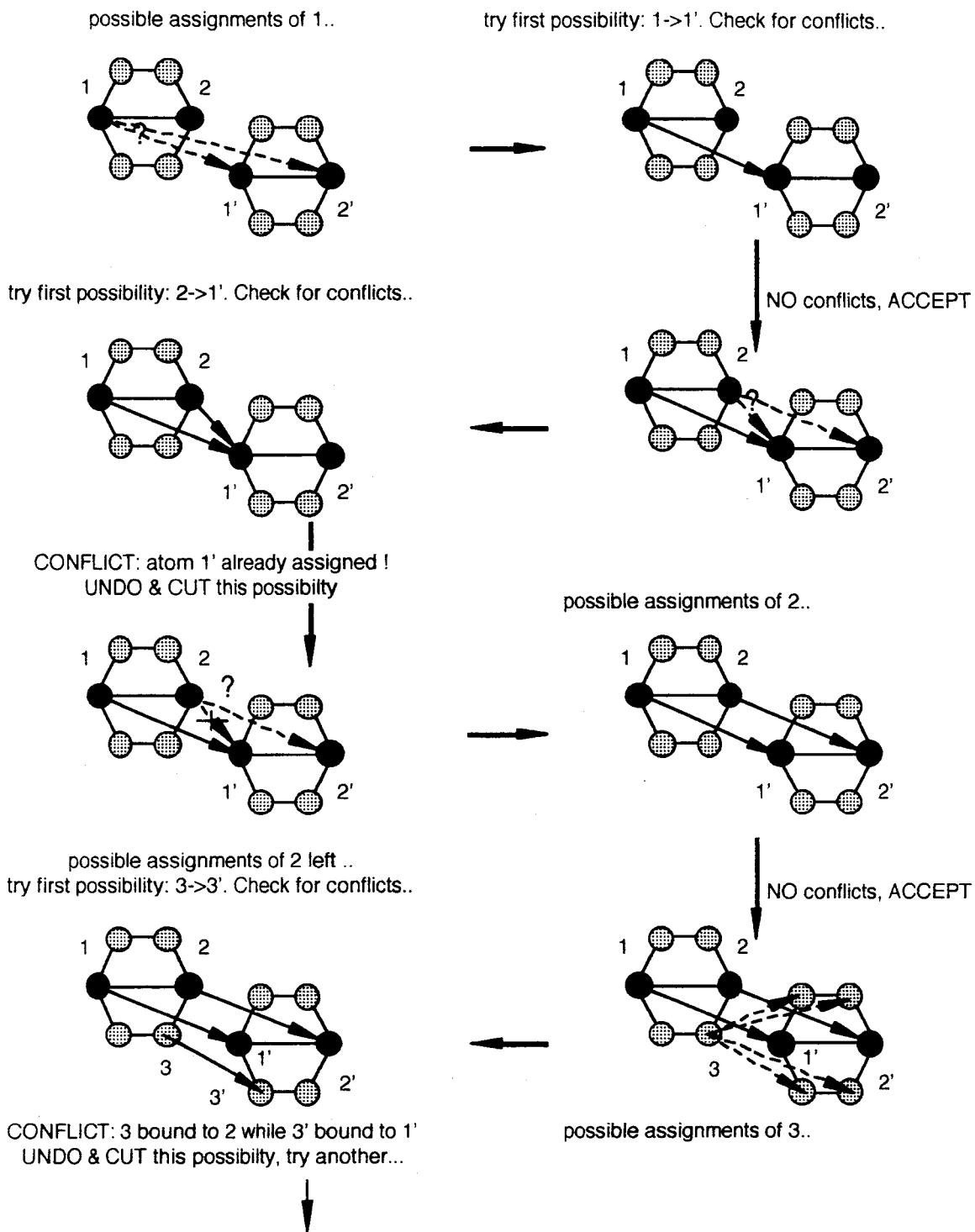


Fig. A-1

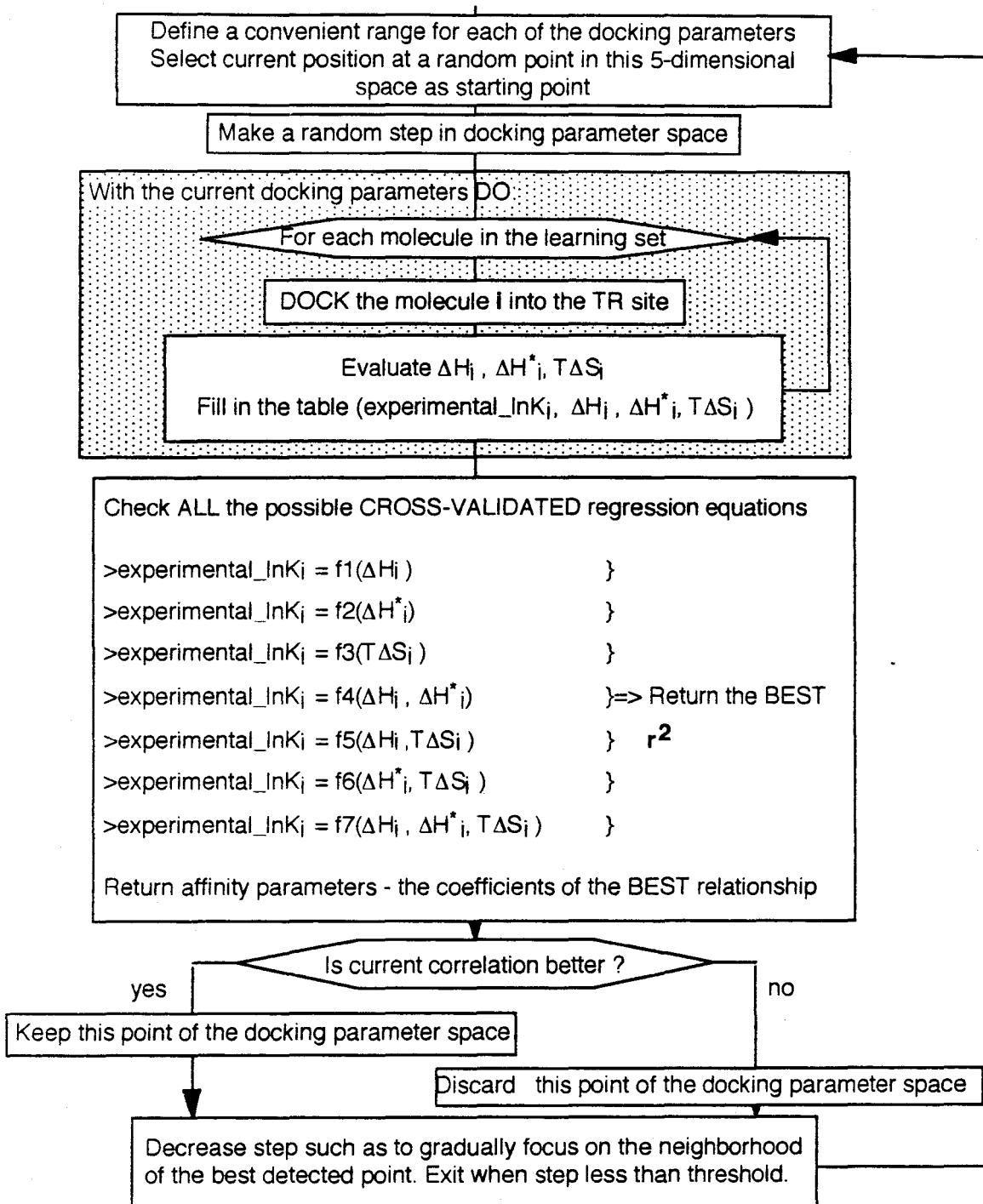


Fig. B-1



Conclusions et perspectives

Nos différents travaux sur les modèles de solvant continu ont montré leurs capacités à évaluer correctement l'intensité des processus de solvation malgré les hypothèses très simples sur lesquelles ils reposent. En plus des corrélations entre les énergies de transfert vide/eau, l'approche du « solvant continu » a démontré son applicabilité aux problèmes liés à l'inhibition des enzymes par des ligands.

Les modèles de solvant continu étudiés font cependant abstraction de certains effets importants, comme la viscosité du solvant et il serait par conséquent intéressant d'étudier de tels modèles dans des conditions dynamiques. Dans ce but, la recherche d'une stratégie permettant l'évaluation rapide du *gradient* du potentiel de solvation par rapport aux coordonnées, basé sur l'intégration de l'équation de Poisson, et d'inclure ce potentiel dans l'Hamiltonien des simulations de dynamique moléculaire, est en cours.

Afin d'obtenir un terme de solvation en accord avec le champ de forces consistant décrivant les interactions interatomiques, il est nécessaire de considérer le potentiel de solvation comme une contribution supplémentaire au champ de forces et donc de l'ajouter aux anciens termes décrivant la molécule en l'absence de solvant. Une évaluation correcte des paramètres utilisés par le potentiel de solvation devient donc très importante pour l'efficacité du champ de force ainsi « corrigé » par la contribution de solvation. L'optimisation de ces paramètres effectuée dans le but de maximiser la concordance entre les énergies de transfert vide/eau des petites molécules a contribué à accroître la fiabilité des modèles de solvant continu proposés dans ce travail. Néanmoins, ce travail

d'optimisation n'a pas été systématique ou exhaustif - des résultats ultérieurs ont en effet démontré que la qualité de ces modèles pouvait encore être améliorée en utilisant des stratégies de paramétrisation plus efficaces.

Le modèle de prédiction des affinités des ligands pour l'enzyme Trypanothione Reductase représente un outil important, pouvant être utilisé par les pharmacochimistes travaillant dans ce domaine. Tout en continuant à cribler les bases de données moléculaires, une application intéressante de ces programmes a été la sélection de nouvelles molécules à synthétiser sur la base des affinités calculées. Ce type d'application continue à faire l'objet de publications en collaboration avec les équipes des pharmacochimistes du laboratoire. Cependant, il reste plusieurs points à éclaircir en ce qui concerne le modèle de prédiction d'affinité. Notamment, la faible participation du terme enthalpique à l'énergie libre calculée, n'a pas trouvé à ce jour une explication définitive. Le caractère non-linéaire des équations rend l'évaluation de l'importance relative de certains paramètres très difficile. Poursuivre la généralisation de ce modèle d'affinité pour des autres enzymes permettrait d'élargir l'échelle des affinités des composés utilisés pour la calibration du modèle et par conséquent de mieux valider les valeurs optimales adoptées par les paramètres suite à cette calibration.

Si l'erreur standard des prédictions des constantes d'inhibition (de l'ordre d'une unité $\log K$) suffit pour faire la différence entre des inhibiteurs potentiels de la Trypanothione Reductase et des composés inactifs, elle est trop importante pour pouvoir expliquer des effets plus subtils concernant les mécanismes d'inhibition de cette enzyme. La prise en compte explicite de la flexibilité du ligand et du site enzymatique pendant le calcul de "docking" pourrait accroître d'une manière significative la qualité des prédictions du modèle. Néanmoins, le grand nombre

de degrés de liberté à prendre en considération pour une telle approche nécessiterait l'utilisation d'algorithmes d'optimisation adaptés, tels que par exemple les algorithmes génétiques.

Une autre direction de recherche envisageable afin d'améliorer les prédictions de l'affinité des inhibiteurs polycationiques pourrait être la reparamétrisation du champ de forces afin de mieux représenter entre autres les interactions entre les cations et les systèmes aromatiques, qui semblent jouer un rôle important dans le processus d'inhibition de la Trypanothione Reductase.

Enfin, tout ceci donne une bonne image des possibilités et des limites des techniques de simulation moléculaire, où, comme dans n'importe quel autre domaine de la science, ce qui reste à réaliser est par définition plus intéressant que ce qui est déjà acquis.